International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 1 | pp. 01-12 | Jan-Mar 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 08 Nov 2025 | Revised: 23 Jan 2026 | Accepted: 01 Feb 2026 | Published: 08 Feb 2026

A Hybrid VGG-ResNet Feature Fusion Network for Object Detection in Side-Scan Sonar Images

[1]Venkata Lakshmi Keerthi K, [2]Vijayalakshmi P, [3]Rajendran V

[1,2,3] Department of Electronics and Communication Engineering, Vels Institute of Science, Technology, and Advanced Studies (VISTAS), Chennai, India.
[1]keerthireddy1123@gmail.com, ORCID ID: 0009-0009-4822-5122

Abstract: Side-scan sonar images present significant challenges for object detection due to high noise levels, limited spatial resolution, and complex seabed structures. These factors, along with speckle noise and acoustic shadowing effects, further complicate reliable target detection. Nonetheless, object detection is extremely important in many areas, including marine archaeology, underwater search and rescue, mine countermeasures operations, and the inspection of critical national infrastructure, to improve safety and operational efficiency. In this work, a hybrid convolutional neural network (CNN) is presented for object detection in challenging side-scan sonar imagery. The proposed model combines the complementary feature-extraction capabilities of VGG-16 and ResNet-50 via feature fusion to improve target discrimination in sonar images. Transfer learning from ImageNet-pretrained backbones is employed to address data sparsity and improve model generalization on sonar datasets. Experimental evaluation demonstrates that the proposed model achieves an overall classification accuracy of 84.2% and a mean Average Precision (mAP) of 88.35%, outperforming several existing methods. The results pave the way to enhance the efficacy and accuracy of underwater surveys, search-and-rescue missions, and seabed mapping.

Keywords: feature fusion, VGG-16, ResNet-50, hybrid CNN, Side-scan sonar, underwater object detection.

1 INTRODUCTION

Object detection in side-scan sonar (SSS) images is important across a variety of underwater fields, including but not limited to marine archaeology and shipwreck studies, mine countermeasures, and the inspection of underwater infrastructure [1, 2]. The absorption and scattering of visible light in marine environments make optical methods useless. With sonar, acoustic waves can be emitted to penetrate deep into the water and accurately map the seabed and submerged objects, often better than optical methods. Therefore, the greater the precision and speed of detecting objects in the sonar dataset, the greater the safety and operational effectiveness. This improves the efficiency of resource scoping, environment monitoring, and humanitarian response activities are conducted [3].

Despite their importance, reliable object detection in side-scan sonar images remains challenging. Sonar imaging systems inherently produce noisy, low-resolution images that are often cluttered with seabed artefacts, such as shadowing and geometric distortions. The underdevelopment of automated target recognition is caused by a lack of annotated sonar data, variability in sonar setup, environment, and objects, as well as data-collection orientation. Although rule-based systems have been applied in sonar image analysis, their adaptability to varying environmental conditions and unseen target appearances is limited.

Over the past few years, advancements in deep learning, particularly the use of Convolutional Neural Networks (CNNs), have revolutionised computer vision and achieved great success in a variety of imaging tasks. These models have learned features from data hierarchically and directly, rather than using traditional methods that relied on sonar image-processing techniques [4]. This shift in methodology has led to notable improvements in object detection in underwater environments, providing reliable methods for object recognition in challenging underwater settings [5]. Motivated by the impact of deep neural networks, researchers have begun developing new architectures to improve the discriminative power and robustness of object-detection systems for side-scan sonar images.

Recognising objects in SSS images is essential for many underwater tasks; however, it remains an extremely complex undertaking. The sonar datasets are challenging because, in addition to various types of artefacts, signal-to-noise ratios are often low, and speckle noise, shadow artefacts, and low lateral and azimuthal resolution introduce interference and attenuation. Additionally, while many factors can create counterfeit signatures, true signal signatures can be attenuated, while artefacts remain prevalent and occur in large volumes. The lack of abundance, poor coverage, and poorly documented SSS collections (pooled with the random nature of sonar backscatter) cause machine-learning algorithms to overfit to the limited variation within the available SSS collections. The SSS tasks of beam patterning, sediment, and water column are difficult to parameterise, and still cause real-time changes in the target appearance.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 1 | pp. 01-12 | Jan-Mar 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 08 Nov 2025 | Revised: 23 Jan 2026 | Accepted: 01 Feb 2026 | Published: 08 Feb 2026

As a result, detection models are often unable to identify low-complex target signatures, a gap in autonomous underwater vehicle capabilities. In many practical side-scan sonar datasets, each image contains a single dominant target, allowing the detection task to be formulated as joint target classification and localisation at the image level. Because deep learning systems can create sophisticated representations from raw data, the need for the rigorous hand-crafting of features, as required by traditional methods, is removed. This is especially useful when little domain expert knowledge is available to define features, as is the case with high-dimensional, highly variable datasets. This ability is one of the reasons these systems excel at object detection, a prerequisite for safe and efficient navigation and manipulation in future robotic systems [6]. In addition, these methods have proven successful in difficult situations characterised by significant occlusions, changes in lighting, and large-scale changes, thereby greatly accelerating the field of computer vision. The contributions of the current research are outlined below.

- Development of a hybrid CNN architecture that integrates VGG-16 and ResNet-50 feature representations for side-scan sonar object detection.
- Enhanced feature discrimination through the fusion of hierarchical and deep residual representations in sonar imagery.
- Application of transfer learning to mitigate data sparsity and improve generalization on limited sonar datasets.
- Comparative evaluation against existing deep learning–based detection methods on side-scan sonar imagery.

The rest of the manuscript is structured as follows: Section 2 reviews the relevant literature in detail, including classical methods, deep learning, and combinations of these approaches for sonar image object detection. Section 3 describes the proposed methodology, including the construction of the datasets, the data preprocessing technique employed, the structure of the original hybrid deep neural network, the loss function, and the training process. Section 4 describes the results of the experiments, including metrics of the network's performance, an assessment against some of the most advanced techniques in the literature, and selected illustrative examples. Lastly, Section 5 concludes the paper, summarises the main contributions of the work, and presents some directions for future work.

2 Related Work

2.1. Traditional Object Detection Methods in Sonar Images

Before the widespread adoption of deep learning, object detection in SSS imagery was primarily achieved using traditional methods of sonar image processing and pattern recognition. These methods often followed a pipeline that included preprocessing the image to remove noise and artefacts, then segmentation, and finally classification or matching [7][8]. Early approaches relied on feature-based methods, where potential targets were represented using manually designed statistical, texture, or geometric descriptors [9][4]. Another prominent category was template matching, which involved correlating a known set of objects (i.e., templates) with an area of interest in the sonar image to detect a similar object. While effective for known predictable targets, the method did not adapt well to changes in object orientation and scale or to the inherent distortion of sonar imagery [10].

