sustainability

MDPI

Article

Enhancing Underwater Object Detection and Classification Using Advanced Imaging Techniques: A Novel Approach with Diffusion Models

Prabhavathy Pachaiyappan [1,*], Gopinath Chidambaram [2], Abu Jahid [3] and Mohammed H. Alsharif [4,*]

[1] Department of Computer Science and Engineering, College of Engineering Guindy (CEG) Campus, Anna University, Chennai 600025, India
[2] Department of Electrical and Electronics Engineering, Sri Venkateswara College of Engineering, Sriperumbudur 602117, India
[3] School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON K1N 6N5, Canada; ajahi011@uottawa.ca
[4] Department of Electrical Engineering, College of Electronics and Information Engineering, Sejong University, Seoul 05006, Republic of Korea
* Correspondence: pprabhavathy@annauniv.edu (P.P.); malsharif@sejong.ac.kr (M.H.A.)

Abstract: Underwater object detection and classification pose significant challenges due to environmental factors such as water turbidity and variable lighting conditions. This research proposes a novel approach that integrates advanced imaging techniques with diffusion models to address these challenges effectively, aligning with Sustainable Development Goal (SDG) 14: Life Below Water. The methodology leverages the Convolutional Block Attention Module (CBAM), Modified Swin Transformer Block (MSTB), and Diffusion model to enhance the quality of underwater images, thereby improving the accuracy of object detection and classification tasks. This study utilizes the TrashCan dataset, comprising diverse underwater scenes and objects, to validate the proposed method's efficacy. This study proposes an advanced imaging technique YOLO (you only look once) network (AIT-YOLOv7) for detecting objects in underwater images. This network uses a modified U-Net, which focuses on informative features using a convolutional block channel and spatial attentions for color correction and a modified swin transformer block for resolution enhancement. A novel diffusion model proposed using modified U-Net with ResNet understands the intricate structures in images with underwater objects, which enhances detection capabilities under challenging visual conditions. Thus, AIT-YOLOv7 net precisely detects and classifies different classes of objects present in this dataset. These improvements are crucial for applications in marine ecology research, underwater archeology, and environmental monitoring, where precise identification of marine debris, biological organisms, and submerged artifacts is essential. The proposed framework advances underwater imaging technology and supports the sustainable management of marine resources and conservation efforts. The experimental results demonstrate that state-of-the-art object detection methods, namely SSD, YOLOv3, YOLOv4, and YOLOTrashCan, achieve mean accuracies (mAP@0.5) of 57.19%, 58.12%, 59.78%, and 65.01%, respectively, whereas the proposed AIT-YOLOv7 net reaches a mean accuracy (mAP@0.5) of 81.4% on the TrashCan dataset, showing a 16.39% improvement. Due to this improvement in the accuracy and efficiency of underwater object detection, this research contributes to broader marine science and technology efforts, promoting the better understanding and management of aquatic ecosystems and helping to prevent and reduce the marine pollution, as emphasized in SDG 14.

Keywords: underwater object detection; Sustainable Development Goal (SDG) 14; diffusion models; Convolutional Block Attention Module (CBAM); Modified Swin Transformer Block (MSTB); marine debris detection

Citation: Prabhavathy P; Chidambaram, G.; Jahid, A.; Alsharif, M.H. Enhancing Underwater Object Detection and Classification Using Advanced Imaging Techniques: A Novel Approach with Diffusion Models. *Sustainability* **2024**, *16*, 7488. https://doi.org/10.3390/su16177488

Academic Editor: Zhen-Zhong Hu

Received: 19 July 2024
Revised: 24 August 2024
Accepted: 27 August 2024
Published: 29 August 2024

1. Introduction

Object and Image Localization are crucial tasks in Computer Vision (CV). The algorithm in Object Localization identifies and pinpoints a specific object within an image. Conversely, Image Localization aims to detect and locate all objects present in the entire image. Researchers use deep learning models to detect potential objects within an image. During the detection phase, region proposal networks are used to identify and highlight areas likely to contain objects. Once objects are detected, precise localization further refines these regions by drawing bounding boxes around the identified objects. Object localization starts with the process of object detection, which applies a deep learning model to identify potential objects within an image. Researchers utilize different techniques to detect and mark regions with objects, such as CNNs, faster R-CNN, or SSD and YOLO.

Deep learning-based methods have significantly excelled in extracting deeper semantic information from images, as demonstrated by their success with the COCO natural image dataset given by (Lin, T.-Y. et al., 2014) [1]. Consequently, applying deep learning object detection technology to marine debris detection is a reliable approach. This technology enables precise identification and localization of marine debris for vision robots and differentiates debris from the surrounding biological environment. This ensures the effective cleanup of marine debris while preserving the integrity of the original ecological setting for Sustainable Development Goals 14: Life Below Water.

The Single Shot MultiBox Detector (SSD) proposed by (Liu, W. et al., 2016) [2] applies a few improvements including multi-scale features and default boxes, which makes improvements in SSD to match the Faster R-CNN's accuracy using lower-resolution images. Redmon, J. et al., 2018 [3] proposed YOLOv3, which predicts an objectness score for each bounding box using logistic regression. Bochkovskiy, A. et al., 2020[4] proposed YOLO v4, which is a one-stage object detection network which has a pretrained convolutional neural network such as VGG16 or CSPDarkNet53 trained on COCO [1] dataset. Wang, C.-Y. et al., 2023 [5] proposedYOLOv7, which outperforms all existing object detectors in terms of both speed and accuracy, operating between 5 FPS and 120 FPS and achieves the highest accuracy of 56.8% AP among all real-time object detectors running at 30 FPS or higher on a GPU V100.

Recent research on underwater image enhancement has been done to address challenges like distorted images and degradation in image quality. Liu, B. et al., 2024 [6] introduced a streamlined model called Rep-UWnet designed to improve underwater images. This model features a fully connected convolutional network and three sequentially linked densely connected RepConv blocks, with input images being connected to the output of each block through a Skip connection. Gong, T. et al., 2023 [7] introduced an underwater image enhancement method that utilizes color feature fusion. By taking advantage of how light propagates underwater, the proposed model implements a multi-channel feature extraction approach. Sun, T. et al., 2020 [8] reconstructed structural information for distorted images using image registration. Yeh, C.H. et al., 2024 [9] introduced a deep network model designed for enhancing single underwater images. Specifically, the framework features a light field module (LFM) and a sketch module, which work together to create a light field map of the target image. This map improves color representation and preserves original image details by supplying contour information. The underwater image is progressively enhanced with guidance from the light field map. Yang, J. et al., 2024 [10] introduced a salient region-guided fusion method for underwater image enhancement. An advanced dark channel prior technique is proposed to minimize haze effects in underwater images, which greatly enhances visibility.

1.1. Background

Marine ecosystems face a myriad of challenges, including pollution, climate change, and habitat destruction, jeopardizing biodiversity and global food security. Monitoring

underwater environments is crucial for understanding these complex ecosystems and implementing effective conservation and management strategies. Advanced imaging techniques play a pivotal role in this endeavor, enabling precise detection and classification of underwater objects. Traditional imaging in underwater environments is fraught with challenges such as light attenuation, water turbidity, and varying environmental conditions, which degrade image quality and hinder accurate object detection. Recent advancements in deep learning and image processing have spurred the development of novel techniques tailored for underwater applications. Addressing these challenges aligns with Sustainable Development Goal (SDG) 14: Life Below Water, which aims to conserve and sustainably use the oceans, seas, and marine resources.

This research focuses on enhancing underwater object detection and classification using cutting-edge imaging technologies. The proposed approach integrates diffusion models, Convolutional Block Attention Module (CBAM), and Modified Swin Transformer Block (MSTB) into a unified framework. Diffusion models simulate light propagation and interaction with underwater objects, effectively reducing noise and enhancing image clarity by capturing the inherent structure within noisy images (Siqi Lu et al., 2023) [11]. The CBAM dynamically recalibrates channel-wise and spatial-wise features, prioritizing informative features while suppressing irrelevant ones, thereby improving feature representation and object detection accuracy (Wang N. et al., 2024) [12]. Meanwhile, the MSTB enhances image quality and resolution through a modified U-Net architecture with transformer blocks, facilitating robust feature extraction and context aggregation (Kim H. et al., 2024) [13].

This novel approach addresses key challenges in underwater imaging by leveraging deep learning to enhance image quality, reduce noise, and improve the reliability of object detection systems. The integration of cutting-edge imaging technologies provides visibility enhancement in underwater scenes, thereby enabling the accurate identification and classification of marine debris, biological organisms, and submerged artifacts. The significance of this research extends across various domains, including autonomous underwater robotics for environmental monitoring, marine research and conservation, underwater archeology, and security surveillance (Tian et al., 2024; S D. et al., 2024) [14,15]. These applications rely on accurate and efficient underwater imaging technologies to gather data, study ecosystems, detect anomalies, and support decision-making processes.

This research introduces an innovative approach to improving underwater object detection and classification by leveraging cutting-edge imaging techniques. Thus, it contributes to broader efforts in marine conservation, resource management, and environmental protection by improving the fidelity and reliability of underwater imaging systems.

1.2. Literature Review

Underwater environments present unique challenges for imaging and sensing technologies due to factors such as light attenuation, water turbidity, and environmental variability. These challenges have spurred significant research into advanced imaging techniques aimed at improving the detection and classification of underwater objects. This literature review explores recent developments in this field, focusing on key methodologies, datasets, and performance metrics.

Effective underwater imaging is hindered by several factors that degrade image quality and complicate object detection. Light attenuation, caused by the absorption and scattering of light in water, reduces visibility and contrast, making it difficult to detect objects at varying depths (Almutiry, O. et al., 2024) [16]. Water turbidity further exacerbates these effects, introducing particulate matter that scatters light and reduces image clarity. Moreover, environmental conditions such as currents, sedimentation, and biological activity contribute to dynamic changes in water clarity and light conditions, posing additional challenges for imaging systems (Shuyun Yuan et al., 2023) [17].

