Towards Robust Object Detection in Underwater Sonar Imagery: A Cross-Modality Transfer Learning Study Leveraging RGB, FLS, and SAS Data

Franziska Auer
TKMS ATLAS ELEKTRONIK GmbH
Bremen, Germany
franziska.auer@tkmsgroup.com

Tobias Meisen
Bergische Universität Wuppertal
Wuppertal, Germany
meisen@uni-wuppertal.de

Heeraj Ayyappan
TKMS ATLAS ELEKTRONIK GmbH
Bremen, Germany
heeraj.ayyappan@tkmsgroup.com

Abstract

Robust underwater object detection is challenged by the complex acoustic nature of sonar imagery and the scarcity of labeled data. This research investigates cross-modality transfer learning to address these issues. Given the scarcity of labeled synthetic aperture sonar (SAS) images, we explore the potential of leveraging the larger and more readily available labeled datasets from RGB and forward-looking sonar (FLS) to improve object detection performance in SAS images. We systematically analyze the impact of sensor differences, data augmentation techniques, and dataset mixing strategies. Our results demonstrate that pre-training on RGB datasets can significantly enhance SAS detection performance, particularly when utilizing convolutional neural networks. However, direct transfer learning is ineffective without sonar-specific adaptation, and combining datasets from disparate sonar sensors can be detrimental due to feature inconsistencies. These findings highlight the need for tailored approaches to data processing and model selection in sonar image analysis, paving the way for more robust and efficient underwater object detection systems.

1. Introduction

Detecting concealed underwater items (camouflaged infrastructure or unexploded ordnances (UXOs)) is challenging under low visibility. Sonar, though common, produces signal-based images that differ markedly from RGB/optical data. Moreover, publicly available seabed-object datasets are small [2] and lack the full variety of mine-like contact (MILCO) shapes, sizes, and burial configurations encountered in real-world scenarios.

This research explores cross-modality transfer learning to address these challenges. We hypothesize that leveraging data from readily available modalities like RGB (e.g., MS COCO 328k images [17]) and forward-looking sonar (FLS) (e.g., Marine Debris 1869 images [33]) can significantly improve object detection performance in the data-scarce synthetic aperture sonar (SAS) domain (e.g., ITMINEX 613 images [19]) by exploiting shared feature representations. FLS provides real-time detection with variable resolution, while SAS offers high-resolution imagery over larger areas, but at the cost of processing time.

This study systematically investigates the interplay between sensing modality, data variability (dynamic range of SAS imagery) and model generalization, directly addressing three key challenges: adapting to the sonar-RGB domain gap, optimizing image processing and augmentation for sonar data, and leveraging combined datasets with sensor awareness. We hypothesize that knowledge transfer from RGB-pretrained models will be limited by the differences between RGB images and sonar images, and that targeted preprocessing and augmentation can mitigate these differences. To validate this, we will analyze these limitations, explore the impact of appropriate techniques, and evaluate strategies for effectively combining different FLS and SAS datasets by accounting for their respective sensor characteristics, expecting that this will lead to improved object detection performance. The key challenges addressed in this paper are:

- Adapting to the Sonar-RGB Domain Gap: We explore the limitations of transferring knowledge from RGB-pretrained models to sonar imagery and explore methods to mitigate this gap.
- Optimizing Image Processing & Augmentation for Sonar:

Paper	Costum Name	Basemodel	Adaptation	Dataset	Sensor	Year
[40]	ScEMA-YOLOv8	YOLOv8 [30]	implemented cross-channel attention mechanism	URPC-2021 [31]	FLS	2024
[3]	YOLOX-ViT	YOLOX [11]	ViT layer added after backbone	SWDD [3]	SSS	2024
[39]	YOLOSonar	YOLOv7 [34]	backbone replaced by attention mechanism	Marine Debris [33]	FLS	2025
[18]	AquaYOLO	YOLOv8 [30]	context-aware feature selection added in neck	UATD [38] & Marine Debris [33]	FLS	2025

Table 1. Overview recent adaptations of CNNs and ViTs to the sonar domain

We investigate how preprocessing (dynamic range compression) and data augmentation can enhance robustness and detection performance, while accounting for potential biases introduced by human-centric image enhancement.

- Leveraging Combined Datasets with Sensor Awareness: We assess the increment of training data via FLS/SAS combination while carefully considering sensor-specific characteristics to ensure performance gains.

The paper is organized as follows: Sec. 2 reviews existing literature. Sec. 3 details our experimental setup, including a description of the models and the datasets used. Sec. 4 presents the outcomes of our experiments: Zero-shot transfer, fine-tuning, augmentation strategies, and dataset mixing. Sec. 5 analyzes these results, and Sec. 6 summarizes our key findings and future directions for research in underwater object detection.

2. Related Work

Our literature analysis, in line with prior surveys [9, 12, 20, 26], shows that sonar-based object detection has largely relied on proprietary datasets [6, 21, 25, 28, 37], which in turn fostered the development of individually trained, non-comparable computer vision (CV) models adapted to the sonar domain [13, 16, 27, 36]. Since 2023, however, we observe a notable shift toward publicly available datasets recorded with a FLS [3, 18, 39] and CV models [10, 16, 23], enabling better reproducibility, standardized benchmarking, and collaborative progress. Our work examines now how we can use this shift to improve detection results on SAS data.

