Sonar-to-RGB Image Translation for Diver Monitoring in Poor Visibility Environments

Bilal Wehbe, Nimish Shah, Miguel Bande and Christian Backe

DFKI - Robotic Innovation Center Bremen

Germany, 28359 Bremen

Email: {bilal.wehbe, nimish.shah, miguel.bande_firvida, christian.backe}@dfki.de

Abstract—This work evaluates a method for generating visual-like images from sonar images using Generative Adversarial Networks (GANs), for the purpose of monitoring technical divers working in low visibility environments. The general goal is to enhance the interpretability of sonar images in order to assist emergency operators that monitor the safety of the divers. To train the models, sonar and visual data were collected over the course of three trials from two different sites, an indoor pool and a lake. We evaluate and compare two different generative models namely, a modified version of *pix2pix* and *vid2vid*. Results show that it is possible to recover visual information from sonar data when the camera image is highly disturbed.

*Index Terms*—Multi-modal learning, GANs, image-to-image translation, marine perception

I. INTRODUCTION

Professional divers still play an important role in many underwater applications related to inspection and maintenance. While even the regular technical or industry diving is inherently dangerous, safety aspects are even more important for diving activities related to civil protection and emergency response. Divers have to perform activities in complex and mainly industrial or artificial environments, such as ports, piers, industrial basins, and channels, which makes it difficult to navigate and self-localize. Additionally, even if visibility is good at the beginning of the dive, activities performed by the divers cause the turbidity of the water to increase, leading regularly to white-out situations, where the team on-land monitoring the divers is not able anymore to detect them visually. Sonar imaging is an alternative to visual sensing which is not affected by bad visibility and poor lighting conditions and can be used to mitigate this issue. However, sonar images are often very hard to interpret even by a trained human eye and suffer usually from low signal-to-noise ratios.

In this work, we aim to learn an end-to-end association between sonar and optical camera images observing the same underwater scene. The idea is to use this learned association to generate realistic visual-like images given only sonar images as input or a combination of a sonar image and a dark or turbid optical image. The purpose of this algorithm 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.

(a) Camera image of a diver at the beginning of a dive
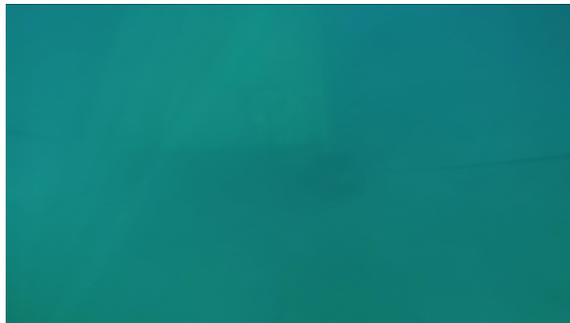
(b) Camera image after working for few minutes

Fig. 1: The difference between clear visibility conditions at the beginning of a dive and strong turbidity after few minutes of the diver performing underwater activities.

A. Related work

A sensor modality is defined as the means by which an instrument perceives or measures the physical world, such as sound, light, pressure, temperature, etc. A research problem is thus considered as multi-modal when two or more sensor modalities are used simultaneously to capture a scene. Combining multi-modal sensory data can benefit to enhance the perception of the environment as well as reduce perceptual ambiguity in challenging conditions such as the ones faced underwater. Three main benefits can be identified [1] i) combining multiple modalities that observe the same phenomenon produces more robust predictions by exploring supplementary and redundant information, ii) having access to multiple modalities might allow to capture complementary information which might not be perceived by a single modality, and iii) a

system that fuses multiple modalities can still operate when one of the modalities is missing, and even, predict the missing modality from the existing ones.

Given the intrinsic heterogeneity of the data, the field of multi-modal learning, i.e., field of machine learning that uses data from different modalities to reason or build models to describe the physical world, brings also some unique challenges for computational researchers [2]. The main challenges faced in this work are translation and fusion, as a clear visual-like image must be generated using a sonar image and a highly distorted optical image if present. In recent years, deep-learning-based multi-modal learning methods have attracted much attention from the research community due to their flexibility and powerful abstraction capabilities.

One of the earliest and most popular image-to-image translation methods is known as *pix2pix* [3]. *pix2pix* makes use of a concept known as Conditional GAN (CGAN) [4], where a condition is imposed on both the generator and the discriminator inputs to compel the network to perform translation tasks. To keep image consistency and better model high-frequency structures during translation the authors use CGANs with the combination of U-Net [5], i.e., skip connection between the encoder and the decoder of the generator, and PatchGAN [6], i.e., division of the output image in $N \times N$ patches that the discriminator must classify as real or fake.

Although *pix2pix* is mainly applied with optical images for cross-domain image translation tasks, it has been also applied in the underwater domain for fish monitoring in [7], where daytime underwater images are generated from an optical underwater camera and an imaging sonar on night-time. Here, the authors added a slight modification to the *pix2pix* architecture in order to fuse the sonar and a darkened camera image at the input of the generator by concatenating the inputs from both modalities at the first layer of the encoder. Hereafter, this algorithm is referred to as Nightvision.

Conditional GANs have enabled a variety of applications, but the results are often limited to low-resolution and lack of details and realistic textures, as its adversarial training might be unstable and prone to failure for high-resolution image generation tasks. *pix2pixHD* [8] is an extension of *pix2pix*, which enables the generation of high-definition images, by adding two new elements to the previous architecture: a coarse-to-fine generator, i.e., decomposition of the generator into a global generator network and a local enhancer network, and multi-scale discriminators, i.e., 3 discriminators that have an identical network structure but operate at different image scales.