Historically, underwater imaging relied on conventional methods such as sonar and acoustic imaging, which are effective for large-scale surveys but lack the resolution

needed for detailed object identification (DinhQuangHuy et al., 2023) [18]. Optical imaging techniques, including cameras and lidar, offer higher resolution but are limited by light attenuation and water turbidity, compromising their effectiveness at greater depths (Zhou, J. et al., 2023) [19]. Underwater images often suffer from color distortion, blurriness, and significant noise due to the scattering and absorption of light as it travels through water. Underwater images often suffer from color distortion, blurriness, and significant noise due to the scattering and absorption of light as it travels through water. Object detection and classification in such images leads to low accuracy. These challenges, along with recent advancements in deep learning, have revolutionized underwater object detection by enabling the extraction of meaningful features from noisy and degraded images through advanced imaging techniques. These techniques make use of neural networks, convolutional blocks, or swin transformer blocks to overcome the challenges found in underwater images. This helps in better training the developed network model to learn the significant features in the underwater images; thereby, underwater objects are detected and classified with high accuracy.

Convolutional Neural Networks (CNNs) have been extensively utilized for object detection and classification in various domains, including underwater environments (Zocco, F. et al., 2023) [20]. CNNs leverage hierarchical feature extraction to identify objects based on patterns and textures in images, overcoming traditional limitations in feature representation (Yang, Y. et al., 2023) [21].

Further, the CBAM enhances feature representation in CNNs by incorporating attention mechanisms that dynamically recalibrate channel-wise and spatial-wise features (Xin, H. et al., 2023) [22]. The informative features are focused while suppressing irrelevant ones, and CBAM improves the accuracy of object detection systems in challenging underwater conditions (Wang, X. et al., 2023) [23]. Inspired by the success of transformer architectures in natural language processing, MSTB integrates transformer blocks into CNNs to capture global dependencies and enhance feature interactions (Guang Yang et al., 2023) [24]. This approach improves the robustness of feature extraction in underwater images, facilitating more accurate object classification and localization. Diffusion models simulate the propagation and interaction of light with underwater objects, effectively denoising images and enhancing visibility (Zhang, H et al., 2024) [25]. Diffusion models improve the quality of underwater images and support precise object detection in varying environmental conditions by capturing the underlying structure within noisy images (Lu, S. et al., 2024) [26].

Also, Yeh et al. [27] introduced a lightweight deep neural network for simultaneous underwater object detection and color conversion, emphasizing computational efficiency in underwater environments. However, challenges such as varying light conditions and water turbidity can impact detection accuracy. Zhou et al. [28] developed "Yolotrashcan," a deep learning model aimed at detecting marine debris to support environmental conservation. This work faces challenges including the diversity of debris types and sizes and the dynamic nature of marine environments, which can introduce noise and occlusions that affect detection accuracy. Dhariwal and Nichol [29] compared diffusion models with GANs for image synthesis, highlighting the potential of diffusion models but noting challenges like training stability and high computational resource demands, especially for complex image synthesis tasks. Saleh and Vámossy [30] proposed BBBD, focusing on occlusion detection and order recovery in object detection, but encountered difficulties in scenarios with complex occlusions or overlapping objects, complicating precise bounding box detection.

Furthermore, Teng et al. [31] improved the YOLOv5 algorithm for detecting underwater garbage and addressing environmental concerns. Challenges include distinguishing garbage from natural underwater elements like reefs and vegetation and variations in garbage types and sizes. Liu et al. [32] explored domain generalization in underwater object detection to improve model adaptability across diverse underwater environments,

facing issues such as domain shift and limited labeled data availability for training generalized models. Sharma et al. [33] proposed a wavelength-based attributed deep neural network for restoring underwater images using spectral information. Challenges include limited spectral data availability and the need for accurate calibration to ensure reliable image restoration. Wang et al. [34] introduced a zero-shot image restoration method using a denoising diffusion null-space model, showcasing effective denoising capabilities but facing challenges with model complexity and the need for fine-tuning for specific restoration tasks. Liu et al. [35] presented DiffYOLO, combining the YOLO and Diffusion models for improved detection performance in noisy environments, encountering challenges in optimizing the fusion of these models and handling complex noise patterns.

Zeng et al. [36] focused on underwater target detection using Faster RCNN and adversarial occlusion networks to improve detection reliability, facing challenges with robustness against varying occlusion types and environmental conditions. Fan et al. [37] developed a dual refinement underwater object detection network aimed at increased accuracy, with challenges including computational complexity and the need for efficient training strategies for refinement. Jia et al. [38] introduced an underwater object detection method based on an improved EfficientDet model, focusing on efficiency and accuracy, facing challenges with handling scale variations and diverse underwater object types. Chen et al. [39] proposed SWIPENET for object detection in noisy underwater images, addressing noise robustness, with challenges in fine-tuning for different noise levels and types commonly encountered in underwater imaging.

Fayaz et al. [40] provided a comprehensive review of architectures and algorithms for underwater object detection, noting that staying updated with rapidly evolving techniques and addressing specific application requirements in diverse underwater scenarios remains challenging. Wu et al. [41] proposed an improved YOLOv5-based method for fish target detection in underwater blurred scenes, facing challenges in accurately distinguishing fish from other underwater elements and handling motion blur effects.

The ICRA dataset comprises a diverse collection of underwater images, including biological organisms, marine debris, and submerged artifacts (Hong, L. et al., 2023) [42]. This dataset serves as a benchmark for evaluating object detection algorithms in real-world underwater scenarios, providing annotated images for training and testing models. The TrashCan dataset focuses specifically on marine debris detection, annotating images with bounding boxes and segmentation labels. This dataset supports the development of robust detection algorithms for identifying and classifying underwater trash, a critical component of marine conservation efforts.

The evaluation of underwater object detection systems relies on metrics such as precision, recall, and Mean Average Precision (mAP). Precision measures the accuracy of positive predictions, recall assesses the proportion of true positives correctly identified, and mAP evaluates the overall detection performance across multiple object classes. These metrics provide quantitative insights into the effectiveness of advanced imaging techniques in underwater environments. Recent case studies highlight the practical applications of advanced imaging techniques in underwater research and exploration. For instance, autonomous underwater vehicles equipped with CNN-based object detection systems have been deployed for environmental monitoring and habitat assessment. These systems enable real-time data collection and analysis, supporting scientific research and conservation efforts in marine ecosystems. While significant progress has been made in underwater object detection using advanced imaging techniques, several challenges remain. Improving the robustness of detection algorithms in complex underwater environments, integrating multi-modal sensor data for enhanced perception, and addressing ethical considerations in marine research are critical areas for future exploration (Deluxni, N. et al., 2023) [43].

Advanced imaging techniques coupled with deep learning have transformed underwater object detection and classification by overcoming traditional limitations and enhancing the accuracy and reliability of detection systems. By leveraging CBAM, MSTB,

and Diffusion models, researchers can effectively navigate the complexities of underwater imaging, contributing to advancements in marine conservation, environmental monitoring, and scientific exploration.

1.3. Challenges and Objectives

The following list encapsulates the core challenges and objectives addressed by this research, focusing on enhancing underwater object detection and classification through advanced imaging techniques.

- Existing methods for underwater object detection suffer from limited accuracy and reliability due to challenges such as light attenuation, water turbidity, and environmental variability.
- Current imaging techniques often fail to provide clear and detailed images necessary for precise object classification and localization in diverse underwater conditions.
- Effective detection and classification of marine debris, crucial for environmental monitoring and conservation efforts, remains a significant challenge due to the complex underwater environment and varying debris types.
- There is a need to integrate advanced imaging technologies, including Convolutional Block Attention Module (CBAM), Modified Swin Transformer Block (MSTB), and Diffusion models, to enhance the quality and clarity of underwater images for improved object detection.
- Defining robust performance metrics such as precision, recall, and Mean Average Precision (mAP) specific to underwater environments is essential to accurately assess the efficacy of detection algorithms.
- Developing real-time object tracking capabilities using integrated approaches like YOLOv7 and DeepSORT to maintain continuous object detection and tracking in dynamic underwater scenarios.

This research article starts with an introduction section providing background knowledge, related works in this research area, challenges addressed, and objectives of this research work. Section 2 provides the model of the proposed system followed by a description of the dataset and techniques adapted in Section 3. Then, results obtained from this system are discussed in Section 4 and are concluded with a discussion of its significance in marine research, conservation, and environmental monitoring in Section 5.

2. Modeling of the System

The structured approach integrates advanced imaging techniques and state-of-the-art algorithms to address the complexities of underwater object detection and classification. Each module and algorithm cited contributes uniquely to enhancing system performance in challenging underwater environments. The proposed system architecture is shown in Figure 1.

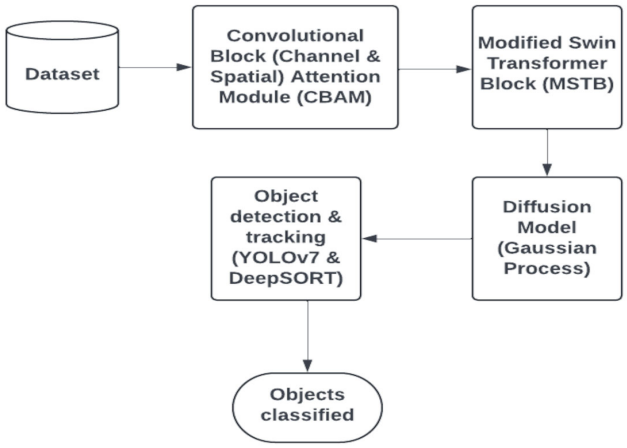


Figure 1. Proposed System architecture.