CV Models for Sonar Image Detection

Recent works have adapted You Only Look Once (YOLO) variants to improve sonar image detection [3, 18, 39, 40]. An overview of these adaptations for FLS and sidescan sonar (SSS) is given in Tab. 1.

Two approaches adapted YOLOv8 to improve feature extraction: ScEMA-YOLOv8 [40] added another feature extraction layer and more connections to enhance the detection of small targets. The additional EMA mechanism is used to counteract the loss of feature information introduced by the additional detection layer. Working towards the same goal, AquaYOLO [18] replaced conventional convolutional neural network (CNN) layers with residual blocks to capture fine details in noisy data. It further integrates dynamic feature aggregation to reduce redundancy and enhance feature correlation. In addition, context-aware feature selection combines adjacent feature levels, which improves object localization accuracy.

Similarly, YOLOX-vision transformer (ViT) [3] introduced ViT layers between the backbone and neck of YOLOX [11], enhancing feature extraction. This increased the mean average precision $(\text{mAP})_{50}$ from 0.18 to 0.20 for YOLOX-L and more than doubled it for YOLOX-Nano (0.19 to 0.42). In addition, the use of knowledge distillation reduced model size while lowering false positives in SSS images by up to 20%.

YOLO-SONAR [39], an adaptation of YOLOv7, integrates competitive coordinate attention and spatial group-enhanced attention to suppress seabed clutter and strengthen semantic–spatial feature extraction. Further, a context feature extraction module improves the detection of small objects, while the Wise-IoUv3 loss function addresses class imbalance in sonar datasets.

Most approaches adapt the backbone of a CV model to enhance feature extraction. Among these, YOLOX-ViT [3] is the only work with publicly available code, allowing us to directly compare its performance against unadapted baseline models. Unfortunately, for the other adaptations, the descriptions lacked sufficient detail for reliable reimplementation, and our requests for code went unanswered, preventing their inclusion in this work.

Datasets

As most models in Tab. 1 rely on FLS data, we begin by reviewing publicly available FLS datasets to contextualize recent model adaptations and results.

There are three public datasets using FLS images: the Marine Debris Dataset [33], the UATD dataset [38], and the UXO dataset [8]. The Marine Debris Dataset [33], pre-

Model name	Backbone	Year	Million parameter	GFLOPs	Estimated size
YOLOv7	E-ELAN	2023	6 - 36.9	13 - 104.7	24 - 148 MB
YOLOv8	CSPDarknet53	2023	11.8 - 45.9	42.4 - 220.1	47.2 - 183.6 MB
YOLOXViT	CSPDarknet-53 + ViT layer	2024	15 - 110	40 - 400	60 - 440 MB

Table 2. Overview models used in this study

sented in 2021, offers 1,868 FLS images. It is focused on marine debris segmentation with eleven object classes plus a background class. This dataset has been used in a few studies [35], [39], [18] focusing on underwater object classification and detection tasks. A year later the UATD dataset [38] was published. It contains over 9,000 Multibeam FLS images featuring ten categories of target objects for underwater acoustic target detection. In contrast to the other mentioned datasets, the UXO dataset [8] focuses on munition instead of everyday items.

In the SAS domain, 89% of the datasets used in the cited papers are proprietary. However, [2] introduced a small multi-sensor dataset with 86 SAS and 82 optical images of the same objects, including manta mines, cylinders, and natural objects. A. Abu and R. Diamant obtained an average of 88% true positives with their proposed multi-modal object classifier.

3. Cross-Modality Study Setup

Building upon the trends identified in Sec. 2, this section details the experimental configuration employed in our cross-modality study. We first give an overview of the models and datasets used in this study and then detail the experiments conducted to access transferability between domains.

Applied Models

As established baselines in sonar adaptation, we include YOLOX-ViT [3] with publicly available code together with the base models YOLOv7 [34] and YOLOv8 [30]. Their specific characteristics are given in Tab. 2.

Within the YOLO family, we use YOLOv7 and YOLOv8 as real-time baselines with complementary design choices. YOLOv7 (E-ELAN and advanced training strategies) offers an excellent speed–accuracy trade-off for object detection, whereas YOLOv8 adopts an anchor-free head, multi-scale prediction, and an updated backbone, typically yielding higher accuracy and faster inference than YOLOv7 and YOLOX [30]. As discussed in Sec. 2, recent work proposes sonar-specific YOLO adaptations. However, because reproducible code is unavailable for them, we use the standard YOLOv8 alongside YOLOv7 as real-time baselines instead in our study. Based on Xie et al. [39] and Lu et al. [18] showing that both models perform well with FLS data and we expect that they will do the same with SAS data.

Datasets

To address the research challenges outlined in Sec. 1, we combine public datasets that enable reproducibility (Marine Debris, UXO, SASOptical) with the proprietary ITMINEX dataset, which, better than the public datasets, reflects realistic operational conditions (adjustable preprocessing & targets obtained in the ocean). This selection balances comparability across studies with the need for representative high-resolution sonar data.