In the past, techniques have included thresholding and other traditional methods that attempt to separate potential objects from the background by varying intensity levels. This often includes adaptive thresholding, region growing, or a snake contour to define the object's boundaries. After segmentation, standard methods such as Support Vector Machines (SVM) [11] or K-Nearest Neighbors (KNN) [12] were used to classify the resulting features into predefined object classes or as clutter. However, when applied to complex side-scan sonar imagery, these methods were not robust across varying environmental conditions. The major challenges were the data's high vulnerability to background noise and operational noise (random noise), the difficulty of creating a stable, generalized set of features that perform reliably in a variety of situations and environments, and the limited capability of the used methods to analyze complex, multi-layered, non-linear, and intricate structures present in the sonar mosaics.

2.2. Deep Learning-Based Object Detection in Sonar Images

The advent of deep learning, especially Convolutional Neural Networks (CNNs), has revolutionized computer vision, and this revolution has now extended to the highly specialized field of sonar image analysis. The initial use of CNNs on sonar datasets focused on classification and demonstrated their ability to automatically learn, layer, and abstract features from raw pixel data, a process that greatly surpassed earlier manual feature-engineering attempts [13]. Motivated by this, researchers began tackling the more complex challenge of object detection, given CNNs' unique ability to localize and classify multiple objects, thereby enhancing the utility of sonar images for real-time operational use.

Sonar data began to be processed using popular deep learning models in computer vision for natural images once the field had matured. Such models are generally categorized as two-stage or single-stage. The former employs a region proposal network (RPN) to yield a sparse proposal set of candidate object positions, and a second stage further refines them for classification. Faster R-CNN has often been adopted as a reference framework due to its localisation accuracy. Its use in sonar imagery commonly entails transferring knowledge from a model pre-trained on a large optical dataset (e.g., ImageNet) and then fine-tuning it on a sparse sonar dataset [5] to achieve exceptional localisation accuracy for several underwater targets.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 1 | pp. 01-12 | Jan-Mar 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 08 Nov 2025 | Revised: 23 Jan 2026 | Accepted: 01 Feb 2026 | Published: 08 Feb 2026

Sonar applications include R-FCN (Region-based Fully Convolutional Networks) and Mask R-CNN. While R-FCN focused on improving efficiency through computation region-coupling, Mask R-CNN enhanced object detection in two ways: by adding instance segmentation to provide pixel-level masks for detected objects, which is beneficial for specific characterization of objects in sonar [14]. The two-stage models' intricate architecture and prolonged inference times, however, are inhibitive for many real-time use cases. Instead of bounding boxes and class probabilities for each part of the image, one-stage detectors produce bounding box and class probability predictions for the entire image in a single pass. This leads to faster detection, which is critical for operational speed in underwater detectors. This is especially true for the various versions of the YOLO detectors (v3, v4, v5, v7, v8). YOLO detectors are particularly popular for their speed-accuracy trade-off. Most YOLO adaptations for sonar include modifying the anchor boxes to better suit the sonar object scale and annotating a specific sonar dataset for network training [15]. Another popular one-stage detector is the Single Shot MultiBox Detector (SSD).

The SSD architecture also aligns well with the varying object scales in the sonar image, as it enables detection across scales via a multi-scale feature map. Among sonar object detection, there is also growing interest in the RetinaNet architecture. This interest is most likely due to the introduction of Focal Loss to address the class imbalance between foreground and background, a common issue in sonar imagery. This makes the SSD architecture particularly applicable to sonar object detection within AUVs, as the RetinaNet framework also enables learning from sparse detections.

The persistence of obstacles remains a challenge in deep learning-based object recognition from sonar images. A major challenge is recognizing small objects, as the low resolution of sonar images fails to capture sufficient detail. Another vital challenge is the need for real-time processing, particularly for systems involved in autonomous navigation and intervention, as advanced model architecture may not meet the stringent latency requirements of these applications. Lastly, the problem of unrivalled dependence on the dataset remains: since sonar datasets with varying complexities, extensive annotations, and high resolution are difficult to obtain, deep learning models might fail to generalize and overfit, leading to poor performance on new input data [12, 13]. Existing deep learning approaches demonstrate improved detection performance in sonar imagery but remain sensitive to noise, data scarcity, and variability in target appearance.

2.3. Hybrid Deep Neural Network Models

The demonstrated strengths of individual deep learning architectures such as VGG, ResNet, and Inception have motivated the development of hybrid deep neural network models. This is because these architectures can learn features under different conditions. In VGG, for example, the addition of multiple convolutional layers allows the network to learn intricate fine-grained spatial details [16], whereby ResNet incorporates a novel residual model to overcome the vanishing gradient issue, allowing for the training of deep neural networks with the capability of learning more abstract and robust representation learning [17]. Besides, Inception models use layers with kernels of different convolutional sizes to learn features at different levels of abstraction simultaneously, improving the model's ability to capture multiple complex patterns [18]. Integrating distinct models with different architectures into a hybrid model deepens the feature space, thereby improving the overall performance of the predictive model.

A hybrid setup provides a more intricate representation of an input signal by integrating feature maps from multiple backbone networks. Such hybrid strategies can improve robustness and diversity of representation, particularly in challenging imaging conditions [19]. Feature fusion may be performed at early, intermediate, or late stages of the network, depending on the application requirements. This strategy may also incorporate early fusion (feature integration at initial layers), late fusion (feature integration at output layers), or, as some would say, intermediate fusion (feature integration at various levels of the network). While hybrid deep learning models have been studied for general computer vision problems, their use in sonar image object detection remains underdeveloped, though with great potential.

The challenges with sonar images, including high levels of noise, low resolution, and the distinct and varied appearance of objects, would certainly make these images a good fit for such multifaceted approaches. For example, recent studies have attempted to integrate CNNs with Transformer models for sonar image segmentation to capture both local convolutional and global representation features. Other studies have attempted to integrate models of different types, fusing sequence models, such as Bi-LSTMs, with feature-learning models, such as Restricted Boltzmann Machines, for underwater object detection. Within this context, combining VGG-16 and ResNet-50 feature representations offers a practical fusion strategy for capturing complementary spatial and semantic information in side-scan sonar imagery. This motivates the exploration of hybrid feature fusion for improving detection reliability in complex sonar environments.