As we are dealing in our case with translation of a continuous stream of sonar and optical images, the generated sequences of images not only must be photo-realistic individually but also temporally consistent as a whole. *vid2vid* [9] is an extension of *pix2pixHD* to address time dependency for a sequence of images. This architecture merges a current source image together with past source images as well as past generated prediction. Using this technique, temporal information between subsequent frames can be preserved which helps to
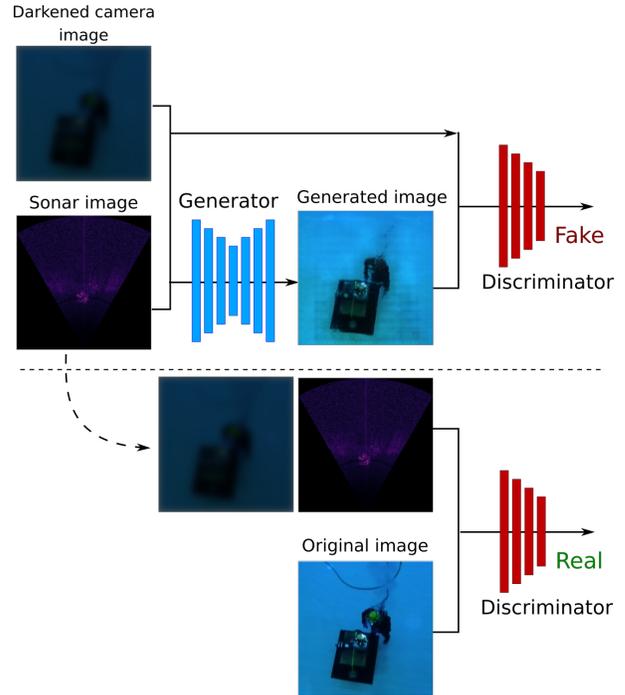

Fig. 2: General network training methodology for CGANs

improve the performance. Additionally, the authors make use of the inherent redundant information of consecutive images, by combining the resulting synthesized intermediate image with an optical-flow warped version of the last generated image by means of a soft occlusion map for attention-based aggregation.

To the best of our knowledge, either *pix2pixHD* or *vid2vid* has not been applied in the underwater domain yet.

II. METHODOLOGY

In this work, we will present the evaluation of nightvision and *vid2vid* methods deployed to solve our task in the underwater domain. Both method are based on CGANs and have been accordingly modified to match the application requirements (see fig. 2 for an general overview of the training methodology used in CGANs).

It is worth noticing that in order to properly combine the sonar and optical images, the sensors were mounted in a specific position, so that the resulting points of view from both sonar and camera were aligned as accurate as possible, and the captures were synchronized during collection for temporal alignment. The alignment of the two points of view is hereby challenging due to the different mechanisms of optical and acoustic imaging. Whereas optical camera has a sensor array of photosites (one per pixel) which captures the light rays bounced by the environment and projects it into a 2D image as does the human eye, imaging sonars simultaneously emit multiple beams by using an acoustic lens and generates a 2D image by mapping the intensity of the reflected waves to the azimuth angle and ranges between the sensor and the reflecting object. In other words, whereas the *x* and *y* cartesian

coordinates in the optical camera are preserved, while the z coordinate gets projected on the image plane, the depth r and swath ϕ spherical coordinates are preserved in the sonar image, while the elevation θ gets projected on the image plane. This inevitably results into a different perspective of the same scene (see fig. 7 for comparison between the capture of the same scene by an optical camera and different multibeam imaging sonars).

A. Nightvision

In the case of Nightvision, both inputs (sonar and camera image) are concatenated and passed through the 3×3 convolutional layer and concatenated, then through seven downsampling CBR (Convolution–BatchNorm–ReLu) layers in the encoder and seven upsampling CBR layers in the decoder. Each CBR layer consists of a 4×4 convolutional neural layer or a 4×4 deconvolutional neural layer and a rectified linear unit (ReLU) activation layer through batch normalization, which is an adaptation of the architecture presented in [10]. The final 3×3 convolutional layer outputs an image. We also add dropout layers in the decoder part to make the model non-deterministic and prevent overfitting. This network is trained for 100 epochs with batch size 8. To train these networks we use Adam optimizer with 2×10^{-4} as a leaning rate and 0.5 as β_1 and 0.999 as β_1.

For a more detailed description about the architecture employed, see appendix A.

B. Modified vid2vid

In the case of *vid2vid*, where the network adopts a coarse-to-fine architecture, the sequential generator uses residual network. We specifically use flownet2 [11] for this purpose. Furthermore, multiple discriminators are trained to mitigate the mode collapse problem during GANs training, i.e. conditional image discriminator and conditional video discriminator to resemble the spatial and temporal dynamics of the original video. This network is also modified to consider sonar images. For this purpose sonar and camera images are passed through convolutional layers and then concatenated.

These concatenated images undergo several residual blocks to form intermediate high-level features (global generator). Intermediate features are added and fed to residual layers to output intermediate image, flow map mask.

For high resolution videos local generator is used in combination with global generator. This local generator downsamples the input and feeds it as an input to global generator. Further the output of global generator is added to intermediate feature layer of local generator. This summed output is then passed through another series of residual blocks to output the higher resolution images.

For a more detailed description about the architecture employed, see appendix B.

