Bridging the Sim-to-Real Gap with Explainability for ML-based Object Detection on Sonar Data*

Şakir Furkan Yöndem[1,*,†], Ramin Tavakoli Kolagari[1,†] and Benedikt Schlereth-Groh[1,†]

[1]Technische Hochschule Nürnberg, Keßlerplatz 12, 90489 Nürnberg, Germany

Abstract

Saving drowning victims is time-critical, but detecting people underwater is highly challenging due to poor visibility and large distances. While side-scan sonar (SSS) is widely used for seafloor mapping and debris detection, human detection in sonar data remains largely unexplored. Training deep neural networks for this task requires a large dataset, but collecting real maritime data is difficult and expensive, making a synthetic data generation approach necessary. We introduce SimWave, a simulation environment designed to generate synthetic data for underwater human detection. We train deep learning models on real, synthetic, and hybrid datasets, evaluating their performance on real sonar images. The contribution of this paper lies in combining synthetic data generation with Explainable Artificial Intelligence (XAI) to systematically refine artificial datasets, addressing the gap between synthetic and natural data to enhance real-world performance—–an approach not previously explored in underwater sonar-based human detection. To gain insight into the model's decision-making process, we apply XAI techniques to analyze how attention shifts between real and synthetic training data. This helps visualize the synthetic-real data mismatch, refine synthetic data, and enhance model performance in real-world conditions. Our experimental results show that models trained on hybrid datasets, supported by XAI-based analysis, achieve notable performance improvements and better generalization. XAI helps identify domain gaps between real and synthetic data, allowing for dataset refinement and improved model accuracy. These findings highlight the effectiveness of synthetic generated data in training deep learning models for underwater human detection and emphasize the critical role of XAI in optimizing training data for real-world conditions.

Keywords

Underwater Human Detection, Side-Scan Sonar, Synthetic Data Generation, Explainable AI (XAI), Domain Bridging, Domain Gap, Deep Learning

1. Introduction

Human search and recovery operations in underwater environments are time-consuming and complex due to poor visibility, hazardous conditions, and operational challenges, placing a significant burden on specially trained divers [1]. Side-scan sonar is widely used to support such searches, but it still relies heavily on manual control and remains dependent on trained sonar operators [2, 3]. Artificial Intelligence (AI)-based automated analysis methods offer a promising alternative to speed up the process and reduce human intervention. However, the effectiveness of these methods depends on the availability of large, high-quality training datasets [4, 5]. In specialized areas such as sonar imaging, collecting real-world data is particularly challenging, and the lack of diverse data limits model generalization, leading to performance degradation in unseen or underrepresented scenarios. The key barriers to data collection include high costs, security risks, legal restrictions, and operational complexities [6].

To address these challenges, synthetic data generation is increasingly used for training deep learning models. Simulation-based approaches enhance model generalization by generating datasets that simulate diverse real-world conditions [7, 8]. However, structural differences between synthetic and real data (domain gap) can cause models to underperform in real-world applications. Optimizing synthetic data

Late-breaking work, Demos and Doctoral Consortium, colocated with the 3rd World Conference on eXplainable Artificial Intelligence: July 09–11, 2025, Istanbul, Turkey

*Corresponding author.

†These authors contributed equally.

sakirfurkan.yoendem@th-nuernberg.de (F. Yöndem); ramin.tavakolikolagari@th-nuernberg.de (R. T. Kolagari); benedikt.schlereth-groh@th-nuernberg.de (B. Schlereth-Groh)

0000-0002-0067-1379 (F. Yöndem); 0000-0002-7470-3767 (R. T. Kolagari); 0009-0002-6621-9542 (B. Schlereth-Groh)

and integrating hybrid datasets can significantly improve generalization and model robustness [9].

The limitations of real-world side-scan sonar datasets for human detection and the low quality of available sonar images are the main motivation for this work. In order to overcome these issues, we introduce SimWave, a synthetic dataset designed for underwater human detection. We train the YOLOv8 and YOLOv11 models using both real and synthetic data and compare their performance on real sonar images. By leveraging XAI, we optimize synthetic images to enhance model decision-making and reduce errors caused by excessive brightness and reflections through a dynamic clipping technique applied to SimWave. Experimental results show that the dynamically clipped dataset improves detection accuracy and recall on real sonar images, demonstrating the effectiveness of synthetic data refinement.