3 METHODOLOGY

The proposed methodology frames the side-scan sonar target identification task as a single-object detection problem, in which each input image contains a single dominant target. The model jointly learns target classification and bounding-box localisation using a hybrid feature-extraction and fusion strategy.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 1 | pp. 01-12 | Jan-Mar 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 08 Nov 2025 | Revised: 23 Jan 2026 | Accepted: 01 Feb 2026 | Published: 08 Feb 2026

This methodology details the engineering of a hybrid convolutional neural network focused on resilient object detection in side-scan sonar images. There are three main parts to this initiative. The first involves creating a complete dataset and constructing a data preprocessing pipeline to prepare the sonar images for model training. Each image is annotated with a single bounding box tightly enclosing the dominant object and an associated class label. During this phase, we implement a range of preprocessing techniques to reduce sonar noise and compensate for the lower resolution and spatial stretching common to sonar data. These preprocessing steps help stabilize feature learning and improve localisation performance in low-contrast sonar imagery. The second phase focuses on creating a hybrid deep neural network model. This is the focus of the methodology, as shown in Fig. 1. This architecture integrates the complementary feature-extraction capabilities of two pre-trained CNN backbones, VGG-16 and ResNet-50, within a unified detection framework.

VGG-16 is effective in learning spatial features and is a widely used model in this domain. ResNet-50 also outperforms its competitors in learning deeper, more robust representations by adding so-called residual connections. Among these models, one of the more effective deep learning architectures is the Discriminative Feature Fusion (DFF) network, which will help create more complex and discriminative representations of submerged targets. The model will be subjected to a rigorous training and evaluation process and will utilise the best available loss functions and training methodologies to achieve elevated levels of accuracy and to provide validation of the model and establish the model's position relative to the present, empirical state-of-the-art models in the domain of sonar object detection.

The core idea behind hybrid models is that no single deep learning architecture is universally optimal for all tasks or data types. By combining different architectures, a hybrid model aims to exploit diverse feature-extraction capabilities, overcome individual limitations, improve robustness and generalisation, meet specific task requirements, and optimise transfer learning. VGG-16's strength is in hierarchical feature learning through uniform, small convolutional filters. ResNet-50's ability to train very deep networks effectively due to residual connections, addressing the vanishing gradient problem.

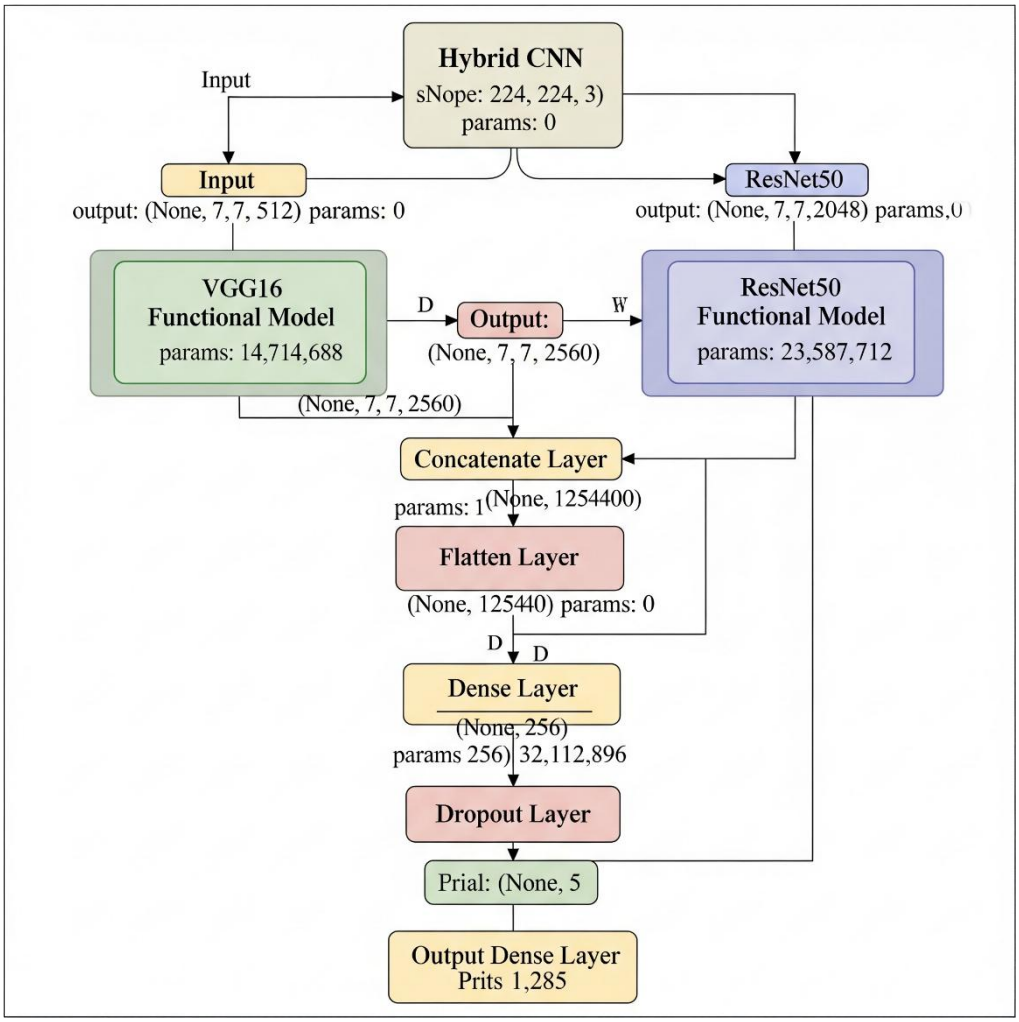

Fig. 1. The proposed hybrid CNN for object detection in side-scan sonar Images

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 1 | pp. 01-12 | Jan-Mar 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 08 Nov 2025 | Revised: 23 Jan 2026 | Accepted: 01 Feb 2026 | Published: 08 Feb 2026

The input layer serves as the model's entry point and defines the expected input dimensions. The (None) in the batch dimension signifies flexibility for any batch size. The (224, 224) indicates that the model is designed to process square images of 224x224 pixels, and (3) denotes that these are RGB colour images (Red, Green, Blue channels). Being a placeholder for input data, it contains no trainable parameters. The core of the hybrid approach lies in its parallel feature extraction, utilising two distinct and powerful pre-trained CNNs: VGG16 and ResNet50.