C. Evaluation Metrics

As evaluation metrics for comparison, we use the structural similarity (SSIM) index, which is the most popular perceptual
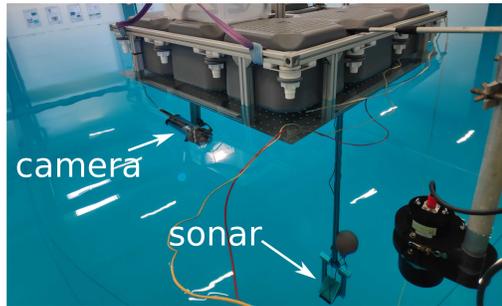

Fig. 3: Sensor setup used for data gathering.

image similarity metric. SSIM values range from 0 (dissimilar) to 1 (similar) [12]. As SSIM is calculated by comparing statistical values for each local region, it is possible to perform a robust similarity evaluation under the influence of noise and distortion.

III. DATA ACQUISITION AND MANAGEMENT

A. Data acquisition setup and locations

Data of divers performing technical underwater tasks were collected using a multi-beam imaging sonar and a stereo camera. The field of view of both sensors was aligned to observe the same scene simultaneously. An image showing the sensor setup is depicted in fig. 3.

This sensor setup was used in three data acquisition sessions between August 2021 and April 2022, at two locations in Germany: Two sessions under laboratory conditions at the Maritime Exploration Hall [1] at DFKI-RIC in Bremen, one field session at the Kreidesee in Hemmoor. In all cases, a typical underwater working environment for the divers was simulated, including a workbench and tools. Divers performed various tasks such as mounting pipe flanges, wrenching nuts and bolts, or building armature.

The Maritime Exploration Hall contains a saltwater basin measuring $23\,\mathrm{m} \times 19\,\mathrm{m} \times 8\,\mathrm{m}$. Fig. 4 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 $50\,\mathrm{m}$. While these are comfortable circumstances for experiment setup and execution, additional measures are required to fully cover the intended application which focuses on difficult working conditions with low visibility in turbid water. This was addressed in part by the post-processing techniques discussed in section II, and in part by the field data acquisition session in Hemmoor. Additional field sessions are scheduled to follow.

The Kreidesee (chalk lake) is a former chalk mine with a surface area of $33\,\mathrm{ha}$ and a maximum depth of $60\,\mathrm{m}$. Fig. 5 shows an aerial image of the entire lake. The visual range under water is up to $25\,\mathrm{m}$, assuming ideal weather conditions. Our field experiments were conducted on a submerged flat grid platform with a $3\,\mathrm{m} \times 3\,\mathrm{m}$ surface area, $5\,\mathrm{m}$ underwater, $50\,\mathrm{cm}$

Fig. 4: Diver entering the saltwater basin in the Maritime Exploration Hall.
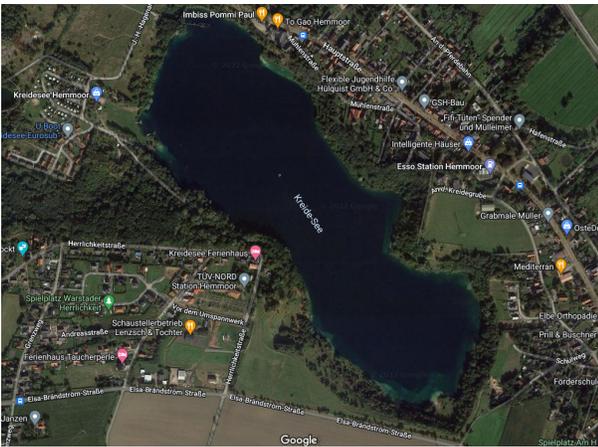

Fig. 5: Aerial image of the Kreidesee im Hemmoor (Google Maps).

above the floor of the lake, near the shore. The divers were able to perform different tasks than in the basin, including wood cutting with a chain saw, drilling, and shoveling. This field experiment also included a night session.

B. Metadata generation and data management

To facilitate the replication of our results, and to support the potential re-use of our data in other applications, the data collection was performed following the FAIR principles for scientific data management [13]. These require payload data to be accompanied by comprehensive metadata, in order to help re-users evaluate whether the data matches their own use case and integrate it into their work environment.

Since there are currently no universal metadata standards, we created a custom design, and collaborated with the project NFDI4Ing [14], which collects and aims to harmonize different metadata approaches from all engineering sciences.

Different types of metadata were collected before, during, and after the acquisition of the payload data. The workflow is depicted in fig. 6
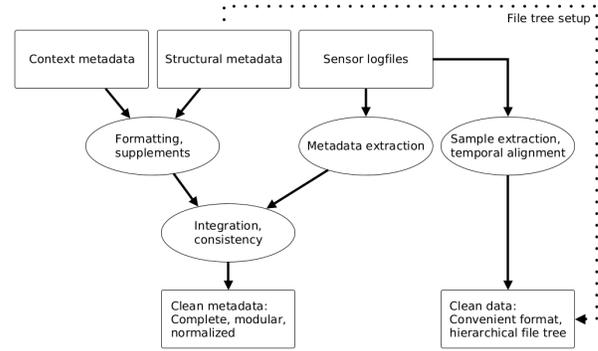

Fig. 6: Workflow for metadata generation and processing.