The remainder of our paper begins by summarizing existing research on underwater human detection and situating our contribution within this broader context. We then detail the design and implementation of the SimWave simulation system, introducing the synthetic sonar data generation mechanism and the different datasets we use. We explore the complex relationship between real and synthetic data, pushing the boundaries of deep learning models. Using an innovative data optimization technique, we achieve noticeable improvements in model performance. Finally, we provide an in-depth analysis of our findings, draw critical insights from our experimental evaluations, and outline future directions to improve sonar-based detection.

2. State of the Art

In order to understand and solve the problem of detecting people in sonar images, important research areas are analysed below and suitable methods are presented.

2.1. Detection on Sonar images

Sonar images based on acoustic wave reflections can be effectively used in underwater object detection tasks with deep learning models such as convolutional neural networks (CNNs), as they can be represented as Red-Green-Blue (RGB) image [10]. Especially in the detection of submerged objects in turbid waters, object detection on sonar images has been successfully applied by various research groups. For example, Lu et al. improved the YOLO Network by replacing convolutional neural network layers with residual blocks and were able to detect different objects in sonar images [5]. Similar detection work on side-scan sonar as well, like the YOLOv7 Model [11] or YOLOv9 [12]. Humans can also be detected on sonar images shown by Hu and Liu by detecting an underwater rescue target on multi-beam imaging sonar [13].

2.2. Domain Gap

Since data collection is particularly expensive and time-consuming, expanding the dataset through a simulation environment to generate synthetic data can be promising. However, the synthetic-to-real domain gap, i.e., the differences between the simulation and the real world, remains a challenge. Kiefer et al. [14] demonstrated poor performance on real world data if trained on only synthetic data, but improvements with synthetic and real world data. Showcasing a gap for object detection trained on either simulated or natural data. While synthetic sonar images are generated for underwater mine-like objects [15] or wrecks on the seabed [16]. Both show how the combination of synthetic and real data can improve training results on sonar images as well, but do not show how big the gap is between training purely synthetic and purely real data [15, 16].

2.3. Explainable AI

Explainable artificial intelligence (x-AI) for object detection algorithms can help to understand at what features are important for the deep neural networks. Visual explanations can be provided by model-agnostic or model-dependent algorithms. By visualizing the last convolutional layer of the target

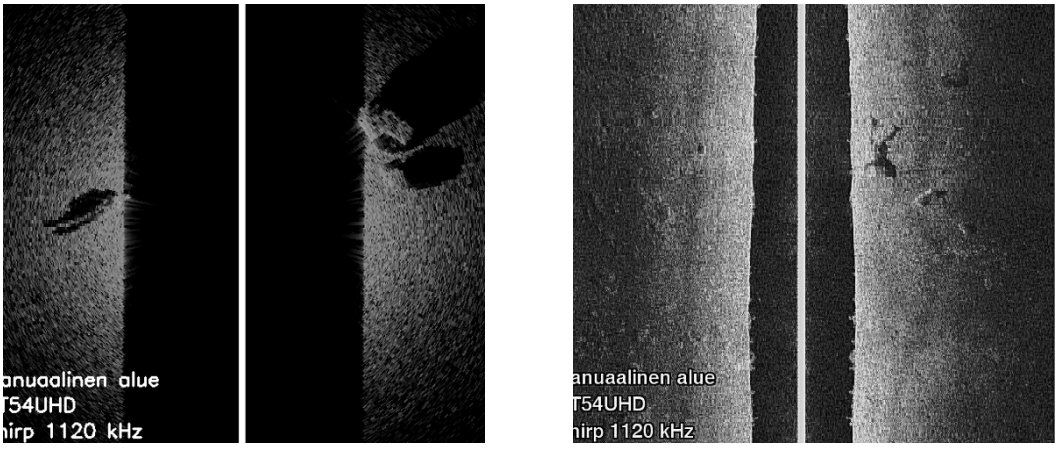
class, Grad-CAM can provide an explanation for the predicted class [17]. Petsiuk et al. improved the explanation by applying random masks and calculating the similarity to the original detection, and can explain localization and classification on images [18]. Since only the predictions are needed to calculate the saliency maps, this approach is also model-agnostic and can be used for different object recognizers [18].