- VGG16 Branch (VGG16): This branch integrates the convolutional base of the VGG16 architecture. VGG16 is renowned for its uniform architecture with 3x3 convolutional filters, enabling efficient extraction of hierarchical and localised features. It processes the raw input image, downsampling it to 7x7 and generating 512 feature maps. The parameters are inherited from the pre-trained ImageNet weights, which can be frozen or fine-tuned depending on the specific application.
- ResNet50 Branch (ResNet50): Running in parallel to the VGG16 branch, this component incorporates the pre-trained ResNet50 architecture. ResNet50 is distinguished by its innovative "residual connections," which mitigate the vanishing gradient problem, allowing for the training of significantly deeper networks. It also processes the input image, producing a 7x7 feature map with 2048 channels. This suggests its ability to extract a richer, more diverse set of deep, abstract features. Its parameters are also derived from pre-trained ImageNet weights.

Following feature fusion, the shared feature representation is forwarded to two parallel output branches. The first branch performs target classification using fully connected layers and softmax activation, while the second branch predicts the target location's bounding box coordinates. This multi-task formulation enables simultaneous learning of object category and spatial localisation. The feature combination layer combines feature maps from branches trained on VGG16 and ResNet50. The concatenation is performed along the channel axis, resulting in a consolidated feature map with 2560 channels (512 channels from VGG16 and 2048 channels from ResNet50).

This fused representation serves as a shared feature space for both classification and localisation tasks. Since this operation is simply merging the feature maps, there are no trainable parameters. The intent is to leverage the feature-learning potential of both backbone networks to create a more robust, more elaborate feature representation. After the feature maps have been extracted and merged, this layer converts the 3D feature map with shape (None, 7, 7, 2560) into a 1D vector. This step is necessary to connect the convolutional section of the model to the fully connected (dense) layers that follow, which require a one-dimensional input. The layer performs no computations and is therefore parameter-less. The output size is computed to be $7\times7\times2560 = 125440$.

The classification head handles the final classification and interpretation of the retrieved features. This is the first fully connected layer, which takes in the 125440 flattened features. It consists of 256 neurons, each linked to every input feature. The substantial number of parameters $(125440+1)\times256=32,112,896$ indicates that it can learn more sophisticated, complex nonlinear relations within the integrated feature space. A Dropout layer is used to minimise overfitting. It is placed after the first dense layer. During training, this layer randomly deactivates a proportion of the input units, which in turn helps the network learn more robust features that do not rely heavily on any single neuron. This technique, known as regularisation, does not introduce any new trainable parameters. The output shape is the same as the input shape.

The output dense layer is the last layer of the proposed model. It contains 5 neurons. This shows that the model is tailored to handle a 5-class classification problem. The parameters in this case are given as $(256+1)\times5 = 1285$. For multi-class classification problems, it is standard to use a softmax activation at the output layer, yielding class probabilities that facilitate interpretation of the model's outputs. The hybrid CNN model leverages the distinct and complementary feature extraction capabilities of VGG16 and ResNet50 to enhance overall robustness. Running both architectures in parallel and merging feature maps yields a more detailed and diverse representation of the input image. After this stage, the merged features are forwarded to parallel fully connected layers for classification and bounding-box regression. This setup increases the model's accuracy by integrating diverse embeddings, thereby enhancing its discrimination power to tackle the intricate task of object detection. Model training is guided by a composite loss function that combines categorical cross-entropy for classification and Smooth L1 loss for bounding-box regression. This joint optimisation strategy balances classification accuracy and localisation precision during training. To reduce overfitting on the limited sonar dataset, transfer learning is employed, with the early layers frozen during initial training and selective fine-tuning thereafter.

4 Results and Discussion

Developing a hybrid CNN architecture for object detection requires a structured, systematic approach to ensure consistent, reliable results. The training process begins with the creation of a custom dataset containing 2,220 images, each with tightly fitted bounding boxes and classification labels to provide optimal object localisation. As the training hybrid CNN is assumed to combine a high-capacity backbone for hierarchical feature extraction with a task-specific output head designed to predict bounding boxes and class scores, this phase of training focuses on fine-tuning hyperparameters to reduce a defined task-specific composite loss.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 1 | pp. 01-12 | Jan-Mar 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 08 Nov 2025 | Revised: 23 Jan 2026 | Accepted: 01 Feb 2026 | Published: 08 Feb 2026

The continual updating of weights and biases is designed to build richer feature-space hierarchies and is necessary for the system to learn to identify and describe intricate spatial and spectral patterns required for dependable image localisation and object classification. The evaluation focuses on assessing both classification reliability and bounding box localisation accuracy for dominant targets in side-scan sonar images. After training, the model's generalization ability is evaluated on a separate, unseen test dataset of 1,486 images and their associated ground-truth segmentations. Each image is analyzed by the trained model, which assigns values to object category classes and bounding box parameters. Detection metrics: Precision, Recall, F1 score, and, most importantly, mAP (Mean Average Precision) are used to calculate the overall performance of the model. These metrics assess the model's ability to detect, locate, and classify entities from distributions different from those in the training data, confirming its robustness for real-world applications. The dataset is designed so that all 1,486 images are allocated for testing. This includes images from all categories shown in the image below, such as aerial and marine vehicles, some polygonal solids, and submerged debris (Fig. 2 and Fig. 3). All reported results are obtained exclusively on the unseen test set, with no overlap between the training and test sets.