The Marine Debris dataset [33] available at [32], offers a diverse set of objects. Fig. 1 shows man-made objects (a tire and a bottle). To test combining different datasets, we add the UXO dataset [8] available at [7] to this study. The three objects within the UXO dataset are more challenging to distinguish than the items in the Marine Debris dataset due to their similar shapes (Fig. 2). Both datasets where acquired with the same sensor (ARIS Explorer 3000 [24]). Thus, we hypothesize that combining them will increase model training success.

The ITMINEX dataset [19] serves as the main SAS dataset, due to its representative nature of the SAS imagery found in our application domain of detecting intricate objects blending into its underwater environments. It has several varified high-resolution targets presented within the dataset. An example SAS image showing a manta mine is given in Fig. 3. The ITMINEX dataset features cylinder mines and manta mines, as well as other man-made objects and clutter. Based on common practice [4, 5, 15, 21, 22, 28] the dataset features two annotation classes: MILCO and non-MILCO. For consistency across datasets, ITMINEX images were converted to PNG format and cropped into 640x640 patches (standard input size of YOLOv7 and YOLOv8). To ensure that each object, including its shadow, appears fully within at least one patch, we applied a 100-pixel (2.5 m) overlap. While this can duplicate objects across patches, their differing positions effectively act as translation augmentation. We further refined the dataset to primarily include patches containing objects, making it more comparable to public datasets.

To provide a public reference for the SAS domain, we also evaluate on SASOptical dataset [2] available at [1], which pairs SAS with optical imagery of similar objects (manta/cylinder mines and boulders). For consistency, we remap its labels to the two-class scheme used in ITMINEX

Dataset name	Sonar Type	Number of Images	Objects viewed	Objects types	Resolution	Sonar Used
Marine Debris [32]	FLS	1869	2364	man-made objects, clutter	2.3 mm - 10 cm	ARIS Explorer 3000 [24]
UXO [7]	FLS	1500	1 per image	3 different UXOs	not given	ARIS Explorer 3000 [24]
ITMINEX [19]	SAS	613	808	cylinder & manta mines, man-made objects, clutter	25 mm	Vision SAS Mk1 [29]
SASOptical [1]	SAS	57	72	cylinder & manta mines, clutter	not given	Kraken SAS[14]

Table 3. Overview datasets used in this study

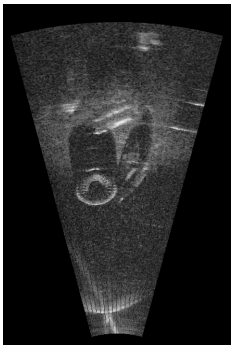

Figure 1. Example Marine Debris dataset [32]

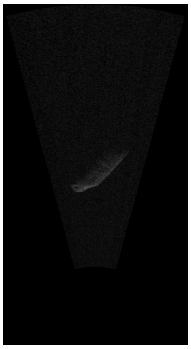

Figure 2. Example UXO dataset [7]

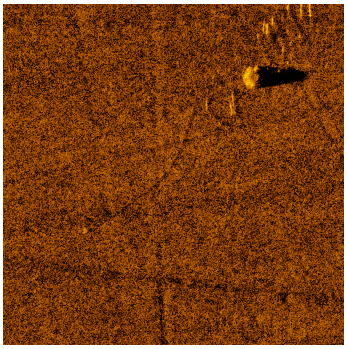

Figure 3. Example ITMINEX dataset [19]

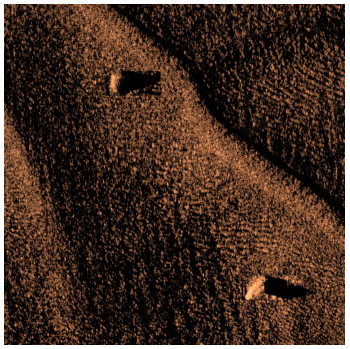

Figure 4. Example SASOptical dataset [1]

(MILCO/non-MILCO). Tab. 3 shows the differences and similarities between the four datasets.

The examples in Fig. 1–4 highlight key differences between the two sensor types. In SAS images (Fig. 3, 4), shadows consistently appear on one side of the object, whereas in FLS images the shadow position depends on object placement (Fig. 1) or may vanish if the object is elevated above the seabed (Fig. 2). FLS imagery also differs in geometry. Captured forward-looking, it produces narrow images in polar coordinates which were than transferred to Cartesian coordinates. By contrast, SAS imagery is side-looking. During the preprocessing, the original long images are cropped into smaller patches suitable for GPU processing. All datasets were subsequently split into training (70%), validation (20%), and test (10%) sets.

Software Setup

All experiments ran in Python 3.9 with PyTorch (2.5.1/2.7.0), torchvision (0.20.1/0.22.0), and Ultralytics (8.3.134). Hardware specifications are listed in Tab. 4.

Storage capacity	512 GB
GPU Memory	46 GB
Processor model	AMD EPYC 9554
Graphics card model	NVIDIA L40S

Table 4. Our hardware characteristics

Preliminary Dataset Size Analysis

To initially assess the influence of dataset size on detection performance, we conducted experiments using the largest available dataset, UXO, and the YOLOXViT_S model (see Tab. 5). We observed a clear correlation between dataset size and mAP_{50}. Increasing the training data from 1,000 images to 5,000 images resulted in an improvement of 1.41%. As expected, further increasing the dataset size to 10,000 images yielded an additional, albeit smaller, gain of 0.3%.