Before data acquisition, we designed a hierarchical data model as a framework for structural metadata, which describes the different data elements, their relationships, and their locations within the data repository. Structural metadata ensures interoperability between data producers and consumers on different stages of the processing pipeline. Its format was designed early to avoid costly re-adjustments over the data life-cycle. The main elements of our data model are

- `session`: a collection of datasets captured within a certain time frame, at the same location, by the same people, using the same type of equipment, showing the same kind of scenes.
- `dataset`: a continuous, uninterrupted collection of samples, devoted to a specific experiment or research goal, using a subset of the available equipment, processable by a certain kind of algorithms, publishable as a unit.
- `sample`: a collection of data files within the same dataset that are captured at approximately the same instant in time, may have different modalities (e.g. camera, sonar), and may be associated with labels (e.g. class name, object names) for different machine learning tasks.
- `modality`: a type of sensor associated with a certain data file format and encoding, and possibly a set of calibration and configuration parameters.
- `data file`: the foundational element of samples and datasets.

For the collection of metadata during data acquisition, we created an entry tool containing fields for structural metadata items (e.g. names and descriptions of sessions and datasets), as well as contextual metadata. The latter describes the environment conditions during data acquisition, e.g. visibility, used equipment, potential malfunctions, or other unforeseen events. This type of metadata was created collaboratively by machine learning experts, divers, and data managers to ensure its relevance and consistency. The entry tool was based on multiple online spreadsheets, to support collaboration over the internet, and to facilitate later post-processing and relational database import. The tool was especially helpful under field conditions, where capturing context metadata is very important, because the environment cannot be fully controlled and must therefore be documented, but where scientists tend to

prioritize safe operation and reliable recording of the payload data, so there are fewer resources to deal with additional tasks.

In addition to the metadata items collected manually with the entry tool, parts of both structural and contextual metadata were captured automatically within the payload data. This included e.g. sensor configuration parameters, sampling rate per modality, or file size. These were extracted after the data acquisition, and integrated with the other metadata. The entire metadata set was cleaned, normalized, and complemented where necessary. The internal consistency between all metadata items was re-checked. Parts of free-text descriptions that should be machine-readable were extracted and encoded.

Another structural metadata task performed after data acquisition was the temporal alignment of the samples of different modalities. The inputs to the translation training must be pairs of sonar and camera images that were captured at approximately the same time. For practical reasons, these were not already matched during data acquisition. The matching done afterwards had to account for different sampling rates of sonar and camera, and occasional dropouts in either modality.

IV. RESULTS

In this section we present results results obtained using aforementioned Nightvision and modified-*vid2vid* algorithms. These models were trained on dataset collected at DFKI's Maritime Exploration Hall at Bremen and on field session at the Kreidesee in Hemmoor.

These images were collected using Zed stereo camera processed (added noise) to look dark and blurry with 0.2 to 1 as darkness factors. Corresponding sonar images were captured using either Gemini 720i or Oculus M1200d Sonar. fig. 7 shows representative images from collected and processed dataset.

A. Nightvision

We train both sonar and without sonar versions of Nightvision in two manners *i)* training various darkness levels individually, and *ii)* combining camera and sonar images of all darkness levels, thereby expecting network to be less sensitive to introduced noise. For data collected in Hemmoor, the network is trained only in combination of sonar and camera image with mixing various darkness levels.

Dataset Description:
1) Gemini 720i in maritime exploration hall (cf. 7a):
   - image size: $1024\,\text{px} \times 512\,\text{px}$
   - training samples: 1000×7 darkness levels
   - testing samples: 427×7 darkness levels
   - train and test samples from same dataset
2) Oculus M1200d in maritime exploration hall (cf. 7b):
   - image size: $512\,\text{px} \times 512\,\text{px}$
   - training samples: 3078×7 darkness levels
   - testing samples: 770×7 darkness levels
   - train and test samples from same dataset
3) Oculus M1200d at Hemmoor (cf. 7c):
   - image size: $1024\,\text{px} \times 512\,\text{px}$
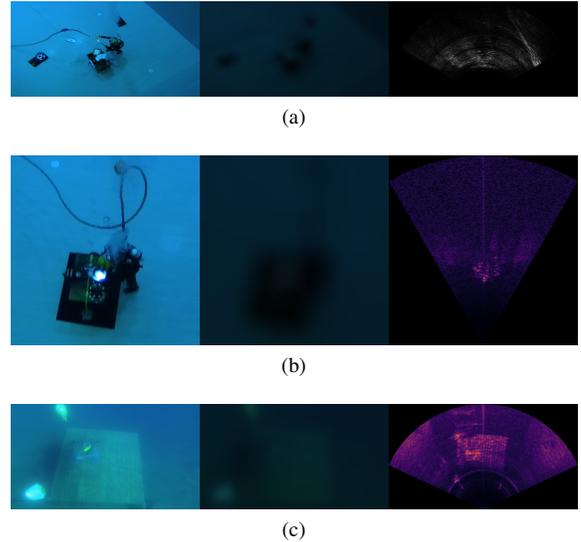
(a)

(b)

(c)

Fig. 7: Representative image of images in the datasets. (a) Using Gemini 720i, (b) Using Oculus M1200d sonar in Maritime Exploration Hall, (c) Using Oculus M1200d at Kreidesee. *(left)* Original Camera Image, *(middle)* processed camera image at 0.75 darkness level, *(right)* polar sonar image

   - training samples: 3328×7 darkness levels
   - testing samples: 3232×7 darkness levels
   - train and test samples from different datasets

The fig. **??** to fig, 14 show two of samples predictions, whereas fig. 8 shows comparison between both training methodologies.