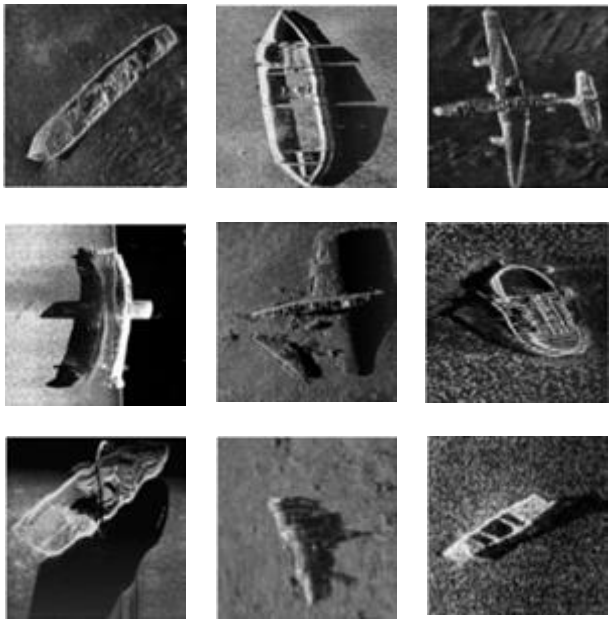

Fig. 2. The training dataset

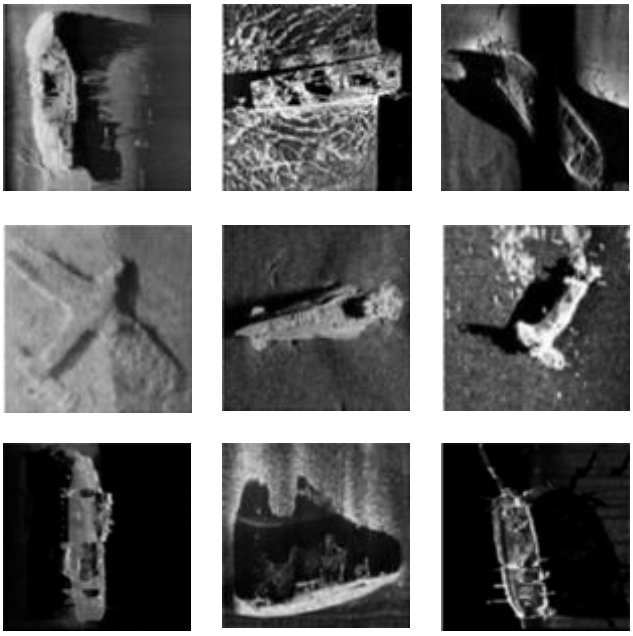

Fig. 3. The test dataset

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 1 | pp. 01-12 | Jan-Mar 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 08 Nov 2025 | Revised: 23 Jan 2026 | Accepted: 01 Feb 2026 | Published: 08 Feb 2026

4.1. Evaluation Metrics

Assessing object-detection algorithms requires metrics that evaluate both object classification and geometric localisation accuracy. Central to this evaluative framework is the Intersection over Union (IoU) metric, computed as the area of overlap between an algorithm-generated and a reference bounding box, normalized by the area of their union. The resulting ratio indicates localisation fidelity: IoU values approaching unity indicate that the generated box closely matches the true object contour. This metric underpins outcome classification by categorizing detections as true positives or false positives, thereby enabling the calculation of precision and recall, which inform the model's overall effectiveness.

A predefined IoU threshold, typically 0.5, determines whether a prediction constitutes a True Positive (TP) (correctly detected object with sufficient IoU), a False Positive (FP) (incorrectly localized detection), or a False Negative (FN) (a missed ground truth object). Unlike general classification, True Negatives are rarely emphasized in standard object-detection metrics because of the overwhelming number of background regions. In this study, Average Precision is computed at an IoU threshold of 0.5, and mAP is computed by averaging across target classes. Key performance indicators for object detection include Precision, which measures the accuracy of detections (TP/(TP+FP)), and Recall, which quantifies the completeness of detections (TP/(TP+FN)). The F1-score, as the harmonic means of Precision and Recall, provides a balanced assessment. The confusion matrix is computed at the image level based on the dominant detected object per image. True Negative values in Table 1 represent images where no target was incorrectly detected as the dominant object.

Table.1 The performance evaluation of the proposed technique using a confusion matrix

Performance Measure	Aircraft	Ship	Overall
Number of SSS images	906	580	1486
True Positive	678	352	1030
False Positive	97	123	220
False Negative	78	53	131
True Negative	53	52	105
Precision	0.897	0.87	0.888
Accuracy	0.856	0.819	0.842
F1 score	0.876	0.844	0.864

4.2. Comparative Analysis with State-of-the-Art

The analysis of the object detection model results focused on the performance on the subclassifications of Aircraft and Ship targets, as well as a combined total assessment. A total of 1486 side-scan sonar images were used in the analysis: 906 were classified as Aircraft and 580 as Ship. The model showed strong performance in detecting actual targets, achieving 678 True Positives (TP) for Aircraft and 352 TP for Ships, for a total of 1030 TP. However, the detection of 220 False Positives (FP) - 97 for Aircraft and 123 for Ships - indicates cases in which the model mistakenly identified non-targets as objects of interest. At the same time, the 131 False Negatives (FN) - 78 for Aircraft and 53 for Ships - indicate the actual targets which were not detected. The relatively low number of True Negatives (TN) - 53 for Aircraft, 52 for Ships, and a total of 105 - reflects the ongoing trend in object detection toward focusing on target localisation and eschewing background classification. The qualitative results are shown in Fig. 4, and the quantitative results are tabulated in Table 1.

The model achieved excellent precision, scoring 0.897 and 0.870 for Aircraft and Ships, respectively, and a mean score of 0.888. This indicates that a high proportion of the targets detected corresponds to true objects of interest. Although accuracy can be affected by class imbalance in detection tasks, the observed values indicate stable classification performance. The model's reliability is further reflected in the independent F1 scores for Aircraft and Ships of 0.876 and 0.844, respectively, yielding a mean score of 0.864, indicating a good balance between identification and correct object localisation for successful target detection.

The proposed model demonstrated competitive and consistent detection performance across both target classes. The steady improvement in Precision and F1 scores for single and combined-target classes demonstrates the model's proficiency at achieving a high true positive rate while maintaining a low false positive rate. The model's demonstrated effectiveness attests to its readiness for real-world applications of automated object detection in complex, cluttered underwater sonar environments. The comparative results for existing methods are reported from their respective publications, evaluated on comparable side-scan sonar detection tasks.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 1 | pp. 01-12 | Jan-Mar 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 08 Nov 2025 | Revised: 23 Jan 2026 | Accepted: 01 Feb 2026 | Published: 08 Feb 2026