However, this performance improvement came at a significant cost in training time. While training with 1,000 images took only 4.5 hours, training with 10,000 images required 1.96 days. This suggests a diminishing return on investment in terms of training time for each additional increment in dataset size, highlighting the need for strategies to maximize the effectiveness of limited data resources and

Dataset size	mAP_{50}	Training time
1,000	97.86	4.5 hours
5,000	99.27	23.0 hours
10,000	99.57	1.96 days

Table 5. Influence of dataset size of detection results (UXO dataset, YOLOXViT_S)

motivating the subsequent investigation into dataset mixing techniques.

Experiments

To tackle the key challenges listed in Sec. 1, we carried out four experiments: transfer learning (to test domain-adaptation), data augmentation (to assess its impact in the sonar domain), and dataset-mixing strategies for sonar image analysis (to expand the training set).

a) Zero-shot transfer We investigated the potential of zero-shot transfer learning by testing several pretrained models directly on SAS imagery, without any subsequent fine-tuning. This included models pretrained on the MS COCO dataset, as well as a YOLOXViT checkpoint specifically trained on SSS data. Furthermore, we also evaluated a YOLOv7 model trained by ourselves on FLS data to assess transferability within the sonar domain on similar objects. This approach aims to determine the extent to which knowledge gained from visible light imagery, or from another sonar modality, can generalize to our target application. Initial expectations suggest limited benefit from MS COCO pre-training on SAS data, as the fundamental image features defining objects in sonar (highlight-shadow patterns) differ significantly from those used in standard computer vision tasks centered around color and texture. However, we hypothesize that the SSS-pretrained YOLOXViT model and the FLS-trained YOLOv7 model may demonstrate improved performance due to the greater similarity within the sonar domains.

b) Fine-tuning To explore the advantages of transfer learning more fully, we compared two training strategies. The first involves training computer vision models from scratch using both FLS and SAS datasets. The second utilizes pretrained weights from MS COCO and then fine-tunes the model specifically on FLS and SAS data. We anticipate positive effects for FLS imagery, as Marine Debris images share characteristics with RGB night scenes present in the MS COCO dataset, potentially reducing the domain gap. However, for SAS imagery, we expect less benefit from MS COCO pre-training due to the fundamental differences in data characteristics.

c) Influence of augmentations Recognizing the importance of dataset variability and model robustness, we evaluated the impact of both augmentation techniques and varying dynamic range compression levels on model performance using SAS data. Given the unique characteristics of sonar imagery, standard augmentations like flips and rotations were excluded due to the preprocessing step of mirroring SAS images to port-side. Instead, we focused on hue saturation value (HSV) augmentation to reduce sensitivity to color scaling, and employed translations, scaling, and mosaic augmentations to account for variations in object position and size. Furthermore, we trained models on SAS images with and without dynamic range compression (-5 dB to +25 dB compared to no compression) to assess its effect on detection, mindful of how preprocessing choices intended for human interpretation may impact algorithmic outcomes. We hypothesize that these augmentations and optimized dynamic range settings will preserve the inherent characteristics of sonar data while improving detection performance as we see it in RGB images.

d) Effect of mixing Addressing the challenge of limited training data, we investigated the potential benefits of combining different datasets to increase both data volume and object diversity. A YOLOv8_L model was trained separately on the Marine Debris dataset, the UXO dataset, and a combined dataset (500 samples of each UXO class were randomly selected and added to the Marine Debris dataset), and the resulting mAP_{50} curves were compared. While we expect this mixing to improve overall performance, we anticipate less favorable results with SAS data. This stems from differences in sensor characteristics and resolutions between the ITMINEX and SASOptical datasets, particularly as only the ITMINEX dataset resolution is known, while both FLS datasets originate from the ARIS Explorer 3000.

4. Results

a) Zero-shot transfer As decribed in Sec. 3, we first focused on establishing a public baseline on the SASOptical dataset, with the goal of identifying a starting point for targeted fine-tuning using our internal SAS data. However, off-the-shelf transfer learning proved ineffective. A YOLOv7 model pretrained on MS COCO failed to produce any valid detections. While a YOLOXViT variant, pretrained on a related SSS data (initialized with wall-detection weights), showed some activity, the results were largely unreliable, yielding pervasive false positives by detecting objects larger than the image patch in 55 out of 57 test images. This indicated that simply leveraging weights trained on SSS data, even when related to underwater environments, was insufficient for accurate detection on SASOptical.

Interestingly, a contrasting result emerged when evaluating a YOLOv7_tiny model trained on FLS data. As shown in Fig. 5, this model successfully detected one MILCO correctly, with no false positives observed. Surprisingly, the successful detection of a MILCO on SAS data from an FLS-trained model, with no false positives, indicates that some level of transfer learning is possible between sonar sensors. While we will not directly pursue further optimization of this FLS-to-SAS transfer, this insight motivates our exploration of fine-tuning RGB-trained models as a promising avenue for SAS object detection.

Figure 5. YOLOv7_tiny model trained with FLS images, tested on SAS images from the SASOptical dataset

b) Fine-tuning Informed by the observed potential for transfer learning, and guided by our hypothesis that MS COCO pre-training would yield greater gains on FLS imagery than SAS imagery, we proceeded to fine-tune models utilizing weights pretrained on this large-scale RGB dataset. Notably, however, the results revealed the opposite trend: pre-training on MS COCO yielded more substantial improvements for the SAS dataset, challenging our initial assumption regarding domain similarity.