3. SimWave

In this study, we developed a simulation environment to overcome the lack of real sonar data and train deep learning models for underwater human detection. We created this environment, which we call SimWave, using the Robot Operating System 2 (ROS2), Gazebo and Blender.

ROS2 is an open source framework that facilitates the management of autonomous systems, sensor-based data processing and the integration of robotic applications thanks to its modular structure. Support for real-time communication and distributed computing allows us to create simulation environments that reflect real-world conditions [19]. Gazebo provides a powerful platform for simulating robotic systems with its realistic physics engine and high-precision sensor modelling capabilities. In particular, its ability to model hydrodynamic interactions makes it ideal for testing the behaviour of sonar systems in an underwater environment. SimWave was developed to generate synthetic side-scan sonar images. First, we integrated the FLS sensor developed in Project Dave (Underwater Sonar Sensor Development Project) [20] into the simulation environment by making it compatible with ROS2. We then configured SimWave with two FLS sensors placed on the left and right sides to provide wide-angle data collection. Finally, we combined the data from these two sensors to produce an image similar to side-scan sonar. The sonar sensor parameters used in the simulation are as follows: The sensor captures 512 vertical samples within a range of -90° to +90°. Horizontally, it records 300 samples within a range of approximately between 0° and 5° to +20°.

To make the simulation environment compatible with the sonar images from the real data set, we used Blender to create a standard-sized 3D human model and various rock and wood objects. Blender is an open-source 3D modeling software that allows the modeling of physical objects, human figures and environmental elements. Figure 1 presents both simulated and real-world sonar images.



SimWave Aaltonen

Figure 1: Visual Representation of Simulated and Real-World Sonar Images

4. Improving Data Quality with XAI

In this section, we show how a dataset for finding submerged bodys in water can be improved by leveraging an XAI analysis. By evaluating the performance of object detection algorithms (YOLOv8

Table 1
Size of Traindatasets

Dataset	original Size	augmented data	total size
Aaltonen	145	206	351
SimWave	145	206	351
Aaltonen + SimWave	351	-	351

Table 2
Performance of trained models

Yolo-Modelle	Dataset	Precision	Recall	mAP50	mAP50-95
Yolov8	Aaltonen	0.937	0.808	0.901	0.561
Yolov8	SimWave	0.878	0.704	0.763	0.391
Yolov8	Aaltonen + SimWave	**0.943**	0.799	0.907	**0.631**
Yolov11	Aaltonen	0.859	**0.881**	**0.923**	0.621
Yolov11	SimWave	0.814	0.712	0.773	0.387
Yolov11	Aaltonen + SimWave	0.925	0.816	0.901	0.626

and YOLOv11) trained on different dataset compositions, the need for improvement of the synthetic data becomes apparent. With the attention of the object detection algorithms visualized using a XAI method, an adaption to the sensor simulator is proposed.

4.1. Datasets

In this study, both real and synthetic data are used to improve human detection in sonar images. The dataset consists of real-world data collected using side-scan sonar and synthetic data generated in the SimWave simulation environment. The real-world data was collected by Aaltonen[21], which provides waterfall images from side-scan sonar, and made available as a public dataset. We partitioned this dataset into 145 training, 61 validation, and 125 test images for model training and evaluation. Due to the limited size of the real-world dataset, we generated additional sonar images using the SimWave simulation environment. We designed this simulated dataset to align with the real-world dataset, incorporating 145 training and 61 validation images. To further enhance the model's generalization capability and increase data diversity, we applied data augmentation techniques tailored for the functionality of sonar sensors. Meanwhile, we constructed the hybrid dataset without augmentation, solely by merging real and synthetic data, resulting in a total of 351 training samples—145 real and 206 synthetic sonar images, as shown in Table 1.