B. Modified-vid2vid

This network is trained on data collected at maritime exploration hall with Gemini 720i sonar (cf. 7a). To train this network, we divide sequentially captured in sequence of 30. The fig. 9a and fig. 9c shows first and fifth frame of the predicted sequence at darkness level of 0.5.

C. Discussion

From SSIM plots it is evident that augmenting camera image with sonar image for Nightvision algorithm yield better results. It is important to note that this improvement is predominantly visible at higher darkness levels. This can be explained by limited information content present in camera images at higher darkness levels. By combining multiple darkness levels in the training datasets, we try to enforce less variance towards darkness in the camera image. This can be observed in the improved performance of the model when trained with a combination of darkness level over the models trained with a single darkness at a time. In case of datasets captured using Gemini 720i sonar, it can also be seen that prediction drastically shifts towards an average image, this is due to lower resolution of the sonar. Additionally, this dataset is relatively smaller when compared to the other datasets and lacks the variety in the captured scenery. In this case, the high SSIM
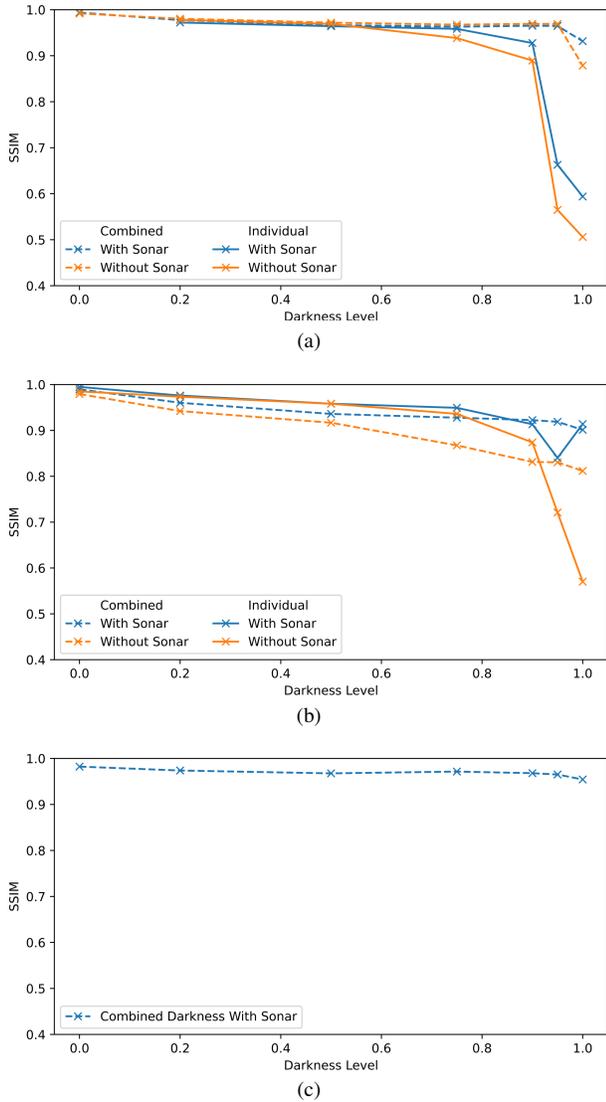
Fig. 8: Structural Similarity Index showing comparison between results when camera image is augmented with sonar data.

index even at high darkness values could be explained by the small size of the subject in the image compared to the plain background which occupies a big portion of the scene.

For *vid2vid*, it can be observed from fig. 9 that predictions, specifically high level features, look spatially and temporally consistent. In fig. 9b, an overlay between two frames is visible. This issue is attributed to some temporal dropouts while capturing images. Another factor which causes this behavior is shift (e.g., diver movements etc.) between two consecutive frames is bigger due to low frequency of image capturing. This drift is increased over time causing a loss of both low and high level features in the predicted images.

V. Conclusion

In this work, we compared two generative methods, namely a modified *pix2pix* and *vid2vid*, for reconstructing underwater
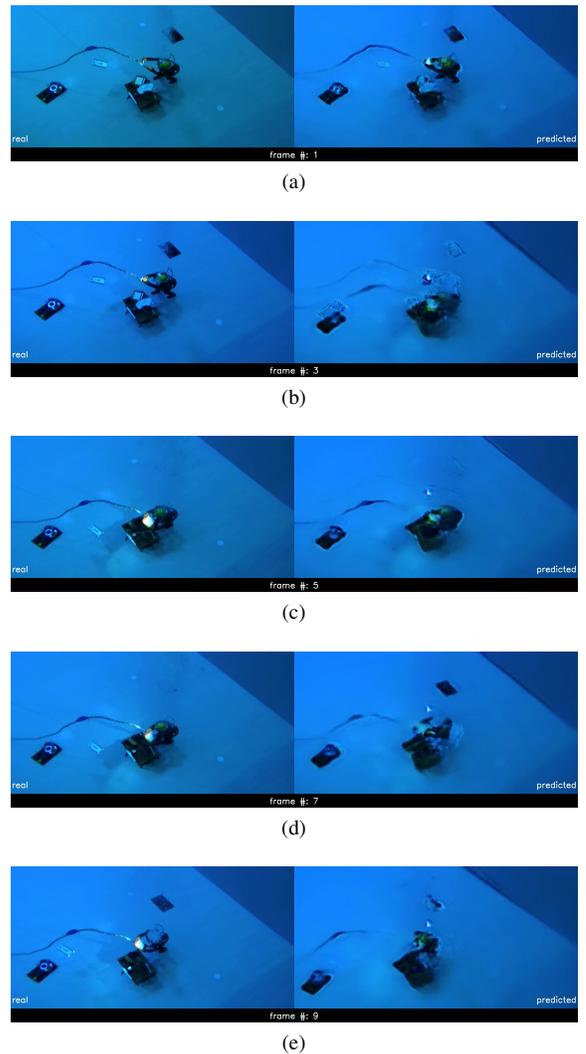