The proposed system uses TrashCan dataset [44], which is a semantically segmented collection consisting of 7212 images, mostly drawn consecutively from 312 distinct video sequences recorded by JAMSTEC in the Sea of Japan since 1982. This dataset represents an enhanced version of the Trash-ICRA19 dataset [45]. The TrashCan dataset is dedicated to detecting marine debris, featuring images annotated with bounding boxes and segmentation labels. It aids in developing effective detection algorithms for identifying and classifying underwater trash, playing a vital role in marine conservation efforts. This dataset has a total of 7212 images with 22 classes, where 5066 images are used for training, 721 images for validation, and 1425 images for testing purposes. Each image class has multiple annotations as labels; thereby, 2426 labels are found for 1425 images. CBAM acts as color correction module by focusing features on spatial and channel attention modules, thereby improving image quality by enhancing color information in underwater scenes. The Convolutional Block Attention Module (CBAM) plays a crucial role in improving image quality by preserving and enhancing color information in underwater scenes. CBAM employs advanced algorithms to mitigate color distortion and enhance visibility, addressing challenges such as light attenuation and water turbidity. CBAM ensures clearer and more accurate representations of underwater environments by emphasizing informative spatial and channel-wise features. The Modified Swin Transformer Block (MSTB) integrates a modified U-Net architecture with transformer mechanisms to enhance image resolution and detail. This module excels in capturing both global dependencies and local context, crucial for improving object detection precisely underwater. MSTB's capability to process high-level features enhances the system's ability to classify and localize objects effectively. The Diffusion module employs Gaussian processes to reduce noise and artifacts in underwater images. This technique enhances image clarity and reduces interference, thereby improving the system's robustness in challenging visual conditions. The Diffusion model gradually adds noise to the original distribution through a Markov chain and gradually recovers from the latent distribution to the original distribution by using a learned denoising process. This contributes significantly to the overall performance of the object detection system by effectively denoising images.

The integration of DeepSORT and YOLOv7 algorithms enhances the system's object tracking and detection capabilities. DeepSORT improves tracking accuracy by associating object identities across frames, while YOLOv7 provides efficient real-time object detection using a single-shot detection approach. In general, YOLOv7 provides a fast and strong network architecture that provides a more effective feature integration method, more accurate object detection performance, a more robust loss function, and an increased label assignment and model training process efficiency. As a result, YOLOv7 requires several

times cheaper computing hardware than other deep learning models. Together, these algorithms enable the proposed AIT-YOLOv7 net to detect, track, and classify underwater objects with high accuracy and efficiency, essential for applications in underwater exploration, research, and surveillance.

3. Dataset and System Description

The dataset encompasses a variety of images showcasing different objects, including marine debris, biological organisms like plants and animals, and remotely operated vehicles (ROVs) as shown in Figure 2. It consists of 7212 images that vary in quality, depth, scene compositions, and camera types utilized.

Figure 2. Selection of images, highlighting the dataset's diversity.

The images depict a range of marine debris captured in real-world environments, showcasing various objects under conditions like decay, occlusion, and overgrowth. They also exhibit significant variations in water clarity and lighting quality across different scenes.

3.1. Convolutional Block Attention Module

Underwater images are processed through a neural network architecture incorporating a VGG16 module for feature extraction, alongside channel and spatial attention modules and a color correction module (CC_Module). This color correction model is trained using a dataset containing both hazy and clean images. The architecture includes initial convolutional layers with varied kernel sizes to capture diverse image features. Following these convolutional operations, local attention modules like CBAM (Convolutional Block Attention Module) are employed to further enhance feature representation.

CBAM (Convolutional Block Attention Module) integrates two essential components: ChannelGate and SpatialGate. These elements dynamically recalibrate features within intermediate feature maps during training. The ChannelGate recalibrates channel-wise features, emphasizing important channels while suppressing less relevant ones. Simultaneously, the SpatialGate recalibrates spatial-wise features, focusing on informative spatial regions while reducing the impact of irrelevant areas. The process flow description for the color correction module of CBAM is shown in Figure 3.

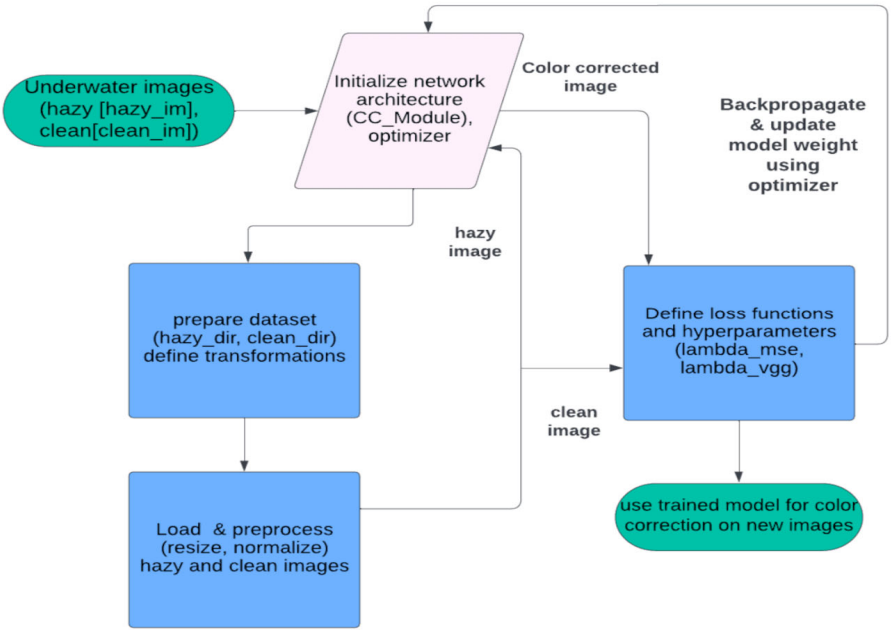


Figure 3. Process flow description for color correction module of CBAM.

This dual recalibration mechanism enables CBAM to enhance feature maps effectively, improving the model's ability to capture relevant information and suppress noise, thereby enhancing the overall performance of the neural network in tasks such as object detection and classification in underwater imaging scenarios. Figure 4 presents unprocessed images directly captured from the data source, representing the raw state of the images without any modifications or enhancements applied.

Figure 4. Unprocessed images.

Figure 5 illustrates the output image after applying color correction to the input image from Figure 4.

Figure 5. Output image after color correction from the input image.

The Convolutional Block Attention Module (CBAM) enhances the quality of color correction by emphasizing significant spatial regions and enhancing global information within the feature maps. This attention mechanism ensures precise retention of colors and tones, resulting in an accurate visual representation of the scene. During the training process, batches of hazy images from the dataset are processed through the CBAM-equipped neural network. The model computes outputs for these images, which are then compared with corresponding clean images to calculate loss. VGG Loss is computed by iterating through the feature maps of the color corrected image and the clean image. The Euclidean distance between the two feature maps is computed for each pair of units. The distances are then summed up and weighed based on the dimensions of the feature maps. The Mean Squared Error (MSE) is a straightforward and widely used loss function, which takes the difference between the actual value and the model prediction, squares it, and then averages it across the entire dataset. An optimizer adjusts the model's parameters iteratively to minimize this loss, enhancing the model's ability to accurately correct colors in underwater environments. Checkpoints are saved to preserve the model's state, allowing for training resumption or inference in subsequent sessions.

Thus, the integration of CBAM in this method plays a crucial role in maintaining accurate color representation in underwater imagery. CBAM improves the fidelity of color information essential for tasks like object recognition and classification in challenging underwater conditions by selectively enhancing spatial and channel-wise features.

3.2. Modified Swin Transformer Block (MSTB)

Underwater images often suffer from degradation caused by absorption and scattering in the medium. To address these challenges, our proposed method enhances a U-Net architecture with Swin Transformer Blocks. This approach aims to capture both global dependencies and local contexts crucial for improving image quality in underwater scenes. Unlike traditional linear layers, which are often used in U-Net architectures, our method replaces them with two convolutions to reinforce channel and spatial attention mechanisms, while preserving the core attention functionality.

The U-Net architecture, originally designed for medical image segmentation, proves to be beneficial due to its ability to operate effectively with limited annotated data. This characteristic is particularly advantageous in underwater imaging contexts, where comprehensive datasets are often scarce. By integrating Swin Transformer Blocks, our enhanced U-Net framework not only maintains speed and accuracy, but also enhances its capability to handle complex underwater image features, thereby improving overall performance in tasks such as object detection and classification. The U-Net architecture comprises a contracting path and an expansive path as shown in Figure 6.

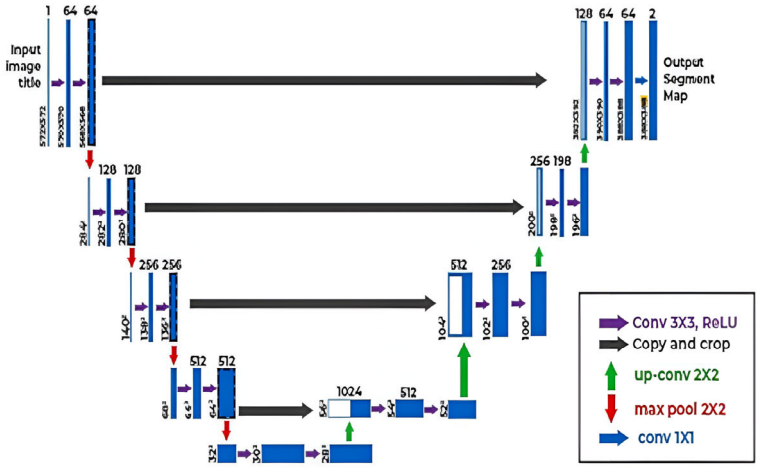


Figure 6. U-Net architecture.

Encoder layers in the contracting path capture contextual information and reduce spatial resolution of the input. These layers perform convolutional operations to deepen feature maps, extracting progressively abstract representations akin to feedforward layers in CNNs. In contrast, the expansive path's decoder layers decode encoded data, maintaining spatial resolution while upsampling feature maps. Skip connections from the contracting path help retain the spatial information lost during contraction, thereby enhancing the accuracy of feature localization using decoder layers.