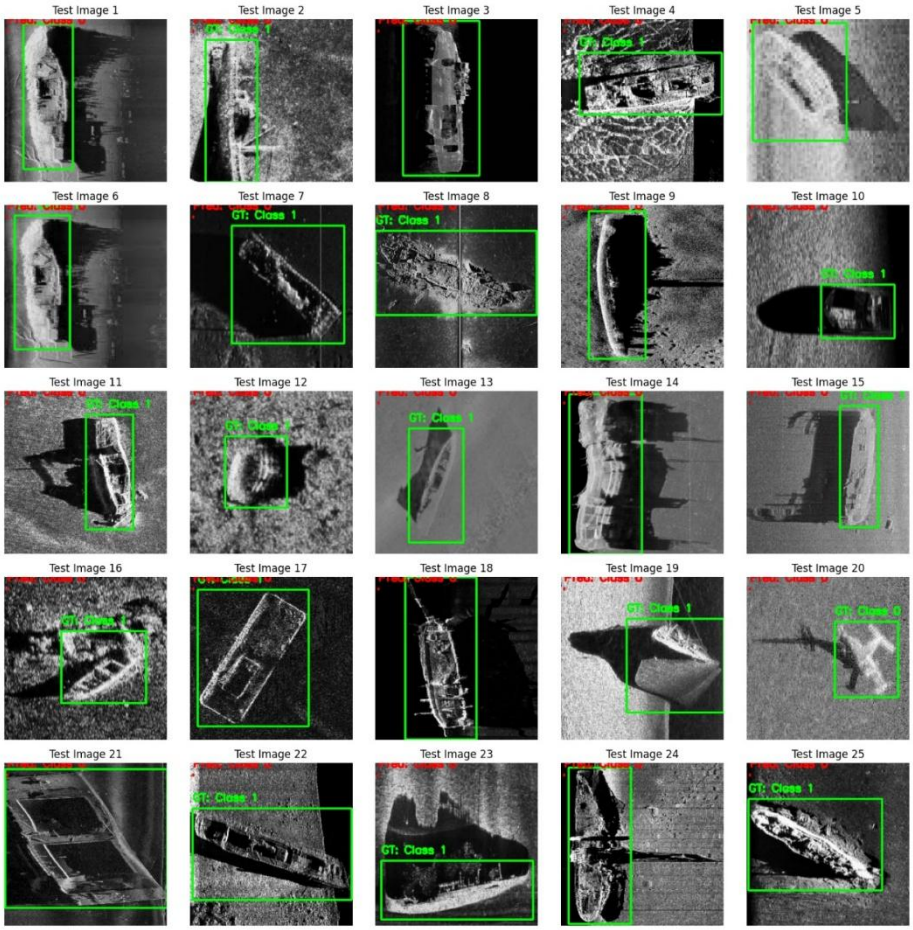

Fig. 4. The proposed technique's object detection qualitative results

Table 2. The proposed method's performance evaluation with the state-of-the-art techniques

Method	Precision		Average Precision
	Aircraft	Ship	
Mask R-CNN [20]	70.92	78.5	74.71
MS R-CNN [21]	72.3	81.4	76.85
SOLO [22]	76.9	77.9	77.4
YOLOACT [23]	78.8	82.3	80.55
LW R-CNN [24]	81.9	82.9	82.4
Proposed	89.7	87	88.35

The performance comparison is summarised quantitatively in Table 2, demonstrating that the proposed framework achieves notable improvements over several state-of-the-art deep learning models, including Mask R-CNN, MS R-CNN, SOLO, YOLOACT, and LW R-CNN. The evaluation of precision performance with standard methods is illustrated in Fig. 5. In the Aircraft detection subclass, the model achieves an Average Precision (AP) score of 89.7, the highest among all models, with the next highest being LW R-CNN at 81.9. Being 7.8 percentage points higher than the next-best approach demonstrates the model's increased effectiveness in accurately detecting and spatially localising aircraft in sonar images. Fig. 6 shows the improvement in accuracy over the course of training, particularly for the proposed method. Fig. 7 examines the relationship between aircraft detection and ship detection performance. Most methods fall below the diagonal, indicating that detecting ships is easier than detecting aircraft. The proposed method is the only exception that specialises in detecting aircraft and also detects ships well.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 1 | pp. 01-12 | Jan-Mar 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 08 Nov 2025 | Revised: 23 Jan 2026 | Accepted: 01 Feb 2026 | Published: 08 Feb 2026

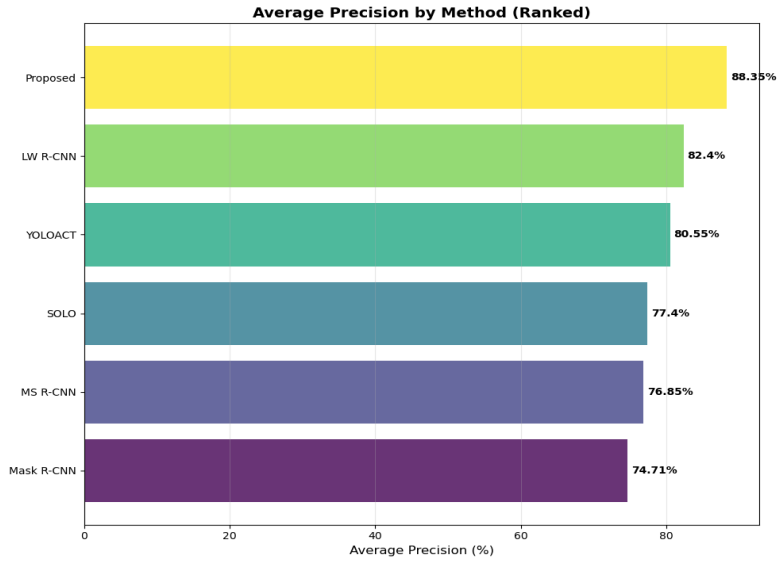

Fig. 5. The proposed technique's precision comparison with the state-of-the-art techniques

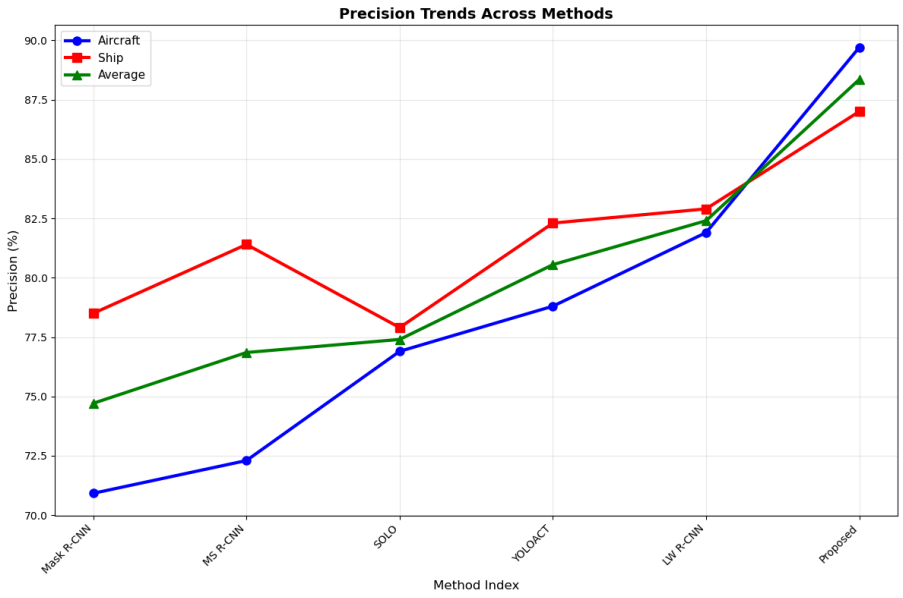

Fig. 6. Line plot showing precision trends and progression across methods

The proposed method demonstrated competitive performance, with an average precision (AP) of 87 in ship detection. Even though LW R-CNN achieves an AP of 82.9 and YOLOACT achieves 82.3, the proposed method remains competitive with Mask R-CNN, MS R-CNN, and SOLO. The value of 87 for ships in the row labelled "Proposed" is consistent with the previously stated precision value for ships (0.870), indicating a consistently competitive position in that category.