Fig. 9: Results of modified-*vid2vid* network with 0.5 darkness level. Image shows first five odd frames.

images of the divers by taking a sonar imaging and a disturbed camera image as input. Three datasets were used in this study that collected from two different sites, one indoor and one outdoor. Regarding the *pix2pix* algorithm, our results show consistently that adding a sonar image and mixing the different darkness levels of camera images on training helps improves the quality of the reconstructed image. Two essential findings can be noticed here. Firstly, the higher sonar resolution resulted in better quality predictions when compared to the lower resolution counterpart. Secondly, a noticeable improvement in the results was observed when combining different darkness factors into the training datasets.

Due to the limited available computational resources, *vid2vid* was only evaluated on the first dataset by the time of submitting this paper. Results show the the predicted images resemble closely the ground truth in the beginning of the sequence, however the quality of the prediction degrades with time. This result could be attributed to the low resolution of the

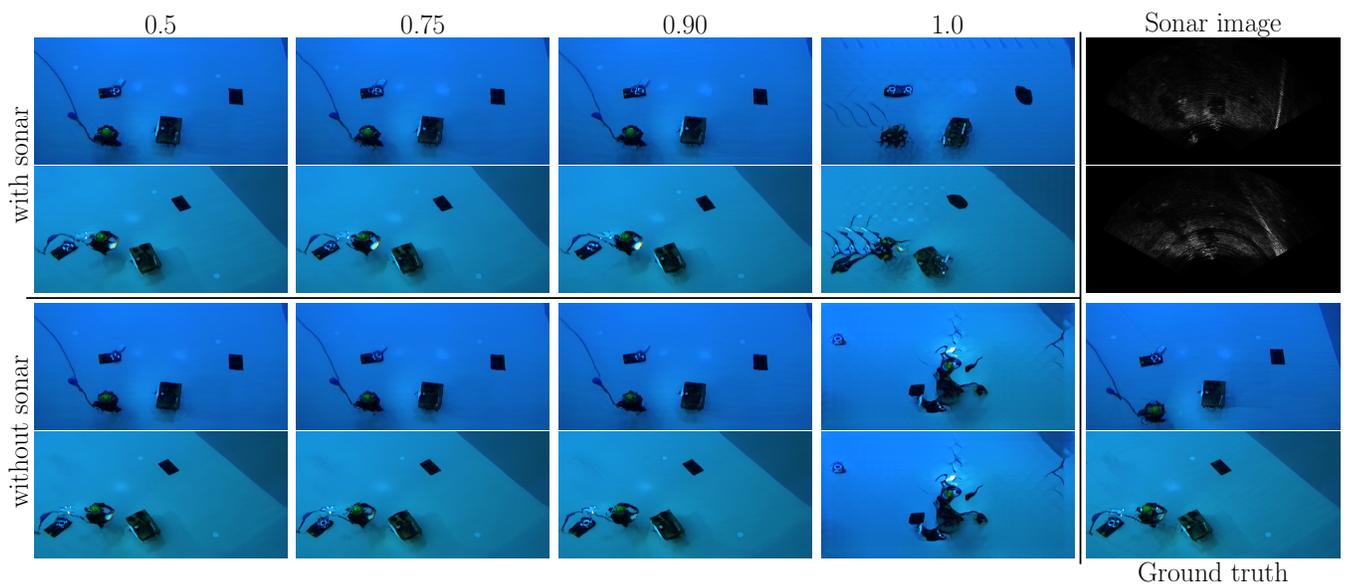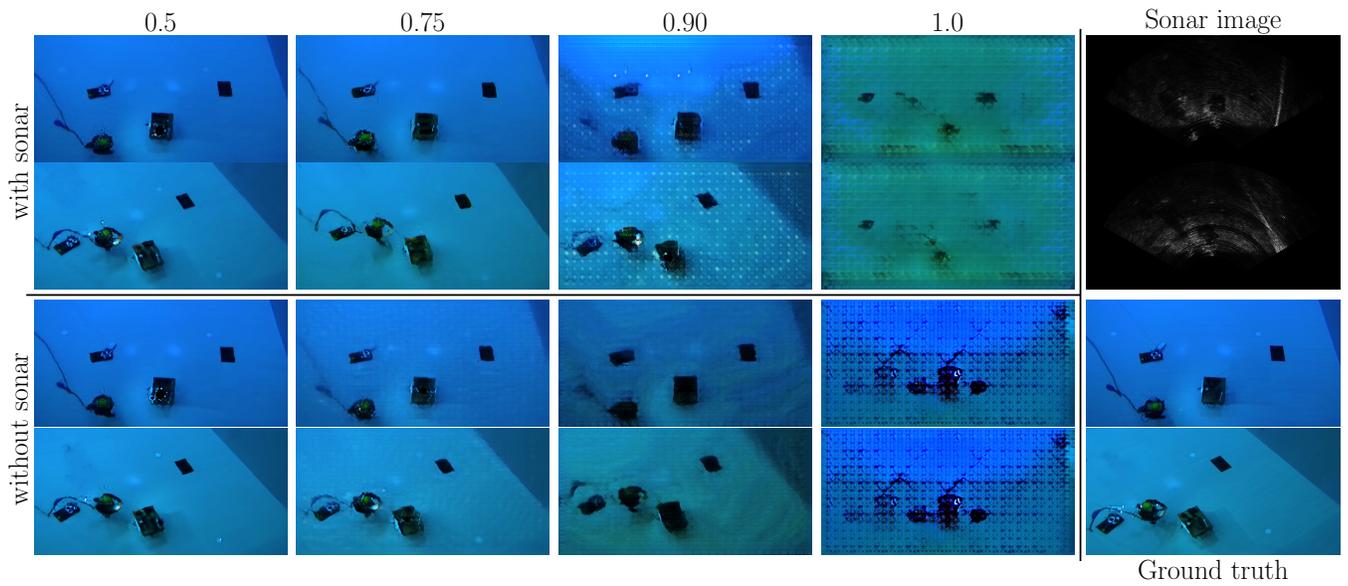(a) Individual darkness levels