The ViT model revolutionizes image recognition tasks like object detection, segmentation, classification, and action recognition by treating images as sequences of patches. Each patch is flattened into a vector by concatenating pixel channels and linearly projecting them to the desired input dimension. The ViT encoder block integrates key components: Layer Norm ensures stable training and adaptation across diverse image characteristics, while the Multi-head Attention Network (MSP) generates attention maps to focus on crucial image regions. The Multi-Layer Perceptron (MLP) head, with Gaussian Error Linear Unit (GELU) activation, serves as the final classifier output, often used for image classification with softmax. GELU weights inputs based on their probabilities under a Gaussian distribution, enhancing ViT's performance in learning intricate image structures independently. Choosing the right activation function is crucial for the success of deep learning models, impacting their ability to learn, maintain stability, and operate efficiently. Recently, the Gaussian Error Linear Unit (GELU) has become a popular choice, often outperforming traditional functions like the Rectified Linear Unit (ReLU) in many applications. Equation (1) states that we scale x by how much greater it is than other inputs using the Gaussian distribution, which is often computed with the error function as follows:

$$GELU(x) = \frac{x}{2}\left(1 + erf\left(\frac{x}{\sqrt{2}}\right)\right) \quad (1)$$

where erf in Equation (2) denotes the error function given by the following:

$$erf(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\, dt \quad (2)$$

The Swin Transformer adopts a hierarchical architecture by dividing the image into non-overlapping patches initially. Unlike Vision Transformers, the Swin Transformer employs shifted windows that enable patches to attend to neighboring patches. This approach facilitates more effective information exchange between local and global features.

The Swin Transformer incorporates hierarchical stages of transformers, where each stage refines features at different resolutions. This design choice allows for the model to capture both local details and global context efficiently. By refining features across multiple stages, the Swin Transformer enhances its ability to understand intricate relationships within the input data, making it particularly effective for tasks like image recognition, object detection, and segmentation in diverse and complex visual environments. The Modified Swin Transformer Block enhances the image's resolution and improves the image's clarity as depicted in Figure 7a. The process flow for image resolution enhancement, as shown in Figure 7b.

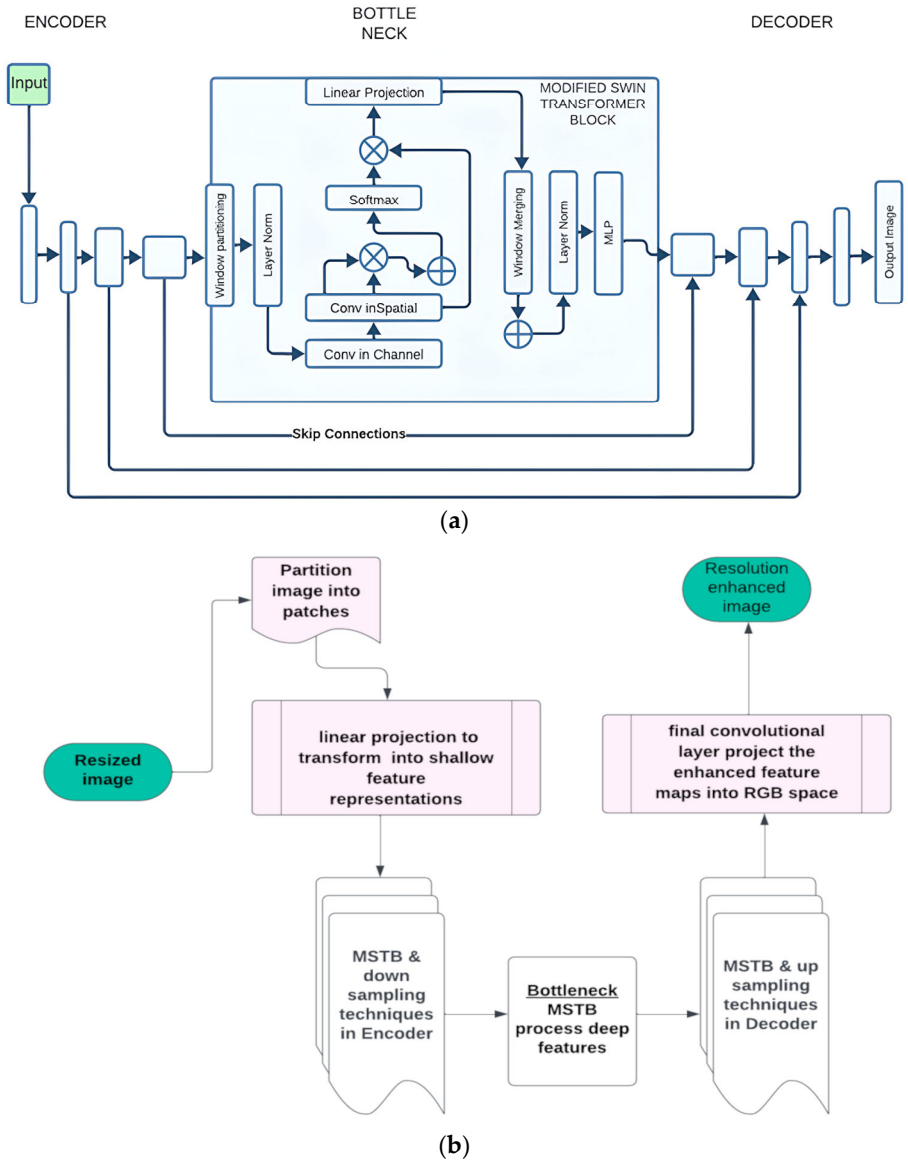


Figure 7. (**a**) Modified Swin Transformer Block (MSTB) for enhancing image resolution. (**b**) Process flow of MSTB in underwater image resolution enhancement.

The proposed AIT-YOLOv7 net adopts a modified U-Net architecture enhanced with convoluted layers inspired by Swin Transformers. It comprises three main components: the encoder, bottleneck, and decoder. The encoder transforms the input into a deeper feature space, reducing spatial dimensions while increasing channel depth. At the network's

core, the bottleneck focuses on learning crucial high-level features while maintaining feature dimensions. In deep networks, extracting additional features becomes ineffective when the network encounters a bottleneck. At this stage, we perform feature extraction once more, aiming to compel the network to consolidate valuable information from the existing features, thereby achieving feature compression rather than merely adding more features. This process helps enhance the global dependencies within the network. We use two convolutions instead of a linear layer to reinforce in the channel and spatial. This is achieved in channel by a 1 × 1 convolution to triple the channels, which is like a linear layer. The spatial is enhanced by a 3 × 3 convolution in channel-wise. The decoder reconstructs the underwater image from the feature space, enlarging spatial dimensions while reducing channels. Task-specific upsampling in the decoder generates both enhanced and high-resolution images simultaneously.

While U-Net effectively captures global and local context, relying solely on it may be insufficient. To enhance global dependencies, standard Swin Transformer Blocks (STBs) replace conventional convolutional blocks. However, given the limitations of small datasets, CNNs are reintegrated to bolster local attention. Moreover, convolutions replace linear layers within STBs to enhance channel and spatial features concurrently, thereby reinforcing the core attention mechanism. The encoder is used to map input into deeper feature space, while the decoder is utilized to reconstruct the image from feature space. The bottleneck can learn useful compression of features. This modification enhances the network's ability to capture intricate relationships within underwater images, improving both the resolution and clarity for enhanced object detection and classification tasks, as depicted in Figure 8.

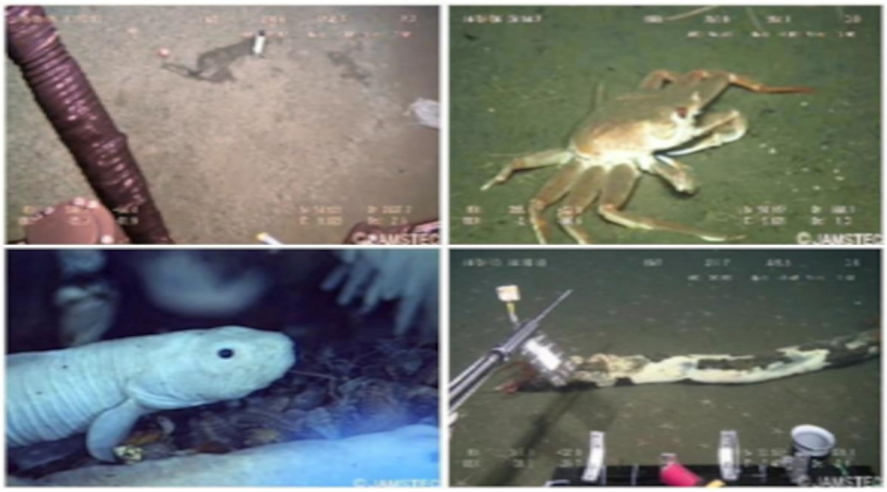


Figure 8. Enhanced images.

3.3. Diffusion Model

Diffusion models operate through two key processes: a forward process and a reverse process. During the forward process, noise is progressively introduced in timesteps, following a Gaussian distribution modeled as a Markov chain. In the reverse process, the neural network learned this noise with time embeddings, allowing it to reverse the noise that was added.

The diffusion process systematically introduces known noise into the input distribution, modeled as a Markov chain with probability density q over T iterations. Gaussian noise is progressively added to the input distribution $q(x^{(0)})$, as shown in Equation (3).

$$q\left(x^{(0\bullet\bullet\bullet T)}\right) = q\left(x^{(0)}\right)\prod_{(t=1)}^{T} q\left(x^{(t)}\middle|x^{(t-1)}\right) \quad (3)$$

where $q(x^{(t)} \mid x^{(t-1)})$ corresponds to a Gaussian distribution during the forward diffusion process. U-Net with ResNet blocks in this neural network reconstructs the original data by gradually reversing this noising process, which is called reverse diffusion. After training, these models can generate new data by starting with random Gaussian noise and applying the learned denoising steps. This neural network's task is to estimate the total noise in an image at a specific timestep. Comparing this estimate with the actual noise added to the image, the network becomes trained. During inference, the network continues to predict the total noise at timestep*t* and then removes a portion of this noise based on the schedule employed.