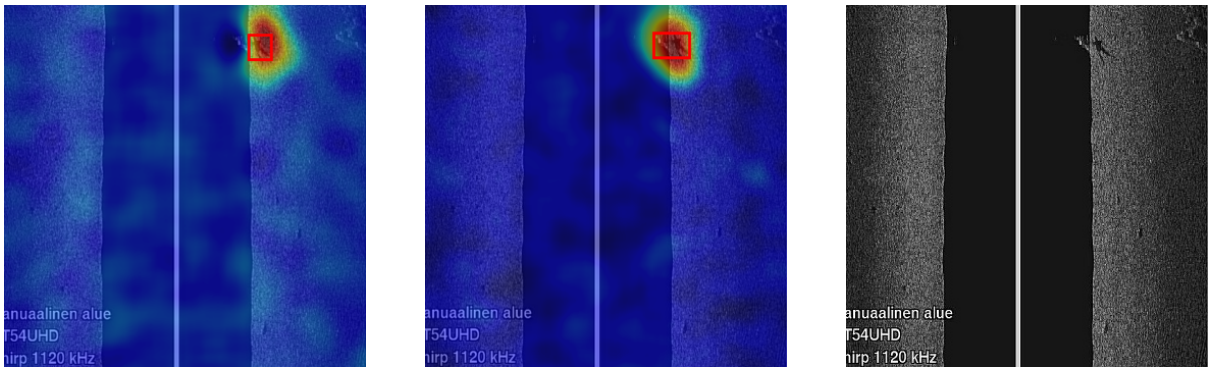
4.2. Model Evaluation

The performance of all models are validated on the same 125 test images from the real-world dataset collected by Aaltonen [21]. As discussed by Lu et al. YOLOv8 model performed well on the Marine Debrise Dataset and the Underwater Acoustic Target Detection Dataset [5]. We therefore decided to test the YOLOv8l [22] model as well and use the newest YOLO Model Yolov11l [23]. The performance of both models on different datasets to detect humans underwater are presented in Table 2.

The YOLOv8 model demonstrated superior precision values, whereas the YOLOv11 model exhibited enhanced performance in terms of recall. The hybrid dataset (Aaltonen + SimWave) facilitated a more balanced performance across both models, leading to an increase in mAP@50-95 and precision, thereby illustrating the capacity of synthetic data to augment the generalization capabilities of the models and provide more consistent outcomes under diverse sonar conditions. However, the models trained exclusively with SimWave experienced a decline in precision and recall, indicating a domain gap between the real and synthetic datasets. In order to understand which differences in detection are important, a XAI approach can provide insights.

4.3. Understanding Sim-to-Real Gap

Although only one sensor modality and a small selection of object detection algorithms are analyzed, the model-agnostic XAI approach called D-RISE is selected to enable an extension to other models and datasets in the future. Employing D-RISE, a methodology posited by Petsiuk et al. [18], calculates a saliency map, indicating important areas for proposed detections. Our analytical procedure involved the utilization of N=5000 masking iterations, with a stochastic masking probability of p=0.5 and a spatial resolution of (h,w)=(16,16) [18]. With the better precision values, the YOLOv8 detections on the same image with different models are visualized with the corresponding saliency maps.



(a) Aaltonen (b) SimWave (c) Hybrid

Figure 2: x-AI with DRISE on different models

In Figure 2 the saliency maps between the models trained on the real-world data set and simulated differ quite a lot. The model trained on the hybrid dataset missed the detection in the image and was therefore not able to provide a saliency map. With those visual analysis and similar comparisons, the following assumption for the synthetic sonar images are made:

- The model trained on real sonar images focuses on shadows during human detection.
- The model trained with synthetic data alone is more sensitive to reflections than to shadows.
- The hybrid model trained with real and simulated data makes errors in human detection because it has difficulty generalizing shadow and reflection information.

4.4. Improving Synthetic Data Generation

Based on the performance of the different models, an adaption to the synthetic data generation is needed. The analysis of the saliency maps indicates over-strong reflection in the synthetic data. To counteract this effect, dynamic clipping is used in the following, which approximates the synthetic data to the real-world data using the following formula:

$$p' = \begin{cases} p - \alpha \cdot (p - T), & \text{if } p > T. \\ p, & \text{otherwise.} \end{cases} \tag{1}$$

Here, p represents pixel intensity, T is the threshold value, and α is the scaling factor. In the synthetic images, the intensity of reflections in sonar images ranges between 0 and 255. Our observations show that the brightest regions reach up to 250, while weak reflections are around 75. To correct over-strong reflections while preserving weak ones, we set $T = 100$. Additionally, to balance reflections from bright and weak regions while maintaining the integrity of features inside the image, we chose $\alpha = 0.5$.