(b) Combined darkness levels

Fig. 10: Consolidated view of results of nightvision experiments; Top rows of each block show results with sonar, while bottom rows show results without sonar. Darkness level varies along column. Rightmost column documents ground truth images.

Gemini sonar that was used in this experiment, where the network was only able to benefit from little information present in this dataset. Another factor the might have contributed to that would be low frame rate of the captured images which caused a noticeable difference between the consecutive frame.

As future work, we plan to cross-validate the different models on the datasets that were not used during the training phase. This will aim to validate the transferability of models. Additionally, the *vid2vid* method will be further investigated to improve the quality and computational efficiency of the model.
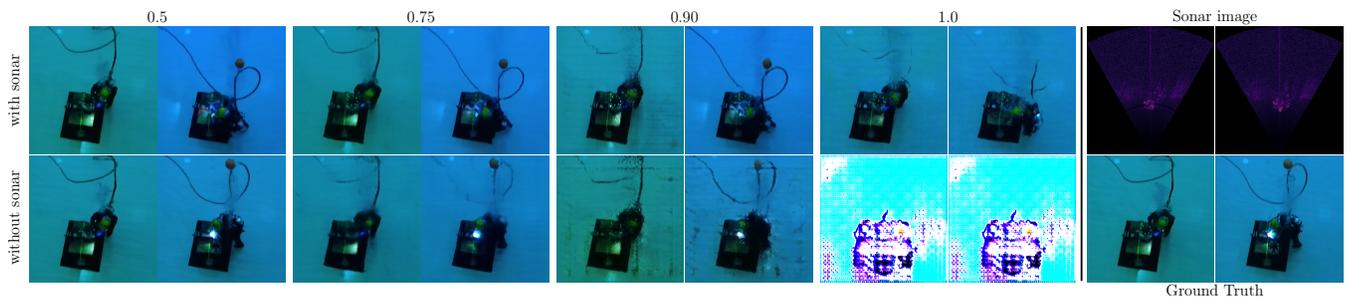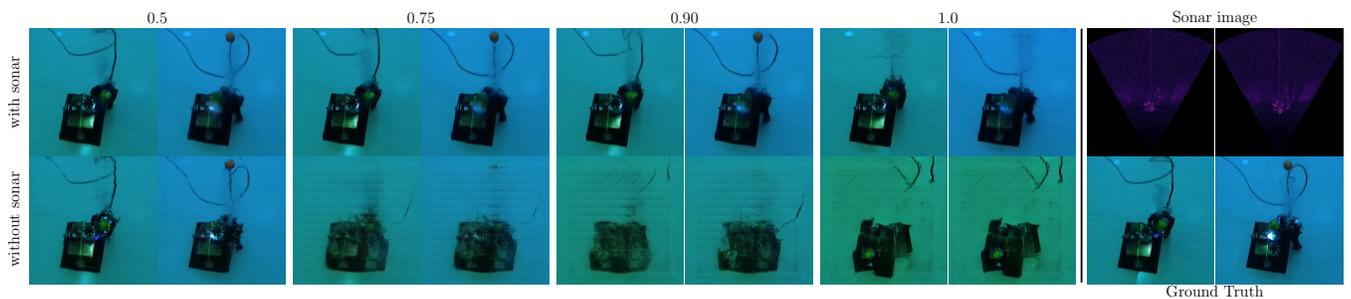
APPENDIX

A. Nightvision Network Architecture

The Nightvision network architecture is inspired by U-Net based Pix2Pix architecture. The generator and discriminator network architectures are shown in fig. 15.

Initially, sonar and camera image is passed through 3×3 convolutional layer and then 7 up- and downsampling layers

(a) Individual darkness levels

(b) Combined darkness levels

Fig. 11: Consolidated view of results of nightvision experiments on Oculus M1200d in maritime exploration hall
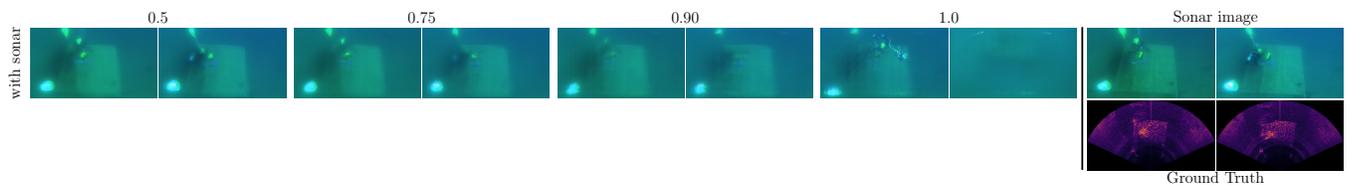

Fig. 12: Consolidated view of results of nightvision experiments on Oculus M1200d at Hemmoor (Combined darkness levels only)

each. Further output of this layer is passed through 3×3 convolutional layer.