Noise is added at each timestep according to a specified pattern with help of scheduler to find the precise amount of noise to be introduced. The noise schedule dictates the process by which diffusion models add and remove noise in an image. Linear schedules are straightforward, but can sometimes lead to lower output quality, whereas cosine schedules enhance results by providing smoother transitions. The algorithmic framework for a diffusion model using Gaussian distribution is shown in Figure 9.

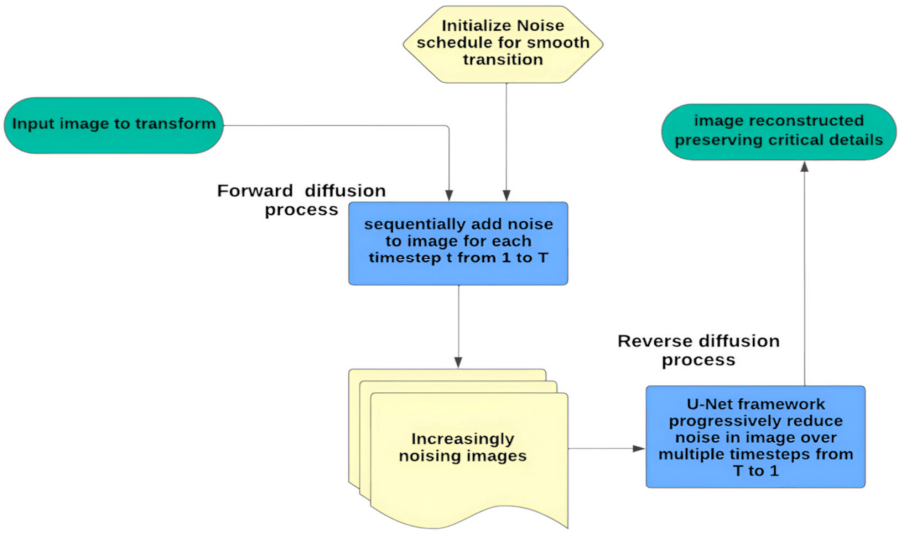


Figure 9. Process flow in diffusion model.

The U-Net architecture employed in this method, featuring an encoder–decoder structure with skip connections that preserve crucial spatial information throughout the network. These connections are vital for maintaining detailed image reconstruction from input to output. Within the U-Net framework, ResNet blocks are integrated into both the encoder and decoder sections. ResNet blocks effectively address the vanishing gradient problem by utilizing shortcut connections, enabling the use of deeper network architectures without a degradation in performance. For the iterative denoising process, the U-Net model is applied sequentially to progressively reduce noise in the input image over multiple timesteps (t) from T down to 1. This iterative approach ensures the gradual refinement and reconstruction of the denoised image. This process begins with an initialized noise-corrupted version of the original image. Figure 10 illustrates this initial corrupted image, which acts as the input for the diffusion process.

Figure 10. Corrupted image.

Each encoder stage includes two Residual Blocks with convolutional down-sampling, except for the final stage. Each decoder stage comprises three Residual Blocks and nearest neighbor up-sampling blocks with convolutions, which are used to reconstruct the input from the previous step. Skip connections link each stage in the decoder path to the corresponding stage in the encoder path. The model incorporates attention modules at a single feature map resolution, and the timestep t is encoded into a time embedding. The noise schedule used in this Diffusion model gives a smooth transition, which stabilizes the reverse diffusion process by ensuring that samples at any timestep are equally valuable to the training process. The output generated by the Diffusion fusion model is shown in Figure 11 and illustrates the results after applying noise reduction techniques.

Figure 11. Image generated using the Diffusion model.

These techniques significantly enhance visual clarity by reducing noise artifacts in the image. By explicitly modeling and training the diffusion process with a neural network, the proposed AIT-YOLOv7 net effectively captures the inherent structure of images while preserving critical details through noise removal.

3.4. Object Detection and Tracking

Combining YOLOv7 and DeepSORT creates a robust system for real-time object detection and tracking in videos. YOLOv7 excels in rapid single-shot detection by leveraging deep neural networks to divide frames into a grid. It efficiently predicts bounding boxes and class probabilities for each grid cell, utilizing anchor boxes to enhance localization accuracy. This approach allows for YOLOv7 to swiftly process frames during inference, making it highly suitable for real-time applications where speed is critical.

YOLOv7 provides greatly improved real-time object detection accuracy without increasing the inference costs. It can effectively reduce about 40% of parameters and 50% computation of state-of-the-art real-time object detections and achieve faster inference speed and higher detection accuracy. Performance metrics shown in [5] indicate YOLO v7's superiority in both accuracy and speed, thereby achieving precise object detection results across diverse object classes. State-of-the-art detection models were employed to test their performance (AP is the average precision) pertained on the COCO [1] dataset, as shown below in Figure 12.

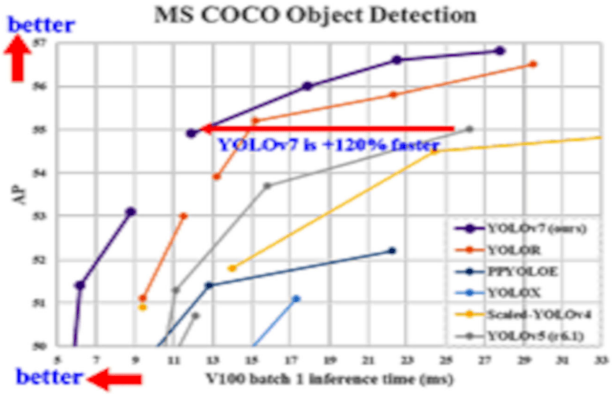


Figure 12. MS COCO object detection from [5].

The benchmarks provided in [5] show that YOLOv7 can effectively reduce about 40% of parameters and 50% computation of state-of-the-art real-time object detections when compared to other known object detectors. It achieves faster inference speed and higher detection accuracy. When compared with YOLOv4, YOLOv7 reduces the number of parameters by 75%, requires 36% less computation, and achieves 1.5% higher Average Precision. Performance metrics indicate YOLO v7's superiority in both accuracy and speed compared to traditional CNNs. YOLOv7 provides a fast and strong network architecture that provides a more effective feature integration method, more accurate object detection performance, a more robust loss function, and an increased label assignment and model training process efficiency. As a result, YOLOv7 requires several times cheaper computing hardware than other deep learning models.

DeepSORT enhances YOLOv7's capabilities by seamlessly tracking detected objects across successive frames. It incorporates appearance features and motion information to associate objects, employing deep learning to accurately represent object appearances. Additionally, DeepSORT integrates Kalman filtering for predicting object positions, particularly effective in handling occlusions where one object obscures another, such as shoals of fish underwater. This combination of YOLOv7 and DeepSORT not only ensures efficient real-time detection and tracking of objects, but also addresses challenges like occlusions, crucial for applications in dynamic environments such as underwater surveillance and monitoring systems.

When integrated, YOLOv7 and DeepSORT form a powerful pipeline for object detection and tracking. YOLOv7 excels in rapid object identification within each frame, leveraging its efficient single-shot detection capabilities. Meanwhile, DeepSORT enhances this process by maintaining continuous tracks of detected objects across frames, thereby improving the system's overall understanding of object movements and interactions over time. Figure 13 showcases the final output image, illustrating accurate object classification achieved through the application of color correction by CBAM, resolution enhancement by MSTB, and diffusion techniques in Figure 11 given as input.

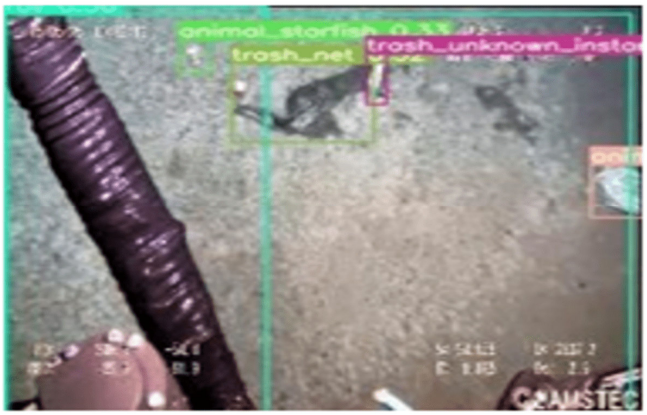


Figure 13. Output image with underwater objects detected and classified using AIT-YOLOv7.

These enhancements play a critical role in improving visual clarity and reducing noise artifacts, thereby enhancing the precision of object detection and classification. This proves highly effective in scenarios necessitating real-time monitoring and management of marine habitats. Capturing videos with detailed insights into the health and dynamics of marine ecosystems, AIT-YOLOv7 contributes significantly to environmental monitoring and conservation efforts in underwater surveillance systems.

4. Results and Discussions

This research uses the TrashCan dataset, which is a semantically segmented collection consisting of 7212 images, mostly drawn consecutively from 312 distinct video sequences recorded by JAMSTEC. This dataset represents an enhanced version of the Trash-ICRA19 dataset. The ICRA dataset encompasses a diverse array of images featuring biological entities like plants, animals, and remotely operated vehicles (ROVs). In contrast, the TrashCan dataset focuses specifically on marine debris sourced from various environments and meticulously annotated with bounding boxes and segmentation labels. This dataset plays a crucial role in the development of robust detection systems for marine debris, addressing the significant environmental threat posed by underwater trash. The total 7212 images have 22 classes, where 5066 images are used for training, 721 images for validation, and 1425 images for testing purposes. These images serve as a comprehensive collection for studying various underwater scenarios.