The proposed model achieves a mean Average Precision (mAP@0.5) of 88.35 on the test dataset. This score is better than all of the baselines, including the ones in the lead, which are LW R-CNN (82.4), YOLOACT (80.55), and MS R-CNN (76.85). The mAP of 88.35 shows that the proposed hybrid CNN has successfully achieved strong performance on sparse categories (Aircraft and Ship) and translated that into significant overall detection performance. This substantial enhancement means that the model's architecture, which is very likely to feature parallel, integrated feature extraction with effective fusion, is a major contributor to the model's performance in complex sonar environments. The qualitative results illustrate accurate localisation of dominant targets despite background clutter and acoustic shadowing.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 1 | pp. 01-12 | Jan-Mar 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 08 Nov 2025 | Revised: 23 Jan 2026 | Accepted: 01 Feb 2026 | Published: 08 Feb 2026

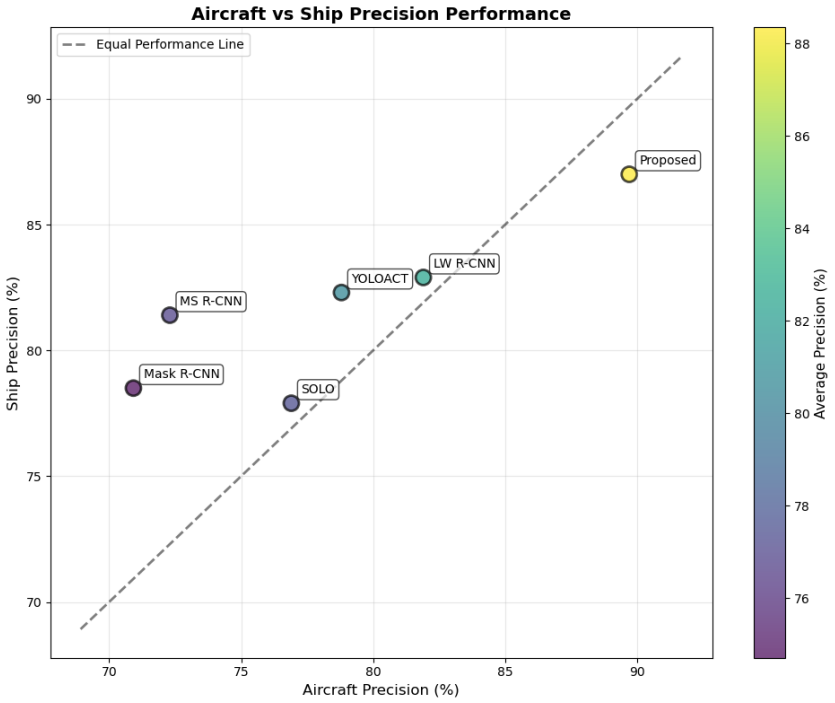

Fig. 7. Scatter plot comparing aircraft vs ship detection precision

While the present study focuses on side-scan sonar images containing a single dominant target and evaluates performance primarily in terms of detection accuracy, future work will extend the framework to multi-object scenarios, broader target categories, and real-time operational constraints.

5 Conclusions

This paper presents a hybrid CNN model for object detection in side-scan sonar imagery, addressing challenges such as noise, low resolution, and complex seabed backgrounds. The proposed model employs transfer learning to address data sparsity and improve feature representation in sonar imagery. This hybrid architecture improves model performance in detecting tight clusters of diverse and intricate seabed structures, thereby supporting reliable target identification in maritime operations. The proposed model is thoroughly and rigorously tested. The experimental results indicate that the proposed model achieves competitive performance compared to several existing methods, with an overall accuracy of 84.2% and a mAP of 88.35%. These results confirm the efficacy of the model's hierarchical features and its ability to operate in high-noise environments and to detect and localize objects, such as aircraft and ships, in the acoustic clutter encountered in offshore and deep-sea surveys. The extensive design process and analysis of the hybrid CNN model indicate the potential to enhance operational efficiency, reliability, and safety for a wide range of underwater activities. Improvements in marine archaeology, humanitarian search and rescue, mine countermeasures, and submerged infrastructure inspections require accurate, timely object recognition to achieve mission success and optimize resource allocation. Future work will explore architectural refinements, evaluate the feasibility of real-time deployment, and extend the model to a broader range of underwater targets and environmental conditions.

FUNDING INFORMATION

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ETHICS STATEMENT

This study did not involve human or animal subjects and, therefore, did not require ethical approval.

STATEMENT OF CONFLICT OF INTERESTS

The authors declare no conflicts of interest related to this study.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 1 | pp. 01-12 | Jan-Mar 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 08 Nov 2025 | Revised: 23 Jan 2026 | Accepted: 01 Feb 2026 | Published: 08 Feb 2026

REFERENCES

[1] X. Wen, C. Cheng, L. Li, F. Zhang, and G. Pan, "A framework for super-resolution of side-scan sonar images: Combination of variational Bayes and regional feature selection," *Engineering Applications of Artificial Intelligence*, vol. 155, p. 111007, May 2025, doi: 10.1016/j.engappai.2025.111007.

[2] S. Jiao, F. Xu, and H. Guo, "Side-Scan sonar image detection of shipwrecks based on CSC-YOLO algorithm," *Computers, Materials & Continua/Computers, Materials & Continua (Print)*, vol. 82, no. 2, pp. 3019–3044, Dec. 2024, doi: 10.32604/cmc.2024.057192.

[3] C. Zhang, Y. Chai, X. Wang, and Y. Chen, "Enhancing sonar image quality for underwater object recognition in marine environments: A review," *Ocean Engineering*, vol. 336, p. 121862, Jun. 2025, doi: 10.1016/j.oceaneng.2025.121862.

[4] U. Anitha, S. Malarkkan, G. D. A. Jebaselvi, and R. Narmadha, "Sonar image segmentation and quality assessment using prominent image processing techniques," *Applied Acoustics*, vol. 148, pp. 300–307, Jan. 2019, doi: 10.1016/j.apacoust.2018.12.038.