The resulting mAP_{50} scores on the validation set are presented in Tab. 6. On the FLS dataset (Marine Debris), the performance gains from MS COCO pre-training were generally small ($< 0.5\%$). YOLOXViT_S showed the largest improvement on this dataset (0.41%).

YOLO version	FLS - Marine Debris		SAS - ITMINEX	
	scratch	pretrained MS COCO	scratch	pretrained MS COCO
7_L	99.40%	99.60%	35.00%	95.76%
7_tiny	99.20%	99.40%	55.50%	**99.50%**
8_L	99.30%	99.00%	**98.80%**	99.20%
8_S	99.20%	99.00%	89.77%	98.44%
XViT_L	**99.70%**	99.40%	95.78%	94.56%
XViT_S	99.30%	**99.71%**	94.55%	83.07%

Table 6. mAP_{50} of the validation set

For the SAS dataset (ITMINEX), pre-training on MS COCO consistently resulted in substantial performance improvements. Particularly YOLOv7_tiny experienced a dramatic increase from 55.50% (scratch) to 99.50% (pretrained). While it is consistent with observations in other CV domains, its impact on SAS imagery is particularly significant given the expected limitations imposed by the domain gap. Similarly, YOLOv8_S improved from 89.77% (scratch) to 98.44% (pretrained), and YOLOv7_L from 35.00% to 95.76%. These results clearly demonstrate the value of leveraging pretrained representations from the MS COCO dataset to overcome the challenges associated with learning from SAS imagery, and contradict our initial expectation of limited benefit. This suggests that SAS data shares more visual commonalities with natural images in MS COCO than previously assumed, indicating a potential underestimation of transfer learning's effectiveness in sonar.

It is worth noting that YOLOXViT_S surprisingly performed worse with pre-training on SAS data, indicating a potential overfitting to the pretrained weights or a mismatch in feature representation between the two datasets.

c) Influence of augmentations We evaluated augmentation techniques on YOLOXViT_S performance using SAS data (ITMINEX) to improve robustness. HSV augmentation and translation augmentation both yielded significant positive effects, increasing the mAP_{50} compared to training without augmentations (Tab. 7). Specifically, using only translation augmentation resulted in an mAP_{50} of 98.99%, while HSV augmentation achieved 98.74%. Combining both HSV and translation yielded an mAP_{50} of 95.69%, suggesting a potential for diminishing returns when applying multiple augmentations simultaneously.

Augmentation Strategy	mAP_{50}	Epochs	Time Elapsed
No Augmentations	94.07%	300	2.754 hr
Only HSV	98.74%	300	2.668 hr
Only Translation	**98.99%**	300	2.550 hr
HSV and Translations	95.69%	300	3.195 hr
All Augmentations (Mosaic off at end)	94.55%	300	4.160 hr

Table 7. Performance Comparison of Different Augmentation Strategies (ITMINEX dataset, YOLOXViT_S)

However, the most significant impact was observed with mosaic augmentation, although its effect was complex. As shown in Fig. 6, while initially beneficial, mosaic augmentation ultimately impacted performance. To mitigate this, we implemented a strategy of disabling mosaic augmentation after 250 epochs. This approach, using all augmentations but turning off mosaic for the last 50 epochs, resulted

in a final mAP_{50} of 94.55%.

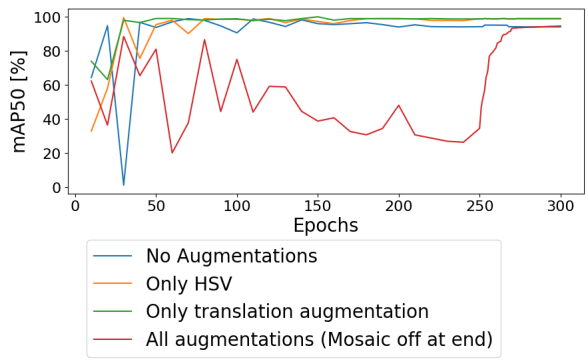

Figure 6. Influence of augmentations (ITMINEX dataset, YOLOXViT_S)

These results demonstrate a nuanced relationship between augmentation strategies and model performance. While HSV and translation consistently improved performance with minimal overhead, the application of mosaic augmentation should only be done with careful consideration, as improvements were marginal.

Furthermore, we investigated the impact of dynamic range compression on detection performance. Training without limiting the dynamic range yielded an mAP_{50} of 95.76%, while limiting the dynamic range to support the human eye resulted in a slightly reduced mAP_{50} of 95.12%.

d) Effect of mixing Motivated by our initial findings (partial transferability from FLS to SAS demonstrated earlier) we investigated whether combining datasets could mitigate the performance limitations of smaller datasets.

Combining the Marine Debris and UXO datasets (Tab. 8) yielded a marginal reduction in mAP_{50} (-0.52%). Despite the introduction of three additional classes to distinguish, the model maintained a performance level that can be considered positive, indicating that the increased dataset size effectively counterbalances the added complexity. This suggests that combining datasets containing objects with differing representations, but recorded by the same sensor (FLS in this case), is a valuable strategy for improving model robustness and generalization. Interestingly, training on the individual datasets converged rapidly (after only 300 epochs for Marine Debris and 200 for UXO) while training on the mixed dataset required the full 500 epochs, highlighting the increased learning demands of a more complex dataset.