Marine debris remains a pressing challenge for aquatic ecosystems, prompting numerous approaches from environmental and governmental bodies aimed at cleanup and mitigation. Despite these efforts, effective solutions remain limited. This research is conducted using a NVIDIA Tesla T4 GPU in a Colab Pro, which has 16GB GDDR6, 65 teraflops of peak performance for FP16, thereby providing substantial computational power to facilitate the implementation and evaluation of advanced imaging techniques for underwater object detection and classification. Compared to the best-performing Cascade-Mask R-CNN models, YOLOv7 achieves 2% higher accuracy at a dramatically increased inference speed. When compared to YOLOR (You Only Learn One Representation), YOLOv7 reduces the number of parameters by 43% parameters, requires 15% less computation, and achieves 0.4% higher Average Precision. In summary, YOLOv7 is designed to be more efficient compared to its predecessors, leveraging architectural innovations to balance accuracy and computational efficiency. The computational complexity of our proposed approach is as follows: FPS as 114, input size as 640, and average time in batch size 16 as 3.8 ms. YOLOv7 has high speed and the best accuracy in achieving precise object detection results across diverse object classes, and these are the underlying reasons for choosing it in the development of Advanced Imaging Techniques (AIT) for underwater object detection and classification in real-world underwater scenarios.

The proposed system methodology leverages the Convolutional Block Attention Module (CBAM), Modified Swin Transformer Block (MSTB), and Diffusion model to enhance the quality of underwater images, thereby improving the accuracy of object detection and classification tasks. The CBAM enhances the clarity and accuracy of underwater environment representations by focusing on informative spatial and channel-wise features. The Modified Swin Transformer Block (MSTB) integrates a revised U-Net architecture with transformer mechanisms to improve image resolution and its detail. This module captures both global dependencies and local context, which is essential for precise object detection in underwater images. MSTB's ability to process high-level features enhances the proposed system's effectiveness in classifying and localizing objects. The Diffusion model uses Gaussian processes to minimize noise and artifacts in underwater images, enhancing image clarity and reducing interference, thereby strengthening the proposed system's performance under challenging visual conditions.

The proposed AIT-YOLOv7 net precisely detects and classifies trash of diverse classes in the TrashCan dataset, thereby significantly contributing to the SDG 14: Life Below water.

4.1. Performance Metrics

1. *Precision*:

Precision measures the proportion of true positive predictions among all positive predictions made by the model. It is calculated using the following formula:

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

where:

True Positives (*TP*) are the number of correct positive predictions.

False Positives (*FP*) are the number of incorrect positive predictions.

2. *Recall*:

Recall measures the proportion of true positive predictions among all actual positive instances in the dataset. It is calculated using the following formula:

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

where:

True Positives (*TP*) are the number of correct positive predictions.

False Negatives (*FN*) are the number of actual positive instances incorrectly predicted as negative.

(a) Mean Average Precision

Mean Average Precision (*mAP*) assesses the accuracy and precision across various classes by computing Average Precision (*AP*) from the precision-recall curve. It is a widely used metric in object detection tasks, including underwater object detection, to evaluate the algorithm's effectiveness. Mean Average Precision is calculated as the mean of Average Precision (*mAP)* values calculated for each class. *AP* for a class is determined from the precision–recall curve, which summarizes the trade-off between precision (the proportion of true positive predictions among all positive predictions) and recall (the proportion of true positive predictions among all actual positive instances).

$$mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k \tag{6}$$

where:

AP_k = the average precision of class k.

n = the number of classes.

In the context of underwater object detection, achieving a high *mAP* indicates that the algorithm can accurately detect and classify underwater objects across different categories or classes, considering varying conditions such as water clarity, object size, and environmental variability.

4.2. Ablation Study

This study emphases the importance of advanced imaging techniques used in this proposed approach such as the following:

(i) Color correction by CBAM with Resolution enhancement by MSTB.
(ii) Diffusion model using Gaussian process.

The training process makes use of the following parameters:

- Input images resized to 640 × 640 pixels.
- Training Epochs is 220.
- Batch size is 16.
- Learning rate is 0.01.
- Momentum is 0.9.
- Weight decay is 0.0005.

Single-shot object detection performs a prediction on the presence and location of objects within an image by processing the entire image in one shot. Mean Average Precision (mAP) is commonly used to assess the performance of object detection systems. Table 1 shows the overall performance of different single-shot object detection techniques using the TrashCan dataset with the mAP@.5 metric.

Table 1. Overall performance of different single-shot object detection techniques.

Sl.No.	Single-Shot Object Detection Techniques (Year)	mAP@.5
1	SSD (2016)	57.19%
2	YOLOv3 (2018)	58.12%
3	YOLOv4 (2020)	59.78%
4	YOLOTrashCan (2023)	65.01%
5	AIT-YOLOv7 (ours)	81.40%

The experimental results given in Table 1 depict that the proposed AIT-YOLOv7 has a significant improvement of 16.39%, 21.62%, 23.28%, and 24.21% when compared with the state-of-the-art object detection techniques YOLOTrashCan [28], YOLOv4 [4], YOLOv3 [3], and SSD [2] respectively.

The combination of the Modified Swin Transformer Block (MSTB) with the Convolutional Block Attention Module increases mAP@.5 by 1.41%, and the Diffusion model combined with them leads to an increase of 3.08% in mAP@.5 value.

The color correction along with resolution enhancement performed by the combined effect CBAM and MSTB techniques has less improvement without the Diffusion model. The Diffusion model significantly learns the intricate structure found in underwater images by adding Gaussian noise in each timestep t, which enables the neural network to capture the inherent features while preserving critical details through the noise removal process. Thus, the high-quality image reconstructed from the Diffusion model achieves significant improvement in object detection. A comparative analysis in Table 2 highlights the efficacy of integrating advanced imaging techniques with state-of-the-art object detection and tracking algorithms, paving the way for more accurate and reliable underwater monitoring systems.

Table 2. Ablation comparison on model performance using TrashCan dataset.

Model	Advanced Image Enhancement Techniques	mAP@.5
YOLOv7	CBAM	76.91%
	CBAM + MSTB	78.32%
	CBAM + MSTB + Diffusion	81.40%

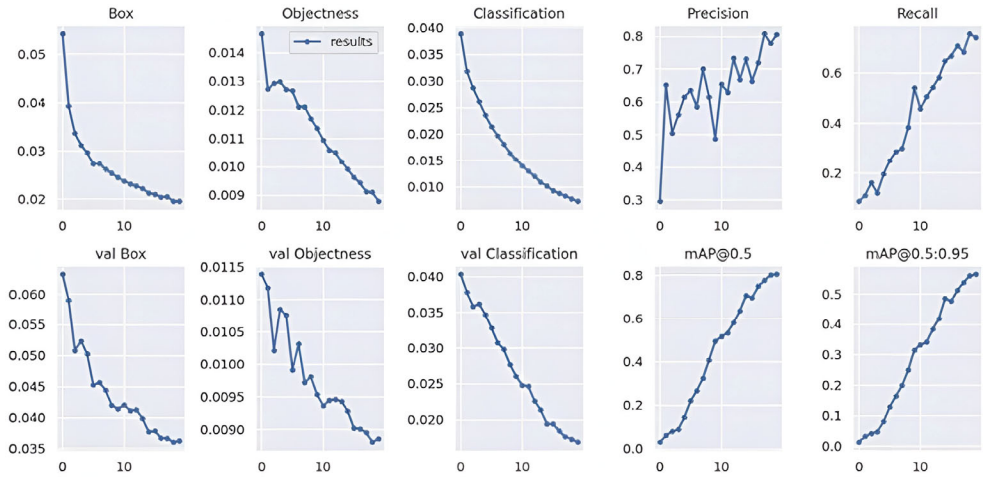
Precision, calculated using Equation (4), measures the proportion of true positive predictions among all positive predictions made by the model, reflecting the accuracy of identifying relevant objects without including false positives.

Recall, determined using Equation (5), measures the proportion of true positive predictions among all actual positive instances in the dataset, indicating the model's ability to detect all relevant objects.

The mAP, assessed using Equation (6), evaluates accuracy and precision across various classes by computing Average Precision (AP) from the precision–recall curve, providing a comprehensive measure of the algorithm's efficiency and effectiveness in detecting underwater objects.

The TrashCan underwater image dataset with 22 classes has a wide range of diversities, which serves as a benchmark for evaluating the performance of image enhancement and acts as the foundation of learning-based methods. Usually, we stop training a model when model loss starts to increase, or accuracy starts to decrease. To decide on the change in these generalization errors, we evaluate the model on the validation set after each epoch. The proper training of model using an appropriate dataset also has an impact on performance along with proposed image enhancement techniques for the betterment. Figure 14 depicts the performance metrics for the proposed AIT-YOLOv7, which shows the better improvement in object detection and classification by using advanced imaging techniques namely Convolutional Block Attention Module (CBAM), Modified Swin Transformer Block (MSTB), and Diffusion models with YOLOv7 and DeepSORT.

The precision, recall, and mAP@.5 metrics collectively offer a robust evaluation framework, highlighting the proposed system's strengths and weaknesses, particularly in challenging underwater environments characterized by occlusion, low visibility, and diverse object types.



Class	Images	Labels	P	R	mAP@.5	mAP@.5:.95: 100%
all	1425	2426	0.775	0.783	0.814	0.581
animal_crab	1425	83	0.66	0.795	0.796	0.478
animal_eel	1425	59	0.589	0.831	0.751	0.465
animal_etc	1425	46	0.645	0.543	0.609	0.392
animal_fish	1425	149	0.742	0.799	0.855	0.597
animal_shells	1425	60	0.731	0.65	0.668	0.431
animal_starfish	1425	59	0.621	0.797	0.738	0.392
plant	1425	94	0.852	0.734	0.757	0.465
rov	1425	700	0.786	0.887	0.903	0.716
trash_bag	1425	174	0.784	0.833	0.882	0.606
trash_bottle	1425	29	0.776	0.862	0.894	0.71
trash_branch	1425	70	0.811	0.957	0.935	0.719
trash_can	1425	83	0.91	0.853	0.93	0.686
trash_clothing	1425	15	0.847	0.933	0.935	0.825
trash_container	1425	106	0.839	0.906	0.904	0.688
trash_cup	1425	17	0.93	0.781	0.959	0.647
trash_net	1425	23	0.765	0.565	0.786	0.434
trash_pipe	1425	28	0.908	0.857	0.917	0.719
trash_rope	1425	16	0.433	0.622	0.437	0.281
trash_snack_wrapper	1425	14	0.795	0.786	0.881	0.665
trash_tarp	1425	24	0.964	0.667	0.724	0.603
trash_unknown_instance	1425	550	0.77	0.82	0.834	0.547
trash_wreckage	1425	27	0.88	0.741	0.817	0.707

Figure 14. Precision, recall, and mAP metrics for the proposed AIT-YOLOv7.