[5] H. F. Tolie, J. Ren, R. Chen, H. Zhao, and E. Elyan, "Blind sonar image quality assessment via machine learning: Leveraging micro- and macro-scale texture and contour features in the wavelet domain," *Engineering Applications of Artificial Intelligence*, vol. 141, p. 109730, Dec. 2024, doi: 10.1016/j.engappai.2024.109730.

[6] M. Soori, B. Arezoo, and R. Dastres, "Artificial intelligence, machine learning and deep learning in advanced robotics, a review," *Cognitive Robotics*, vol. 3, pp. 54–70, Jan. 2023, doi: 10.1016/j.cogr.2023.04.001.

[7] K. Tan, X. Xu and H. Bian, "The application of NDT algorithm in sonar image processing," *2016 IEEE/OES China Ocean Acoustics (COA)*, Harbin, China, 2016, pp. 1-4, doi: 10.1109/COA.2016.7535652.

[8] Q. Wang, S. Du, F. Wang and Y. Chen, "Underwater target recognition method based on multi-domain active sonar echo images," *2021 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, Xi'an, China, 2021, pp. 1-5, doi: 10.1109/ICSPCC52875.2021.9564611.

[9] C. Dong, L. Guo, K. Hu, J. Yin and X. Sheng, "Side-scan Sonar Image Rough Recognition and Feature Matching Based on CNN and SIFT," *2019 IEEE International Conference on Signal, Information and Data Processing (ICSIDP)*, Chongqing, China, 2019, pp. 1-5, doi: 10.1109/ICSIDP47821.2019.9172987.

[10] L. Lindzey and A. Marburg, "Extrinsic Calibration between an Optical Camera and an Imaging Sonar," *OCEANS 2021: San Diego – Porto*, San Diego, CA, USA, 2021, pp. 1-8, doi: 10.23919/OCEANS44145.2021.9705956.

[11] B. Kubicek, A. S. Gupta and I. Kirsteins, "Feature Engineering and Interpretation of Active Sonar Data Using Geometric Wavelets and Support Vector Machines," *OCEANS 2021: San Diego – Porto*, San Diego, CA, USA, 2021, pp. 1-5, doi: 10.23919/OCEANS44145.2021.9705879.

[12] C. Lei, H. Wang and J. Lei, "SI-GAT: Enhancing Side-Scan Sonar Image Classification Based on Graph Structure," in *IEEE Sensors Journal*, vol. 24, no. 15, pp. 24388-24404, 1 Aug.1, 2024, doi: 10.1109/JSEN.2024.3416193.

[13] A. Saenko, A. Mironov and E. Fomina, "Methods for Improving the Efficiency of Object Detection in Sonar Images," *2024 International Russian Automation Conference (RusAutoCon)*, Sochi, Russian Federation, 2024, pp. 578-582, doi: 10.1109/RusAutoCon61949.2024.10694046.

[14] C. -C. Chang, Y. -P. Wang and S. -C. Cheng, "Standardization of Sonar Images by Conditional Random Fields for Fish Segmentation With Mask R-CNN," *2021 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, Hualien City, Taiwan, 2021, pp. 1-2, doi: 10.1109/ISPACS51563.2021.9650982.

[15] M. ERCAN and Ç. Mehmet, "Realistic Sonar Image Generation With Rendering Method for Underwater Object Detection," *2024 11th International Conference on Electrical and Electronics Engineering (ICEEE)*, Marmaris, Turkiye, 2024, pp. 421-425, doi: 10.1109/ICEEE62185.2024.10779281.

[16] Q. Wang and F. Zhu, "Fine-grained Visual Recognition based on Prototypical Mamba," *2025 3rd International Conference on Intelligent Perception and Computer Vision (CIPCV)*, Hangzhou, China, 2025, pp. 25-29, doi: 10.1109/CIPCV65863.2025.00014.

[17] X. Liu, H. Zhu, W. Song, J. Wang, Z. Chai, and S. Hong, "Review of Object Detection Algorithms for Sonar Images based on Deep Learning," *Recent Patents on Engineering*, vol. 19, no. 3, Nov. 2023, doi: 10.2174/0118722121257145230927041949.

[18] M. S. Islam, A. K. M. Jayed, M. R. Islam, A. F. Arnob, and M. a A. Hassan, "Deep Learning-Based Sonar Image Object Detection System - International Journal of Engineering and Advanced Technology Studies (IJEATS)," *International Journal of Engineering and Advanced Technology Studies*, vol. 12, no. 4, pp. 34-47, Dec. 2024, doi: 10.37745/ijeats.13.

[19] S. Sürücü and B. Diri, "A hybrid approach for the detection of images generated with multi generator MS-DCGAN," *Engineering Science and Technology an International Journal*, vol. 63, p. 101969, Jan. 2025, doi: 10.1016/j.jestch.2025.101969.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 1 | pp. 01-12 | Jan-Mar 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 08 Nov 2025 | Revised: 23 Jan 2026 | Accepted: 01 Feb 2026 | Published: 08 Feb 2026

[20] M. Aubard, L. Antal, A. Madureira, L. F. Teixeira and E. Ábrahám, "ROSAR: An Adversarial Re-Training Framework for Robust Side-Scan Sonar Object Detection," *2025 Symposium on Maritime Informatics and Robotics (MARIS)*, Syros, Greece, 2025, pp. 1-8, doi: 10.1109/MARIS64137.2025.11139627.

[21] Y. Feng, S. Rong, C. Feng and B. He, "Side-Scan Sonar Image Segmentation Based on Improved Deeplabv3 Plus," *2023 8th International Conference on Intelligent Computing and Signal Processing (ICSP)*, Xi'an, China, 2023, pp. 1702-1706, doi: 10.1109/ICSP58490.2023.10248456.

[22] F. Zhang, W. Zhang, C. Cheng, X. Hou, and C. Cao, "Detection of small objects in Side-Scan sonar images using an enhanced YOLOV7-Based approach," *Journal of Marine Science and Engineering*, vol. 11, no. 11, p. 2155, Nov. 2023, doi: 10.3390/jmse11112155.

[23] G. Huo, Z. Wu and J. Li, "Underwater Object Classification in Sidescan Sonar Images Using Deep Transfer Learning and Semisynthetic Training Data," in *IEEE Access*, vol. 8, pp. 47407-47418, 2020, doi: 10.1109/ACCESS.2020.2978880.

[24] A. Chen, X. Ye, M. Yu and Z. Wang, "Improved Sonar Target Detection Method Based on YOLOv5," *2023 IEEE 11th International Conference on Computer Science and Network Technology (ICCSNT)*, Dalian, China, 2023, pp. 298-302, doi: 10.1109/ICCSNT58790.2023.10334608.