In contrast to the FLS results, combining the ITMINEX and SASOptical datasets had a markedly negative impact on performance, as illustrated in Tab. 9. Replacing 10% of the data with images from SASOptical caused a substantial drop in mAP_{50} by almost 50%. This significant decrease is likely due to resolution differences between the two SAS datasets, posing a challenge for consistent feature extraction. Furthermore, all training runs, regardless of dataset composition, required the full 300 epochs to complete.

Dataset	mAP_{50}	Early Finish
Marine Debris	99.30%	after 300 epochs
UXO	99.45%	after 200 epochs
mixed datasets	98.78%	no, full 500 epochs

Table 8. Effect of mixing FLS datasets (YOLOv8_L)

Dataset	mAP_{50}	Early Finish
ITMINEX	98.80%	no, full 300 epochs
SASOptical	4.23%	no, full 300 epochs
mixed datasets	48.84%	no, full 300 epochs

Table 9. Effect of mixing SAS datasets (ITMINEX dataset, YOLOv8_L)

5. Discussion

a) Zero-shot transfer Our initial investigation into zero-shot transfer learning directly addresses the challenge of ”Adapting to the Sonar-RGB Domain Gap”. The complete failure of a YOLOv7 model pretrained on MS COCO to detect any objects on the SASOptical dataset starkly illustrates the limitations of simply transferring knowledge learned from natural images to the fundamentally different feature representations present in sonar imagery (highlight-shadow patterns versus color and texture). While a YOLOXViT model pretrained on wall-detection SSS data showed some activity, the high rate of false positives further reinforces this gap. Notably, the successful detection of a MILCO by a YOLOv7_tiny model trained on FLS data demonstrates a promising avenue. Transfer learning within the sonar domain for similar objects can be surprisingly effective, suggesting shared feature characteristics between different sonar modalities. This emphasizes the potential for exploiting knowledge from one sonar sensor to improve performance on another, potentially reducing the need for extensive task-specific training data.

b) Fine-tuning The fine-tuning experiments further contributes to the challenge of ”Adapting to the Sonar-RGB Domain Gap”, showing that while direct transfer from RGB struggles, pre-training on RGB can be extremely effective as a starting point when combined with further training on sonar data. The substantial performance gains on the SAS ITMINEX dataset after fine-tuning on MS

COCO-pretrained models highlight the value of leveraging the broad feature representations learned from large RGB datasets. However, a key observation was the differing performance between traditional convolutional YOLO networks (YOLOv7, YOLOv8) and the transformer-based YOLOXViT. The convolutional architectures generally benefited more consistently from pre-training, while the YOLOXViT models showed more sensitivity to the weights and even experienced performance decreases, likely due to the architectural differences in how they process image features. This underscores the importance of carefully selecting the model architecture suited to the specific data domain and transfer learning strategy.

c) Influence of augmentations This component of our work directly tackles the challenge of "Optimizing Image Processing & Augmentation for Sonar". The positive impact of HSV and translation augmentations on SAS data demonstrates the efficacy of targeted data manipulation techniques in enhancing robustness, mitigating the impact of minor shifts and variations in target position. The initially beneficial, but ultimately detrimental, effect of mosaic augmentation highlights the need for careful consideration of augmentation strategies and emphasizes the potential for introducing human-centric biases. Specifically, while mosaic augmentation can increase data diversity, it can also distort the inherent spatial relationships within sonar images, leading to reduced performance. Our findings regarding dynamic range compression reveal a nuanced trade-off: while compression aids human interpretability by presenting imagery more akin to natural viewing conditions, it might introduce a minor compromise in algorithmic performance, as it alters the fundamental signal characteristics upon which the detection algorithm relies. However, dynamic range compression also facilitates faster initial convergence during training, potentially reducing overall training time.

d) Effect of mixing Our exploration of dataset mixing strategies directly addresses the challenge of "Leveraging Combined Datasets with Sensor Awareness". The positive results from combining FLS datasets (Marine Debris and UXO) demonstrate the value of increased data volume within a single sensor modality. However, the substantial performance drop observed when mixing ITMINEX and SASOptical datasets underscores the critical importance of considering sensor-specific characteristics. The differing resolutions between these SAS datasets likely introduced inconsistencies that hindered model learning. This reinforces the need for careful evaluation and potentially pre-processing to mitigate discrepancies when combining data from various sources. The results clearly show that simply increasing dataset size without considering data quality and sensor-specific attributes can lead to detrimental outcomes.

Our results demonstrate that detection performance in SAS images can be significantly improved through transfer learning, with viable pathways including both directly transferring knowledge from FLS data and leveraging pre-trained weights from large-scale RGB datasets when employing CNNs, while ViTs exhibited greater sensitivity and reduced gains. Targeted data augmentation, particularly HSV adjustments and translations, consistently improved robustness, but caution is warranted with techniques like mosaic augmentation, which can introduce artificial biases. Finally, dataset mixing proved beneficial when combining data from the same sonar sensor modality, yet requires careful consideration and potentially pre-processing, when integrating data from different sonar systems to mitigate inconsistencies.

