IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 25 December 2023, accepted 11 February 2024, date of publication 23 February 2024, date of current version 5 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3369063

APPLIED RESEARCH

HLASwin-T-ACoat-Net Based Underwater Object Detection

S. MANIMURUGAN[1], (Senior Member, IEEE), C. NARMATHA[1], (Member, IEEE), MAJED M. ABOROKBAH[1], (Member, IEEE), NAVEEN CHILAMKURTI[2], (Senior Member, IEEE), SUBRAMANIAM GANESAN[3], (Life Senior Member, IEEE), RAJENDRAN THAVASIMUTHU[4], (Member, IEEE), P. KARTHIKEYAN[5], AND M AMMAD UDDIN[1]

[1]Faculty of Computers and Information Technology, University of Tabuk, Tabuk 71491, Saudi Arabia
[2]Department of Computer Science and IT, La Trobe University, Melbourne, VIC 3086, Australia
[3]Department of Electrical and Computer Engineering, Oakland University, Rochester, MI 48309, USA
[4]Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, Tamil Nadu 600077, India
[5]School of Computer Science and Engineering, RV University, Bengaluru 560059, India

Corresponding authors: S. Manimurugan (mmurugan@ut.edu.sa) and P. Karthikeyan (karthikeyanp@rvu.edu.in)

This work was supported by the Deputyship for Research & Innovation, Ministry of Education, Saudi Arabia, under Project 249-1443-S.

ABSTRACT Due to the limited light penetration in underwater environments, sonar equipment plays a crucial role in various commercial and military operations. However, underwater images often suffer from degradation due to scattering and absorption phenomena, resulting in poor visibility of submerged objects. To address this challenge, image enhancement techniques are essential for enhancing the appearance and visibility of underwater objects. This research proposes a novel approach called HLAST-ACNet, which combines the advantages of a hybrid Local Acuity Swin Transformer and an Adapted Coat-Net for Underwater Object Detection (UOD). The HLASwin-T-ACoat-Net leverages Contrast Limited Adaptive Histogram Equalization (CLAHE) to increase the quality of images. Additionally, it incorporates a path aggregation network to integrate deep and shallow feature maps and utilizes online complicated example mining to improve training efficiency. Furthermore, the algorithm improves Region of Interest (ROI) pooling by introducing ROI alignment, which mitigates quantization errors and enhances object detection accuracy. Compared to existing algorithms, the algorithms based on HLASTACNet demonstrate significant improvements in the URPC2018 and OUC datasets, achieving precision rates of 91.25% and 92.36%, respectively. The research model has a higher computational complexity than four existing methods, as evidenced by its GFLOPs, per-image processing time with a speed of 20ms, and the FPS measures for average processed frames per second reaching 2.28s. The research model effectively addressed the challenges and false detection with varying sizes of objects in complicated underwater environments.

INDEX TERMS UOD, CLAHE, local acuity Swin transformer, ROI pooling and adapted coat-net.

I. INTRODUCTION

In the underwater realm, the restoration of physical entities poses a significant challenge, mainly due to the attenuation of light caused by water absorption and the unpredictable nature of illuminations resulting from the scattering medium [1], [2]. Various approaches were employed to capture visual representations of objects in environments with inadequate lighting, aiming to improve the visual clarity of images taken in underwater settings [3], [4], Polarimetric imaging systems were employed to address backscattering effects and achieve enhanced image resolution. While the application of polarization correction holds the potential to improve the reconstruction process, it simultaneously obstructs a

The associate editor coordinating the review of this manuscript and approving it for publication was Tallha Akram.

portion of incident light, leading to a reduction in the signal-to-noise ratio and introducing complexities in underwater object detection (UOD) imaging. The quality of underwater images is significantly affected by absorption and scattering, both exhibiting wavelength dependence.

This phenomenon inevitably leads to reduced visibility, contrast, and the emergence of color irregularities. These adverse effects pose substantial limitations on the practical utility of underwater images and videos in fields such as marine biology, archaeology, and ecology. Consequently, various algorithms for underwater images enhancement (UIE) serve as an initial pre-processing phase for UOD operations. This aims to enhance the identification accuracy of the system by improving the overall quality of images [5]. Although with the extensive efforts, thorough investigation, and perspicacious analysis of the comparison among UIE and UOD tasks, the insufficient situation persists in this field, owing to the absence of open access UOD image data sets, equipped with reference images and bounding boxes annotations, i.e., images unrestricted by underwater regression. Given the lack of reference images, a prior study [6] solely examined the impact of UIE algorithms on UOD tasks through an assessment of the interplay between the non-reference image quality evaluation metrics [7] and detection accuracy. Nevertheless, metrics to utilize the image quality without utilizing a reference image only provide a limited insight into the image's attributes and their correlation with human perceptual experience is not always coherent, as demonstrated in previous research [8].