These metrics, which are computed using Equation (4) for precision, Equation (5) for recall, and Equation (6) for mAP, demonstrate substantial improvements in performance due to the advanced imaging techniques. The integration of CBAM with MSTB and the Diffusion model significantly enhances the quality of underwater images, leading to better feature representation and more accurate object detection. The precision metric indicates a higher proportion of true positive predictions among all positive predictions, reflecting improved accuracy. The recall metric shows a greater proportion of true positive predictions among all actual positive instances, highlighting the model's enhanced capability to detect all relevant objects. The mAP metric, which evaluates accuracy and precision across various classes by computing Average Precision (AP) from the precision–recall curve, underscores the comprehensive effectiveness of the proposed enhancements in improving the detection and classification of underwater objects.

Figure 15 illustrates the stages of image processing and enhancement that lead to high-precision object detection and classification. Figure 15a presents the raw input image from the dataset. Figure 15b shows the image after applying the Convolutional Block Attention Module (CBAM) with the Modified Swin Transformer Block, enhancing image quality by preserving and improving crucial features and spatial information. Figure 15c depicts the image processed with the Diffusion Gaussian process, which further reduces noise and artifacts for clearer visual representations. Finally, Figure 15d displays the output where objects are detected and classified with high precision using enhanced techniques, demonstrating the effectiveness of the proposed AIT-YOLOv7 in improving underwater object detection and classification.

(**a**)

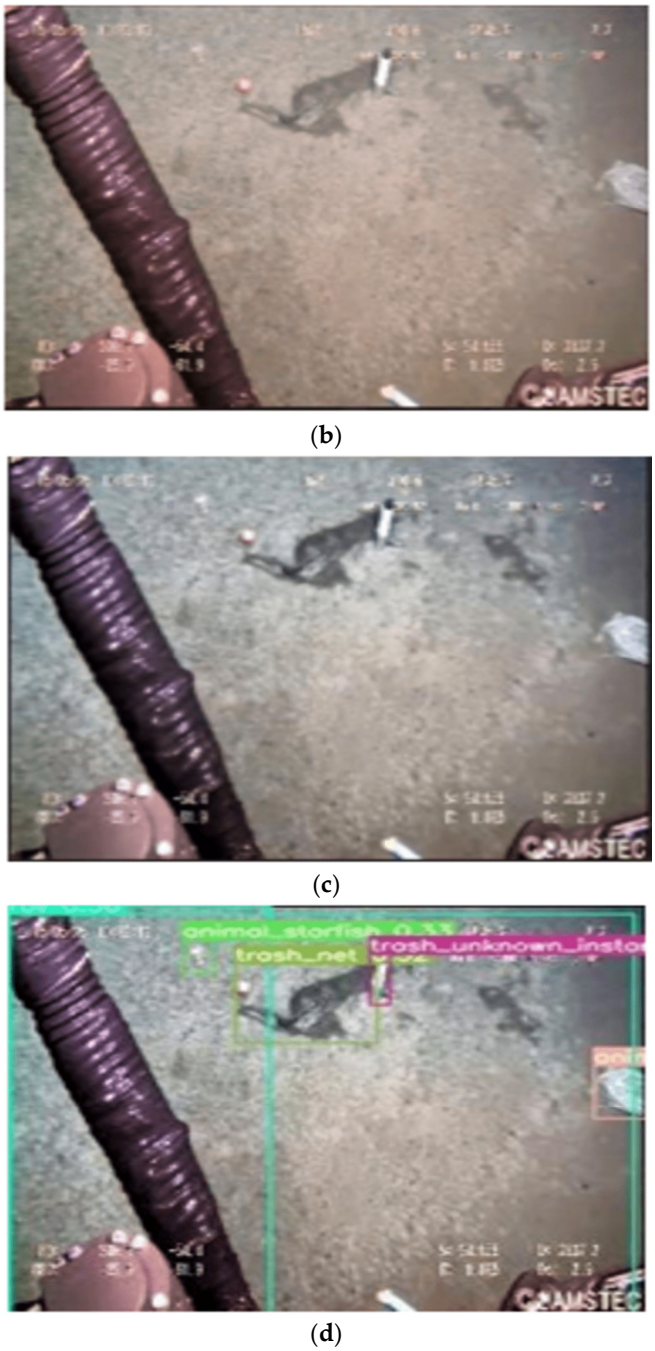


Figure 15. (**a**) Input image from the dataset. (**b**) Convolutional block attention with modified Swin Transformer Block. (**c**) Diffusion model. (**d**) Detected and classified with high precision.

Precision metric quantifies the fraction of true positives out of all detected objects, while recall measures the fraction of true positives out of all actual objects in the image. Mean Average Precision (mAP) score averages the precision and recall scores for each object class to determine the overall accuracy of the object detector.

The proposed AIT-YOLOv7 achieves mAP@.5 as 81.4% for all classes (22 categories) found in the TrashCan dataset. Table 3 shows a significant improvement in mAP@.5 for various trash categories, which clearly depicts the significant contribution of proposed AIT-YOLOv7 in trash removal for Sustainable Development Goal (SDG) 14: Life Below Water.

Table 3. Accuracy of AIT-YOLOv7 in different trash objects detection.

Sl.No.	Class	mAP@.5
1	Trash_bag	88.2%
2	Trash_bottle	89.4%
3	Trash_branch	93.5%
4	Trash_can	93%
5	Trash_clothing	93.5%
6	Trash_container	90.4%
7	Trash_cup	95.9%
8	Trash_net	78.6%
9	Trash_pipe	91.7%
10	Trash_rope	43.7%
11	Trash_snack_wrapper	88.1%
12	Trash_tarp	72.4%
13	Trash_unknown_instance	83.4%
14	Trash_wreckage	81.7%

A performance comparison of baseline YOLOv7 with the proposed AIT-YOLOv7 for the underwater object detection and classification for class ROV (remotely operated vehicle) out of a total of 22 classes found in the TrashCan dataset is shown in Table 4.

Table 4. Performance comparison of AIT-YOLOv7 with baseline YOLOv7 for ROV class.

Model	Images	Labels	mAP@.5
YOLOv7 (baseline)	1425	700	74.9%
AIT-YOLOv7 (ours)	1425	700	90.3%

There is an increase of 15.4% due to the integration of advanced imaging techniques (AIT) with YOLOv7. This showcases significant advancement in underwater object detection and classification, providing a valuable tool for marine research, underwater environmental monitoring, and conservation systems satisfying the SDG 14: Life Below Water.

5. Conclusions

Based on the comprehensive analysis and experimentation conducted in this study, several key conclusions can be drawn regarding the effectiveness of advanced imaging techniques for underwater object detection and classification. Firstly, leveraging state-of-the-art methodologies such as the Convolutional Block Attention Module (CBAM), Modified Swin Transformer Block (MSTB), and Diffusion model has demonstrated significant enhancements in image quality and object detection accuracy. These techniques collectively address challenges such as color distortion, noise, and low visibility inherent in underwater environments, thereby improving the clarity and fidelity of captured images. Secondly, the integration of these advanced techniques with established object detection and tracking frameworks like YOLOv7 and DeepSORT has proven to be highly effective. YOLOv7's rapid single-shot detection capability combined with DeepSORT's robust object tracking across frames creates a synergistic pipeline capable of real-time and accurate object detection and tracking in underwater scenarios. This integration enhances the system's ability to identify and classify marine objects and ensures continuity and precision in monitoring dynamic underwater environments. Moreover, the evaluation metrics, including precision, recall, and mean Average Precision (mAP), consistently showed superior performance after applying the proposed enhancements. The significant improve-

ments in mAP@.5 metric underscore the efficacy of the developed methodologies in handling complex underwater scenes and diverse object classes, crucial for applications in marine research, conservation, and environmental monitoring.

In conclusion, this research underscores the importance of advanced imaging techniques and their integration with robust deep-learning frameworks for advancing underwater object detection capabilities. Addressing these challenges aligns with Sustainable Development Goal (SDG) 14: Life Below Water, which aims to conserve and sustainably use the oceans, seas, and marine resources. Future research directions may focus on further refining these techniques, exploring additional datasets, and adapting the approach for broader underwater monitoring and conservation efforts. These advancements hold promise for addressing ongoing challenges in marine debris detection and ecosystem management, contributing to sustainable underwater resource utilization and preservation.