Each downlampling layer consists of a 4×4 convolutional layer followed by batch normalization and ReLU. Whereas, each upsampling layer consists of 4×4 transpose convolutional layer followed by batch normalization and ReLU. Dropouts are also implemented.

The discriminator forms PatchGAN model which outputs average of probabilities over all the patches.

B. Modified Vid2Vid Network Architecture

This network adopts coarse-to-fine architecture as described in [9]. Initially, camera and sonar images are passed through convolutional layer and concatenated. Further, this concatenated vector is passes through series of residual blocks. Same process is applied to the previously generated image.

Each residual block is a sequence of `padding-convolution-normalization-relu-padding-convolution`. To estimate optical flow flownet2 [11] is used in its vanilla form.

REFERENCES

[1] W. Burgard, A. Valada, N. Radwan, T. Naseer, J. Zhang, J. Vertens, O. Mees, A. Eitel, and G. Oliveira, "Perspectives on deep multimodel robot learning," in *Robotics Research*, pp. 17–24, Springer, 2020.

[2] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[3] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

[4] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[6] C. Li and M. Wand, "Precomputed real-time texture synthesis with markovian generative adversarial networks," in *European conference on computer vision*, pp. 702–716, Springer, 2016.

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

[8] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with condi-
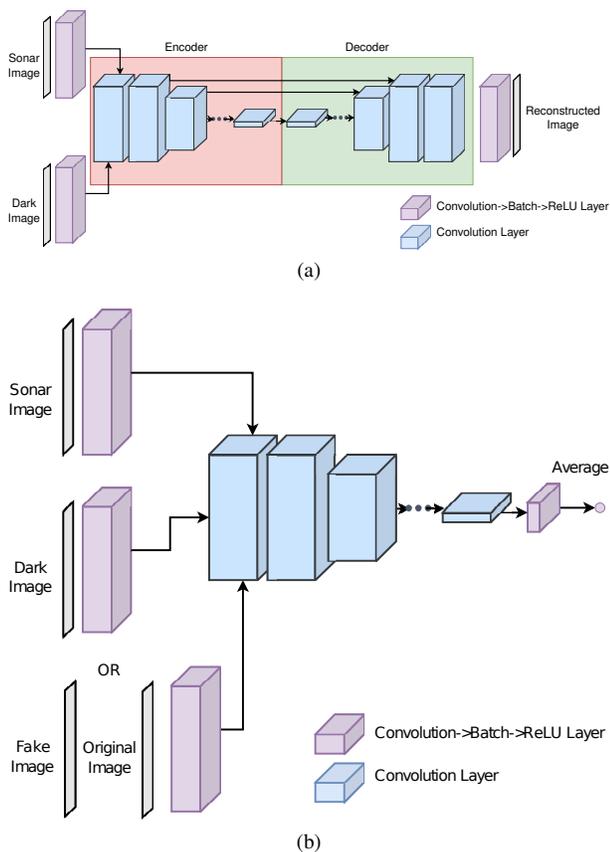
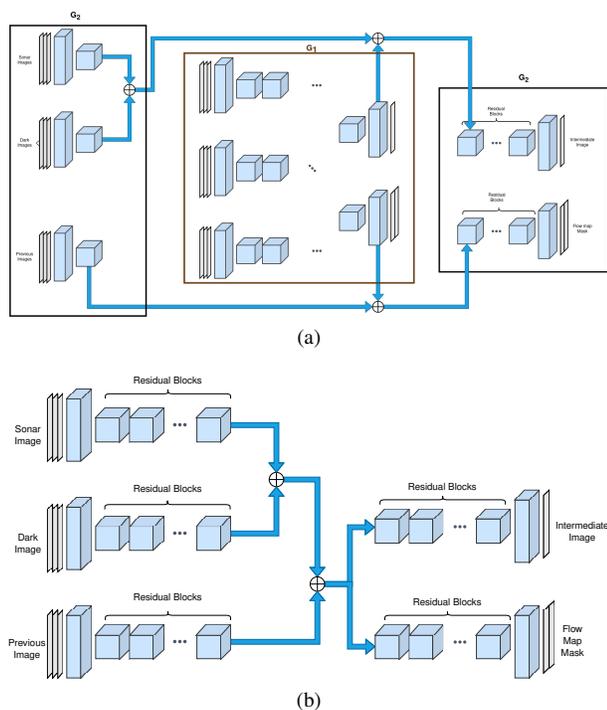Fig. 13: Nightvision architecture (a) generator, (b) discriminator

Fig. 14: Modified Vid2Vid architecture (a) local generator, (b) global generator

tional gans," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8798–8807, 2018.

[9] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro, "Video-to-video synthesis," *arXiv preprint arXiv:1808.06601*, 2018.

[10] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[11] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2462–2470, 2017.

[12] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[13] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[14] "NFDI4Ing homepage." https://nfdi4ing.de, 2022. Accessed: 2022-08-25.