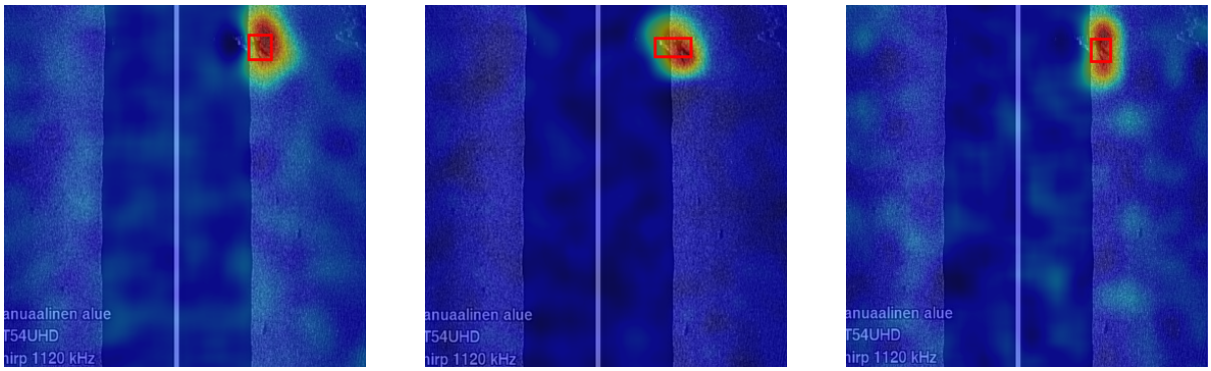
4.5. Evaluating on improved data

The dynamic clipping method is applied to the synthetic generated data and two new datasets (SimWave and Aaltonen + SimWave) are introduced and the training is repeated. The evaluation of the four new

Table 3
Performance after clipping trained models

Yolo-Modelle	Dataset	Precision	Recall	mAP50	mAP50-95
Yolov8	Aaltonen	0.937	0.808	0.901	0.561
Yolov8	SimWave	0.878	0.704	0.763	0.391
Yolov8	SimWave (clipping)	0.788	0.592	0.682	0.401
Yolov8	Aaltonen + SimWave	0.943	0.799	0.907	0.627
Yolov8	Aaltonen + SimWave (clipping)	**0.963**	0.823	0.935	0.631
Yolov11	Aaltonen	0.859	0.881	0.923	0.621
Yolov11	SimWave	0.814	0.712	0.773	0.387
Yolov11	SimWave (clipping)	0.829	0.619	0.728	0.361
Yolov11	Aaltonen + SimWave	0.925	0.816	0.901	0.626
Yolov11	Aaltonen + SimWave (clipping)	0.898	**0.921**	**0.946**	**0.648**

models and the previous results are shown in Table 3 exceeding the previous models on all metrics. YOLOv8 demonstrated higher precision by effectively filtering out false positives, while YOLOv11 improved recall by detecting more people. These results suggest that the clipping method reduces the domain gap and enhances the model's generalization ability for the hybrid dataset. The improvement in detection is also to be underlaid by the saliency maps.