Author Contributions: Conceptualization, P.P. and G.C.; Methodology, P.P. and G.C.; Software, P.P.; Validation, G.C. and M.H.A.; Formal analysis, M.H.A.; Data curation, A.J.; Writing – original draft, P.P.; Writing—review & editing, M.H.A.; Visualization, A.J.; Supervision, G.C.; Project administration, A.J.; Funding acquisition, A.J. and M.H.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Springer: Cham, Switzerland, 2014; pp. 740–755.
2. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot MultiBox detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 21–37.
3. Redmon, J.; Farhadi, A. YOLOv3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
4. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. YOLOv4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
5. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475. https://doi.org/10.1109/CVPR52729.2023.00721.
6. Liu, B.; Yang, Y.; Zhao, M.; Hu, M. A Novel Lightweight Model for Underwater Image Enhancement. *Sensors* **2024**, *24*, 3070. https://doi.org/10.3390/s24103070.
7. Gong, T.; Zhang, M.; Zhou, Y.; Bai, H. Underwater Image Enhancement Based on Color Feature Fusion. *Electronics* **2023**, *12*, 4999. https://doi.org/10.3390/electronics12244999.
8. Sun, T.; Tang, Y.; Zhang, Z. Structural Information Reconstruction of Distorted Underwater Images Using Image Registration. *Appl. Sci.* **2020**, *10*, 5670. https://doi.org/10.3390/app10165670.
9. Yeh, C.-H.; Lai, Y.-W.; Lin, Y.-Y.; Chen, M.-J.; Wang, C.-C. Underwater Image Enhancement Based on Light Field-Guided Rendering Network. *J. Mar. Sci. Eng.* **2024**, *12*, 1217. https://doi.org/10.3390/jmse12071217.
10. Yang, J.; Huang, H.; Lin, F.; Gao, X.; Jin, J.; Zhang, B. Underwater Image Enhancement Fusion Method Guided by Salient Region Detection. *J. Mar. Sci. Eng.* **2024**, *12*, 1383. https://doi.org/10.3390/jmse12081383.
11. Lu, S.; Gua, F.; Zhang, H.; Lai, H. Underwater image enhancement method based on denoising diffusion probabilistic model. *J. Vis. Commun. Image Represent.* **2023**, *96*, 103926.
12. Wang, N.; Zhang, Z.; Hu, H.; Li, B.; Lei, J. Underground Defects Detection Based on GPR by Fusing Simple Linear Iterative Clustering Phash (SLIC-Phash) and Convolutional Block Attention Module (CBAM)-YOLOv8. *IEEE Access* **2024**, *12*, 25888–25905. https://doi.org/10.1109/ACCESS.2024.3365959.

13. Kim, H.; Yim, C. Swin Transformer Fusion Network for Image Quality Assessment. *IEEE Access* **2024**, *12*, 57741–57754. https://doi.org/10.1109/ACCESS.2024.3378092.
14. Tian, T.; Cheng, J.; Wu, D.; Li, Z. Lightweight underwater object detection based on image enhancement and multi-attention. *Multimed. Tools Appl.* **2024**, *83*, 63075–63093. https://doi.org/10.1007/s11042-023-18008-8.
15. Desilva, S.; Karthik, R.; DV, K.R.; Akilandeswari, J. A Deep Learning Framework for Detecting Underwater Trash. In Proceedings of the 2024 International Conference on Computing and Data Science (ICCDS), Chennai, India, 26–27 April 2024; pp. 1–6. https://doi.org/10.1109/ICCDS60734.2024.10560433.
16. Almutiry, O.; Iqbal, K.; Hussain, S.; Mahmood, A.; Dhahri, H. Underwater images contrast enhancement and its challenges: A survey. *Multimed. Tools Appl.* **2024**, *83*, 15125–15150. https://doi.org/10.1007/s11042-021-10626-4.
17. Yuan, S.; Li, Y.; Bao, F.; Xu, H.; Yang, Y.; Yan, Q.; Zhong, S.; Yin, H.; Xu, J.; Huang, Z.; et al. Marine environmental monitoring with unmanned vehicle platforms: Present applications and future prospects. *Sci. Total Environ.* **2023**, *858 Pt 1*, 159741.
18. Huy, D.Q.; Sadjoli, N.; Azam, A.B.; Elhadidi, B.; Cai, Y.; Seet, G. Object perception in underwater environments: A survey on sensors and sensing methodologies. *Ocean. Eng.* **2023**, *267*, 113202.
19. Zhou, J.; Yang, T.; Zhang, W. Underwater vision enhancement technologies: A comprehensive review, challenges, and recent trends. *Appl. Intell.* **2023**, *53*, 3594–3621. https://doi.org/10.1007/s10489-022-03767-y.
20. Zocco, F.; Lin, T.-C.; Huang, C.-I.; Wang, H.-C.; Khyam, M.O.; Van, M. Towards More Efficient EfficientDets and Real-Time Marine Debris Detection. *IEEE Robot. Autom. Lett.* **2023**, *8*, 2134–2141, April *2023*. https://doi.org/10.1109/LRA.2023.3245405.
21. Yang, Y.; Chen, L.; Zhang, J.; Long, L.; Wang, Z. UGC-YOLO: Underwater Environment Object Detection Based on YOLO with a Global Context Block. *J. Ocean Univ. China* **2023**, *22*, 665–674. https://doi.org/10.1007/s11802-023-5296-z.
22. Xin, H.; Li, L. Arbitrary Style Transfer with Fused Convolutional Block Attention Modules. *IEEE Access* **2023**, *11*, 44977–44988. https://doi.org/10.1109/ACCESS.2023.3273949.
23. Wang, X.; Xue, G.; Huang, S.; Liu, Y. Underwater Object Detection Algorithm Based on Adding Channel and Spatial Fusion Attention Mechanism. *J. Mar. Sci. Eng.* **2023**, *11*, 1116. https://doi.org/10.3390/jmse11061116.
24. Yang, G.; Liu, S.; Zhang, Y. An underwater image enhancement method based on Swin transformer. In Proceedings of the SPIE 12971, Third International Conference on Optics and Communication Technology (ICOCT 2023), Changchun, China, 15–17 September 2023; Volume 129710B. https://doi.org/10.1117/12.3017413.
25. Zhang, H.; He, R.; Fang, W. An Underwater Image Enhancement Method Based on Diffusion Model Using Dual-Layer Attention Mechanism. *Water* **2024**, *16*, 1813. https://doi.org/10.3390/w16131813.
26. Lu, S.; Guan, F.; Zhang, H.; Lai, H. Speed-Up DDPM for Real-Time Underwater Image Enhancement. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 3576–3588. https://doi.org/10.1109/TCSVT.2023.3314767.
27. Yeh, C.H.; Lin, C.H.; Kang, L.W.; Huang, C.H.; Lin, M.H.; Chang, C.Y.; Wang, C.C. Lightweight deep neural network for joint learning of underwater object detection and color conversion. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 6129–6143.
28. Zhou, W.; Zheng, F.; Yin, G.; Pang, Y.; Yi, J. YOLOTrashCan: A deep learning marine debris detection network. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 1–12.
29. Dhariwal, P.; Nichol, A. Diffusion models beat GANs on image synthesis. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 8780–8794.
30. Saleh, K.; Vámossy, Z. BBBD: Bounding Box Based Detector for Occlusion Detection and Order Recovery. *arXiv* **2022**, arXiv:2204.12841.
31. Teng, X.; Fei, Y.; He, K.; Lu, L. The Object Detection of Underwater Garbage with an Improved YOLOv5 Algorithm. In Proceedings of the 2022 International Conference on Pattern Recognition and Intelligent Systems, Wuhan, China, 29–31 July 2022; pp. 55–60.
32. Liu, H.; Song, P.; Ding, R. Towards domain generalization in underwater object detection. In Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 25–28 October 2020; pp. 1971–1975.
33. Sharma, P.; Bisht, I.; Sur, A. Wavelength-based attributed deep neural network for underwater image restoration. *ACM Trans. Multimed. Comput. Commun. Appl.* **2023**, *19*, 1–23.
34. Wang, Y.; Yu, J.; Zhang, J. Zero-shot image restoration using denoising diffusion null-space model. *arXiv* **2022**, arXiv:2212.00490.
35. Liu, Y.; Zhang, H.; Gao, D. DiffYOLO: Object Detection for AntiNoise via YOLO and Diffusion Models. *arXiv* **2024**, arXiv:2401.01659.
36. Zeng, L.; Sun, B.; Zhu, D. Underwater target detection based on Faster R-CNN and adversarial occlusion network. *Eng. Appl. Artif. Intell.* **2021**, *100*, 104–109.
37. Fan, B.; Chen, W.; Cong, Y.; Tian, J. Dual refinement underwater object detection network. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XX 16; Springer: Cham, Switzerland, 2020; pp. 275–291.
38. Jia, J.; Fu, M.; Liu, X.; Zheng, B. Underwater object detection based on improved EfficientDet. *Remote Sens.* **2022**, *14*, 4487.
39. Chen, L.; Zhou, F.; Wang, S.; Dong, J.; Li, N.; Ma, H.; Zhou, H. SWIPENET: Object detection in noisy underwater images. *arXiv* **2020**, arXiv:2010.10006.
40. Fayaz, S.; Parah, S.A.; Qureshi, G.J. Underwater object detection: Architectures and algorithms–a comprehensive review. *Multimed. Tools Appl.* **2022**, *81*, 20871–20916.

41. Wu, F.; Cai, Z.; Fan, S.; Song, R.; Wang, L.; Cai, W. Fish Target Detection in Underwater Blurred Scenes Based on Improved YOLOv5. *IEEE Access* **2023**, *11*, 122911–122925.
42. Hong, L.; Wang, X.; Zhang, G.; Zhao, M. USOD10K: A New Benchmark Dataset for Underwater Salient Object Detection. *IEEE Trans. Image Process.* **2023**. https://doi.org/10.1109/TIP.2023.3266163.
43. Deluxni, N.; Sudhakaran, P.; Kitmo; Ndiaye, M.F. A Review on Image Enhancement and Restoration Techniques for Underwater Optical Imaging Applications. *IEEE Access* **2023**, *11*, 111715–111737. https://doi.org/10.1109/ACCESS.2023.3322153.
44. Hong, J.; Michael, F.; Sattar, J. TrashCan: A Semantically-Segmented Dataset towards Visual Detection of Marine Debris. *arXiv* **2020**, arXiv:2007.08097. https://conservancy.umn.edu/handle/11299/214865.
45. Fulton, M.; Hong, J.; Islam, M.J.; Sattar, J. Robotic detection of marine litter using deep visual detection models. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA) (2019), Montreal, QC, Canada, 20–24 May 2019. https://doi.org/10.1109/ICRA.2019.8793975.