An extensive inquiry into the interrelationship between two given tasks necessitates an examination of the correlation among the full-reference images and identification accuracy metrics [9]. These metrics possess the capability to evaluate various characteristics of the quality of an image, including but not limited to its textures, colors, structures, and contents. Though, the utilization of reference image is deemed imperative in conducting comprehensive assessments of image quality through the full-reference approach. Numerous research endeavors on the detection of submerged objects have substantially contributed to a wide range of ecological applications. The generic methodologies formulated are advantageous in the identification of entities in complex environments. The field of deep learning, an area of emerging research, has demonstrated notable efficacy in image processing tasks, with examples including image recognition, object detection, and person pose estimation [10], [11].

In the integration of neural networks with Signal Processing Information (SPI) systems, two modalities exist for obtaining the ultimate target images. One leverages the resilient neural networks to restore the integrity of the object images, while the another one employs the neural network for training and forecasting images after the process of reconstruction. The utilization of side-scan and forward-looking sonars has become ubiquitous in the field of oceanography as they are invaluable imaging systems employed for capturing expansive seafloor images. Their usage has amplified exponentially owing to their increased deployment of autonomous underwater vehicles. The extraction of quantitative information from images resulting from said processes poses a considerable challenge, primarily concerning the detection and extraction of information about the objects contained therein. The recent progressions in the field of machine learning have facilitated the automated identification process, which can be efficiently utilized in real-time scenarios. Nevertheless, the tool's capacity to effectively manage dynamic backgrounds is limited. The requirement for the creation of efficient schemes for underwater detection that are appropriate for difficult scenarios is evident.

As discussed earlier, the standard of subaquatic images is based upon the aquatic surroundings. Consequently, the efficacious application of the deep learning methodology is contingent upon the acquisition of an ample quantity of sonar images across diverse aquatic environments as well as varying operational parameters. UOD faces unique challenges due to the specific characteristics of underwater images, such as low visibility, colour irregularities, and background complexities. Traditional approaches, including polarimetric imaging systems, have shown some improvements in enhancing underwater image quality, but they still have limitations in handling dynamic backgrounds and achieving robust detection performance. The present investigation employed deep learning (DL) methodology utilizing the URPC2018 and OUC datasets.

The main contributions of the research paper are summarized as follows.

- The research paper introduces a novel approach for UOD in challenging underwater frames. After restoring the images, they are subjected to the LASwin-T model, which is tailored for handling the intricacies of underwater environments. This contribution aids in addressing the complexities of underwater scenes and advances the state-of-the-art in UOD.
- Another noteworthy contribution involves the integration of a path aggregation network, serving as a mechanism to merge deep and shallow feature maps. Through this integration, the research paper enhances the representation of image features, resulting in a more comprehensive and informative feature map that significantly contributes to improved detection performance.
- The effectiveness of the proposed UOD method is rigorously examined through a series of comprehensive experiments. The evaluation results underscore the superiority of the system over existing state-of-the-art methods, both in terms of detection accuracy and overall performance.

The remaining of the article is organized as follows. Section II covers existing DL UOD methods. Section III outlines the proposed UOD framework with CLAHE, Coat-Net, and local Swin transformer-based schemes. Section IV presents the

experimentations, dataset, results, and analysis. Finally, conclusions and future work are discussed in Section V.