6. Conclusion

This work investigated the challenges of applying deep learning to sonar image analysis, focusing on adapting models from the RGB domain, optimizing data processing, and leveraging combined datasets. We found that while direct transfer from RGB is limited, pre-training can be beneficial when combined with sonar-specific fine-tuning. In our study CNNs proved more adaptable than transformers, likely due to the specific domain and the small dataset used for fine-tuning . Carefully designed augmentations improved robustness, but required nuanced application. Combining datasets from different sonar types proved detrimental due to feature inconsistencies.

These findings highlight the need for sonar-specific architectures, tailored data augmentation (HSV: 4.67% mAP_{50} increase, translation: 4.92% mAP_{50} increase) and strategies to effectively combine data from heterogeneous sensors (-50% mAP_{50} for combining SAS dataset, while marginal mAP_{50} reduction combining FLS datasets). Bridging the gap between computer vision and underwater acoustics requires continued research into sensor-specific model architectures and data augmentation techniques tailored for the unique challenges of sonar data.

Acknowledgment

This research was supported by the Bundesministerium für Wirtschaft und Energie (BMWE) within the IRAV project (Industrielle Räumung von Altlasten in Verklappungsgebieten). We gratefully acknowledge the CMRE for the opportunity to participate in sea trials on board the Alliance. Furthermore, generative AI tools were utilized to assist with code development and text drafting. All content has been critically reviewed, edited, and remains the sole responsibility of the authors.

References

[1] Avi Abu. Database2. https://drive.mathworks.com/sharing/f32d2544-4c4e-4574-a396-6abf8522f812/, 2021. Accessed on April 25, 2025. 3, 4

[2] Avi Abu and Roee Diamant. Underwater object classification combining sas and transferred optical-to-sas imagery. *Pattern Recognition*, 2023. 1, 3

[3] Martin Aubard, László Antal, Ana Madureira, and Erika Ábrahám. Knowledge distillation in yolox-vit for side-scan sonar object detection. *arXiv*, 2024. 2, 3

[4] Fleur Bouwman, David W. Ecclestone, Alexander L. Gabriëlse, and Alexander M. van Oers. Synthetic side-scan sonar data for detecting mine-like contacts. In *Artificial Intelligence for Security and Defence Applications II*, 2024. 3

[5] Abdesselam Bouzerdoum, Philip B Chapple, Mark Dras, Yi Guo, Len Hamey, Tahereh Hassanzadeh, Thanh Hoang Le, Omid Mohamad Nezami, MA Orgun, Son Lam Phung, et al. Improved deep learning-based classification of mine-like contacts in sonar images from autonomous underwater vehicles. In *UACE2019-Conference Proceedings*, 2019. 3

[6] Jesper Haahr Christensen, Lars Valdemar Mogensen, and Ole Ravn. Side-scan sonar imaging: Automatic boulder identification. In *OCEANS 2021: San Diego – Porto*, 2021. 2

[7] Nikolas Dahn, Miguel Bande Firvida, Proneet Sharma, Leif Christensen, Oliver Geisle, Jochen Mohrmann, Torsten Frey, Prithvi Kumar Sanghamreddy, and Frank Kirchner. An acoustic and optical dataset for the perception of underwater unexploded ordnance (uxo). https://zenodo.org/records/11068045, 2024. Accessed on November 11, 2024. 3, 4

[8] Nikolas Dahn, Miguel Bande Firvida, Proneet Sharma, Leif Christensen, Oliver Geisle, Jochen Mohrmann, Torsten Frey, Prithvi Kumar Sanghamreddy, and Frank Kirchner. An acoustic and optical dataset for the perception of underwater unexploded ordnance (uxo). In *OCEANS 2024 - Halifax*, 2024. 2, 3

[9] Lucas C. F. Domingos, Paulo E. Santos, Danilo G. S. Barros, and Eduardo Todt. A survey of underwater acoustic data classification methods using deep learning for shoreline surveillance. *Sensors*, 2022. 2

[10] Xing Du, Yongfu Sun, Yupeng Song, Lifeng Dong, and Xiaolong Zhao. Revealing the potential of deep learning for detecting submarine pipelines in side-scan sonar images: An investigation of pre-training datasets. *Remote Sensing*, 2023. 2

[11] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 2

[12] Stanisław Hożyń. A review of underwater mine detection and classification in sonar imagery. *Electronics*, 2021. 2

[13] Wanzeng Kong, Jiajie Hong, Minghao Jia, Jinling Yao, Wei Cong, Hao Hu, and Hongying Zhang. Yolov3-dpfin: A dual-path feature fusion neural network for robust real-time sonar target detection. *IEEE Sensors Journal*, 2019. 2

[14] Kraken Robotics. MINSAS Synthetic Aperture Sonar. https://www.krakenrobotics.com/products/synthetic-aperture-sonar/. Accessed on April 17, 2025. 4

[15] Joseph T. Kuhner, Roger W. Meredith, and Casey C. Taylor. Automated contact calling visual aid using sequential mathematical processes: Using textural analysis for mine-like contact detection. In *2012 Oceans*, 2012. 3