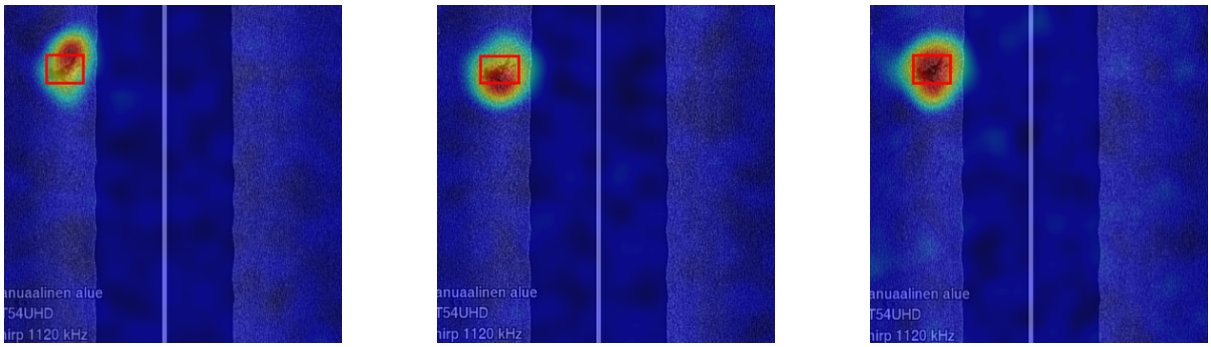


Aaltonen | SimWave (clipping) | Aaltonen + SimWave (Clipping)

Figure 3: x-AI with DRISE on different models with Hybrid-clipping

The Figure 3 shows the saliency maps on the same image as before in Figure 2, but with the last two images trained on the improved data. The model trained on the clipped hybrid dataset is now able to detect the object in the upper right corner, as the detection rate increased. It is also worth mentioning the difference between the models trained on the synthetic data set. The attention of for the object changes from focusing on both the reflection and shadow in Figure 2 to align more with the saliency map trained on real-world data (Aaltonen) in Figure 3.

To further emphasize this point, the difference in the attention trained on real-world and synthetic data is of most interest. The different saliency maps of another example are shown in Figure 4. In this example, the model trained on real-world data focuses on the legs, while the model trained on the hybrid dataset focuses on another region. After approximating the synthetic data to the real-world data by applying the proposed dynamic clipping process, the saliency map focuses even better on the object. This improvement confirms our assumption that the performance of the model was further improved by the selected data adjustment procedure, derived from an XAI methodology.



Aaltonen | Aaltonen + SimWave | Aaltonen + SimWave (clipping)

Figure 4: Effect of Clipping on Person Detection

5. Conclusion and future work

This study investigates the integration of real and simulated data to improve human detection in side-scan sonar images. Due to the limitations of real sonar datasets, we developed a simulation-based approach called SimWave to generate synthetic side-scan images and evaluated its impact on real-world data. Our study demonstrates how XAI methods can be effectively used to improve synthetically generated sonar data. The experimental results show that an object detection can perform better when trained on real-world and synthetic generated data. With the help of an XAI method of individual objects, an assumption could be made about the difference between synthetic and real images. This proposal could be used to improve the data provided by the simulation and improve the domain generalization to enhance the detection. Further qualitative analyses showed that the model trained with clipped data successfully identified people that the hybrid model without clipping and the model trained with real data failed to detect. The testing conducted in this paper is based on a small dataset consisting of 125 side-scan sonar scans. For a solid investigation of our assumption, a larger dataset is required. In the field of search and rescue, a larger dataset is currently not available, but the simulation could be transferred to a similar problem with larger datasets at hand. In future work, the real world data set will be enlarged in order to be able to train a better model. In addition, the simulated sonar images can be improved by using articulated 3D human models to create more realistic scenarios. With the model agnostic explanation method used other object detection algorithms like Detection Transformer (DETR) or Region-based Convolutional Neural Networks (R-CNN) can be tested as well. All developments are intended to contribute to the development of a robust and generalizable object detection algorithm based on side-scan sonar to assist human search and recovery operations.

Acknowledgments

This work was accomplished within the project KI-S, FKZ
03DPS1124A, funded by the German Federal Ministry of Education and Research

Declaration on Generative AI

The author has not employed any Generative AI tools.

References

[1] R. F. Becker, S. H. Nordby, J. Jon, Underwater forensic investigation, CRC Press, 2013. doi:https://doi.org/10.1201/b14765.