II. RELATED WORKS

Abu and Diamant [12] utilized a Support Vector Machine (SVM) to distinguish between background and shadow-related pixels by extracting statistical characteristics from the pixels in the ROI. This algorithm's strength lies in its ability to display resilience as a result of its primary parameters being configured on the spot. Furthermore, the algorithm has a broad scope as it can be utilized across various forms of sonar detection without requiring any prior details regarding the dimensions or arrangement of the object being detected. To evaluate how well the detection works, this research employed a unique self-governing underwater device to capture a total of 270 sonar images, which are also being shared with the broader public. A detection strategy for pond-raised river crabs was introduced by Ji et al. [13], which prioritizes high accuracy and swift results. The successful identification of the underwater river crab target can be achieved through the utilization of both the MobileCenterNet model and a multi-scale pyramid fusion image enhancement technique. The suggested strategy involves applying a multi-level pyramid merging method to improve the quality of crab images captured underwater, where blurriness and uneven lighting are common issues. By utilizing CLAHE to boost contrasts and Underwater Dark Channels Prior (UDCP) to eliminate haze, the image's overall quality is enhanced. Moreover, a MobileCenterNet model-based approach to identify crab targets is presented. This approach not only achieves simplicity and agility but also focuses the model's sensitivity on relevant characteristics specific to crabs. The Features Fusion Model (FFM) was developed for extracting multi-scale feature maps data, as explained in this study's description. In addition, the utilization of Atrous Spatial Pyramids Pooling (ASPP) enables the integration of diverse context information from separate receptive fields. Based on the experiment's results, it can be inferred that MobileCenterNet displays moderate-precision scores and F1 values, both reaching 97.86% and 97.94%, respectively. Moreover, the dimensions of the model are compact at just 24.46 M, and it can detect with a rapid speed of 48.18 frames per second.

Hua et al. [14] have introduced a new type of subaquatic object detector that employs YOLOv5s. At first, a module was created to improve or decrease various hierarchical properties in a controlled way and reduce the interference of acoustic signals from intricate underwater environments during feature fusion. A novel approach called FMSPP involving a rapidly mixing pool layer of equal size has been proposed, aimed at spatiotemporal pyramid pooling. This architecture can improve the descriptive abilities of texture and contour features in a network, reducing the parameters and consequently augmenting the network's general performance and accuracy in classifying. The effectiveness of the suggested method is validated by conducting ablation and multi-method comparison experiments on both the URPC and DUT-USEG datasets. The detector showcased in this research exhibits notable advantages in detecting precision and productivity when compared to existing detectors.

Panda and Nanda [15] introduced the SKDE model, which utilizes a method of spatial kernel density estimation for learning within the SKDE feature space. Creating a model of the surroundings and gaining an understanding of it largely involves the use of individual pixels in an analytical approach. Using the histogram representation, the model histograms gain an understanding of the new pixel. The current research has presented an innovative method for learning models and categorizing pixels, which employs a similarity measurement technique based on correntropy. Camera model parameters are determined by utilizing a 2D optimization approach that employs highly accurate corner features of an object. These features are measured to subpixel precision. The use of subpixel characteristics in the pipelining system enables the smooth and accurate determination of model parameters. The model's estimated parameters are used as a tool to convert the input frame, which is then employed for both acquiring and organizing the model. The suggested plan has undergone experimental testing by using six different datasets of video frames captured underwater to validate its effectiveness.

Yeh et al. [16] put forward a compact DL model for identifying objects in images captured underwater. The network was intentionally created to learn both colour conversion and object detection simultaneously, with a focus on underwater environments. The main aim of the module concerned with converting image colors is to solve the problem of colour absorption underwater by transforming coloured images into grayscale ones. The objective is to maximize the identifying object's accuracy while limiting the overall computational resources required. The experiment's results indicate that using the lightweight collaborative learning model on the Raspberry Pi platform is a remarkably effective approach for detecting underwater objects, especially when compared to current methods. In their study, Sun et al. [17] presented a convolutional neural network (CNN) based knowledge transfer framework to recognize objects underwater. This framework aimed to address the challenge of extracting distinctive features from images that possess comparatively low contrast. Despite the inadequacy of the training dataset, the transfer framework can effectively acquire a recognition model tailored to the specific task of underwater object recognition, through the aid of data augmentation. To enhance the identification of objects in underwater videos, this study proposes the implementation of a weighted probabilities decision mechanism that is capable of identifying objects across multiple frames. To verify the efficacy of the approach, the evaluations were computed on a publicly available dataset. The findings indicate that the method put forth demonstrates favorable outcomes for recognizing objects in underwater environments based on the analyses of the test image datasets and underwater videos.

Pan et al. [18] used a new CNN design to train video frames captured underwater. The current methodology revolves around a modified version of ResNet, a neural network, which has been customized to ease the process of identifying submerged objects. The M-ResNet is an advanced method that effectively boosts productivity by integrating multi-scale strategies for accurately detecting objects of different sizes, especially those that are relatively smaller. The results gathered from the trial showcase that the model achieved a recognition accuracy rate of 96.5% (mean average precision), indicating its effectiveness. From now on, there is a recently suggested gadget that can independently recognize items primarily in aquatic environments.