[16] Frederik Kühne, Bastian Kaulen, Christian Kanarski, Finn Röhrdanz, Karoline Gussow, and Gerhard Schmidt. Detektion und klassifikation von objekten aus von sonar-systemen erstellten plots mithilfe von künstlicher intelligenz. In *DAGA 2024*, 2024. 2

[17] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014. 1

[18] Yanyang Lu, Jingjing Zhang, Qinglang Chen, Chengjun Xu, Muhammad Irfan, and Zhe Chen. Aquayolo: Enhancing yolov8 for accurate underwater object detection for sonar images. *Journal of Marine Science and Engineering*, 2025. 2, 3

[19] Vincent Myers, Johannes Groen, Holger Schmaljohann, Isabelle Quidu, and Benoit Zerr. Multi-look processing for coherent change detection with synthetic aperture sonar. In *UACE2017-4th Underwater Acoustics Conference and Exhibition*, 2017. 1, 3, 4

[20] Dhiraj Neupane and Jongwon Seok. A review on deep learning-based approaches for automatic sonar target recognition. *Electronics*, 2020. 2

[21] Narcís Palomeras, Thomas Furfaro, David P. Williams, Marc Carreras, and Samantha Dugelay. Automatic target recognition for mine countermeasure missions using forward-looking sonar data. *IEEE Journal of Oceanic Engineering*, 2022. 2, 3

[22] Nuno Pessanha Santos, Ricardo Moura, Gonçalo Sampaio Torgal, Victor Lobo, and Miguel de Castro Neto. Side-scan sonar imaging data of underwater vehicles for mine detection. *Data in Brief*, 2024. 3

[23] Advaith Sethuraman, Anja Sheppard, Onur Bagoren, Christopher Pinnow, Jamey Anderson, Timothy Havens, and Katherine Skinner. Machine learning for shipwreck segmentation from side scan sonar imagery: Dataset and benchmark. *Journal: International Journal of Robotics Research*, 2024. 2

[24] Sound Metrics. Aris explorer 3000. http://www.soundmetrics.com/products/aris-sonars/aris-explorer-3000. Accessed on April 17, 2025. 3, 4

[25] Yannik Steiniger, Dieter Kraus, and Tobias Meisen. Generating synthetic sidescan sonar snippets using transfer-learning in generative adversarial networks. *Journal of Marine Science and Engineering*, 2021. 2

[26] Yannik Steiniger, Dieter Kraus, and Tobias Meisen. Survey on deep learning based computer vision for sonar imagery. *Engineering Applications of Artificial Intelligence*, 2022. 2

[27] Zhaohui Sun, Jianhu Zhang, Zhenyu Zhang, Xin Ren, Jianping Xie, and Xiang Xiao. Dp-vit: A dual-path vision trans-

former for real-time sonar target detection. *Remote Sensing*, 2022. 2

[28] Olga Lopera Tellez. Human-in-the-loop for autonomous underwater threat recognition. In *OCEANS 2018 MTS/IEEE Charleston*, 2018. 2, 3

[29] TKMS ATLAS ELEKTRONIK. TMKS ATLAS ELEKTRONIK UK awarded contract to deliver minehunting autonomous underwater vehicles to the Royal Navy. https://www.atlas-elektronik.com/newsroom/article/atlas-elektronik-uk-awarded-contract-to-deliver-minehunting-autonomous-underwater-vehicles-to-the-royal-navy. Accessed on April 17, 2025. 4

[30] Ultralytics. Yolov8: Real-time object detection and segmentation. https://github.com/ultralytics/ultralytics, 2023. Accessed on April 10, 2025. 2, 3

[31] URPC2021. Underwater robot professional competition 2021 dataset. https://www.urpc.org/, 2021. Accessed on April 10, 2025. 2

[32] Matias Valdenegro and Deepak Singh. Marine-debris datasets. https://github.com/mvaldenegro/marine-debris-fls-datasets/, 2021. Accessed on April 23, 2025. 3, 4

[33] Matias Valdenegro and Deepak Singh. The marine debris dataset for forward-looking sonar semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021. 1, 2, 3

[34] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3

[35] Lin Wang, Xiufen Ye, Liqiang Zhu, Weijie Wu, Jianguo Zhang, Huiming Xing, and Chao Hu. When sam meets sonar images. *IEEE Geoscience and Remote Sensing Letters*, 2024. 3

[36] Qi Wang, Yixiao Zhang, and Bo He. Intelligent marine survey: Lightweight multi-scale attention adaptive segmentation framework for underwater target detection of auv. *IEEE Transactions on Automation Science and Engineering*, 2024. 2

[37] David P. Williams. On the utility of multiple sonar imaging bands for underwater object recognition. In *OCEANS 2022, Hampton Roads*, 2022. 2

[38] Kaibing Xie, Jian Yang, and Kang Qiu. A dataset with multibeam forward-looking sonar for underwater object detection. *Scientific Data*, 2022. 2, 3

[39] Jian Yang, Kaibin Xie, and Kang Qiu. Yolo-sonar: Semantic-spatial feature guided yolo for object detection in forward-looking sonar images. *Frontiers in Marine Science*, 2025. 2, 3

[40] Linhan Zheng, Tao Hu, and Jin Zhu. Underwater sonar target detection based on improved scema yolov8. *IEEE Geoscience and Remote Sensing Letters*, 2024. 2