[2] J. Rutledge, W. Yuan, J. Wu, S. Freed, A. Lewis, Z. Wood, T. Gambin, C. Clark, Intelligent shipwreck search using autonomous underwater vehicles, in: 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2018, pp. 6175–6182.
[3] A. Ruffell, Lacustrine flow (divers, side scan sonar, hydrogeology, water penetrating radar) used to understand the location of a drowned person, Journal of hydrology 513 (2014) 164–168.
[4] Y. Z. Nga, Z. Rymansaib, A. Anthony Treloar, A. Hunter, Automated recognition of submerged body-like objects in sonar images using convolutional neural networks, Remote Sensing 16 (2024) 4036.
[5] Y. Lu, J. Zhang, Q. Chen, C. Xu, M. Irfan, Z. Chen, Aquayolo: Enhancing yolov8 for accurate underwater object detection for sonar images, Journal of Marine Science and Engineering 13 (2025) 73.
[6] F. Zhang, W. Zhang, C. Cheng, X. Hou, C. Cao, Detection of small objects in side-scan sonar images using an enhanced yolov7-based approach, Journal of Marine Science and Engineering 11 (2023) 2155.
[7] S. R. Richter, V. Vineet, S. Roth, V. Koltun, Playing for data: Ground truth from computer games, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14, Springer, 2016, pp. 102–118.
[8] B. Kiefer, D. Ott, A. Zell, Leveraging synthetic data in object detection on unmanned aerial vehicles, in: 2022 26th international conference on pattern recognition (ICPR), IEEE, 2022, pp. 3564–3571.
[9] N. Mital, S. Malzard, R. Walters, C. M. De Melo, R. Rao, V. Nockles, Improving object detection by modifying synthetic data with explainable ai, arXiv preprint arXiv:2412.01477 (2024).
[10] S. Lee, Deep learning of submerged body images from 2d sonar sensor based on convolutional neural network, 2017 IEEE Underwater Technology (UT) (2017) 1–3.
[11] X. Wen, J. Wang, C. Cheng, F. Zhang, G. Pan, Underwater side-scan sonar target detection: Yolov7 model combined with attention mechanism and scaling factor, Remote. Sens. 16 (2024) 2492.
[12] X. Yuan, J. Li, W. Wang, X. Zhou, N. Li, C. Yu, Improved yolov9 for underwater side scan sonar target detection, The Computer Journal (2024).
[13] S. Hu, T. Liu, Underwater rescue target detection based on acoustic images, Sensors (Basel, Switzerland) 24 (2024).
[14] B. Kiefer, D. Ott, A. Zell, Leveraging synthetic data in object detection on unmanned aerial vehicles, 2022 26th International Conference on Pattern Recognition (ICPR) (2021) 3564–3571.
[15] A. Agrawal, A. Sikdar, R. Makam, S. Sundaram, S. K. Besai, M. Gopi, Syn2real domain generalization for underwater mine-like object detection using side-scan sonar, ArXiv abs/2410.12953 (2024).
[16] K. Basha, A. Nambiar, S3simulator: A benchmarking side scan sonar simulator dataset for underwater image analysis, ArXiv abs/2408.12833 (2024).
[17] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, International Journal of Computer Vision 128 (2016) 336 – 359.
[18] V. Petsiuk, R. Jain, V. Manjunatha, V. I. Morariu, A. Mehra, V. Ordonez, K. Saenko, Black-box explanation of object detectors via saliency maps, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 11438–11447.
[19] S. Macenski, T. Foote, B. Gerkey, C. Lalancette, W. Woodall, Robot operating system 2: Design, architecture, and uses in the wild, Science Robotics 7 (2022) eabm6074. URL: https://www.science.org/doi/abs/10.1126/scirobotics.abm6074. doi:10.1126/scirobotics.abm6074.
[20] W.-S. Choi, D. R. Olson, D. Davis, M. Zhang, A. Racson, B. Bingham, M. McCarrin, C. Vogt, J. Herman, Physics-based modelling and simulation of multibeam echosounder perception for autonomous underwater manipulation. frontiers inrobotics and ai8 (2021), 279, 2021.
[21] T. Aaltonen, Consumer class side scanning sonar dataset for human detection, 2023 46th MIPRO ICT and Electronics Convention (MIPRO) (2023) 1161–1166.
[22] G. Jocher, A. Chaurasia, J. Qiu, Ultralytics yolov8, 2023. URL: https://github.com/ultralytics/ultralytics.
[23] G. Jocher, J. Qiu, Ultralytics yolo11, 2024. URL: https://github.com/ultralytics/ultralytics.