Li et al. [19] developed modified YOLOv8 DL model aimed at detecting fish. To assess the precision and effectiveness of the models, YOLOv8 becomes better equipped to handle occlusion-related issues. To optimize the model's training on the occlusion dataset, employed a dedicated loss function designed for such scenarios Repulsion Loss. This tailored approach contributes to the improved performance of the model in addressing occlusion challenges during fish detection. These approaches were extremely lightweight when correlated to updated versions, yet they exhibit similar levels of performance.

An improved YOLOv5 model was designed in [20] for detecting underwater target wakes in multi-source images. The model, enhanced with linear feature detection, distinguished between underwater and surface targets, as well as optical and infrared images. By optimizing the feature layer through parameter and image space conversion, the proposed model demonstrated superior performance over the original YOLOv5 in experimental results. An enhanced Faster RCNN model was developed for UOD. The VGG16 structure was replaced with Res2Net101 in the backbone network, improving expressive ability. Online Hard Example Mining (OHEM) addressed sample imbalance, and Generalized Intersections Over Union (GIOU) with Soft Non-Maximum Suppressions (Soft-NMS) optimized bounding box regression. Multi-scale training enhanced model robustness, with results confirming the method's effectiveness in UOD [21]. The research in [22] enhanced the YOLOv5 model and its subset version through training on diverse data sets characterized by varying levels of image qualities. The hyperparameters during the extraction of features phase were initially set using momentum and learning rate, and subsequently refined through the utilization of the ADAM optimizer and the implementation of a reducing-learning-rate-on-plateau function. The optimized YOLOv5 models demonstrated improved performance, enhancing the precision of UOD.

EfficientDet-Revised, an improved model for UOD was implemented in [23]. Changes included adding the Channel Shuffle module to the MBConvBlock for enhanced information exchange, replacing the attention modules fully connected (FC) layer with convolution to reduce parameters, and introducing an Enhanced features extraction module for multi-scale features fusion. Results showed that EDR outperformed other algorithms in detection efficiency. An enhanced small target detection algorithm based on YOLOv7 was developed for underwater scenarios. The approach improved detection accuracy by concentrating on crucial features of small targets, reducing model complexity through the integration of the SENet attention mechanism, enhanced FPN network topology, and the EIoU loss function. Results showed superior performance compared to other networks, achieving heightened detection accuracy on the test set [24]. A two-stage UOD network based on Faster R-CNN, leveraging the Swin Transformer as the backbone was developed. It enhanced feature fusion with a path aggregation network and improved training efficiency through online hard example mining. Additionally, ROI pooling was upgraded to ROI alignment, maximizing identification performances, and reducing quantization errors. Experiments demonstrated enhanced detection outcomes in complex underwater environments for the enhanced FR-CNN model [25].

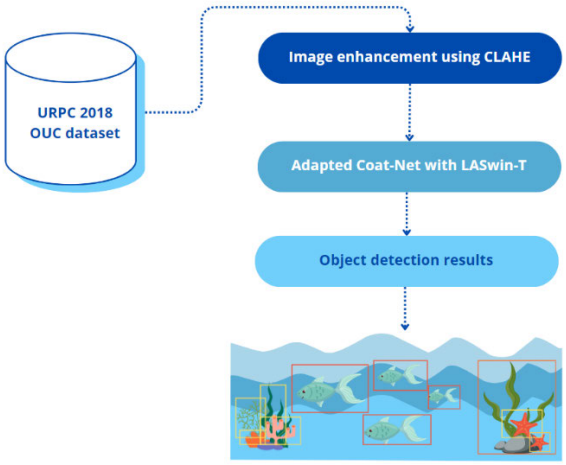


FIGURE 1. Overview of proposed UOD based hybrid deep learning method.

The existing research landscape in UOD employs a variety of ML and DL models. Approaches range from traditional methods like SVM and CNN to more recent advancements such as YOLOv5, EfficientDet, and Swin Transformer-based models. Researchers have addressed challenges such as low contrast, shadow-related pixels, and the need for accurate detection in complex underwater environments. Despite the progress, several research gaps persist. The literature shows a diversity of models and methodologies, but there is a lack of standardized evaluation metrics and benchmark datasets, making it challenging to compare the performance of different models comprehensively. Additionally, many studies focus on specific marine organisms or objects, leaving room for the development of more generalized and versatile UOD models. Furthermore, there is limited exploration of real-world applications and deployment scenarios, raising

TABLE 1. Research gap of existing UOD methods based on ML and DL.

METHODS	ADVANTAGES	DISADVANTAGES
SVM [12]	This method detects shadows using a likelihood ratio test.	A shorter sampling interval means more accuracy but also more complexity.
MobileCenterNet [13]	This method is adept at learning and precise in output.	Manual selection and extraction lead to poor robustness and low efficiency.
YOLOv5s [14]	The method can handle complex underwater backgrounds.	Small object detection remains problematic with missed and false detections.
SKDE [15]	Use SKDE feature space to handle background complexity.	Underwater conditions hinder the detection of small objects.
Lightweight Deep Neural Network [16]	This method overcomes data-hungry issues.	Underwater images degrade due to lighting and low-quality optical devices.
CNN [17]	This method uses transfer learning to address small underwater training data.	Low-light and high-noise in underwater imaging are challenging.
ResNet [18]	This lowers underwater inspection costs.	Underwater videos often have low resolution and saturation, which limits object recognition.
YOLOv8 [19]	This method achieves illumination invariant and overcomes low-quality video challenges.	Many things to avoid in the Fish4knowledge dataset.

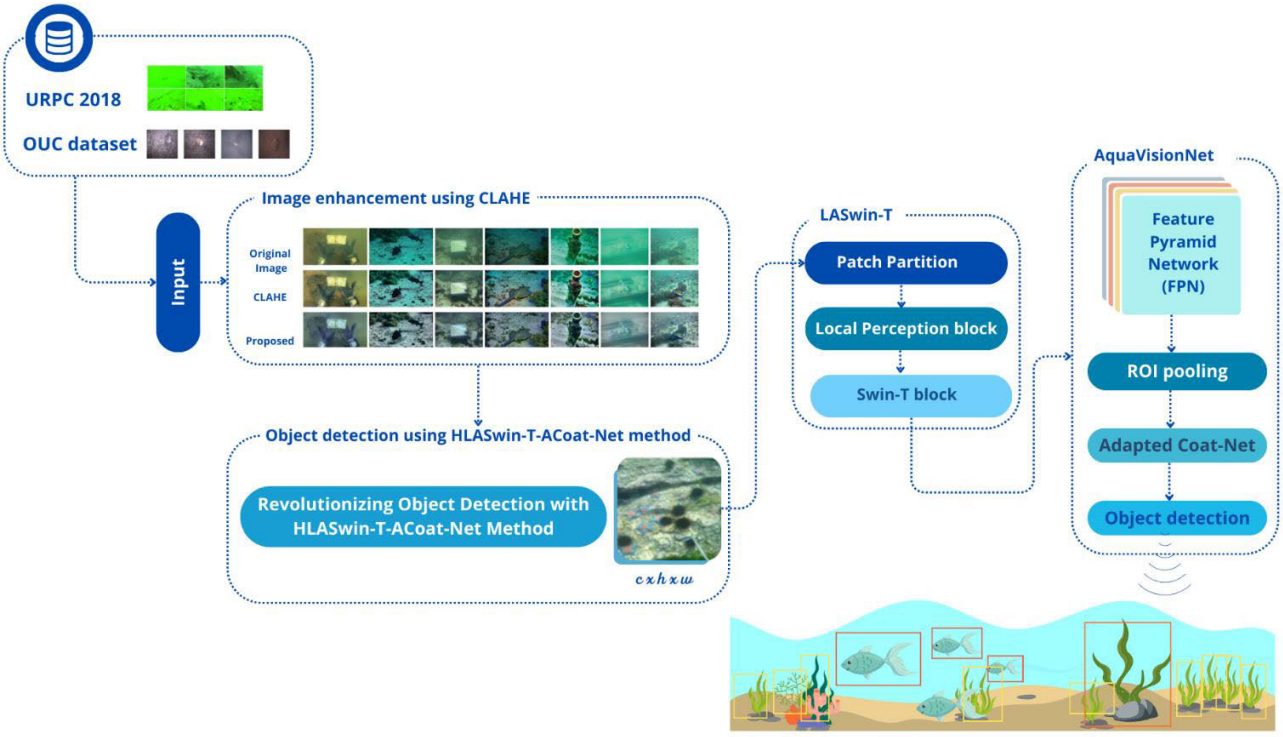


FIGURE 2. Proposed UOD based on HLAST-ACNet.

questions about the scalability and adaptability of these models in practical underwater settings. Table 1 presents the research gap of existing UOD methods. The problem statement in UOD research revolves around the need for improved detection accuracy, particularly in challenging underwater conditions. Existing models often grapple with issues related to feature extraction, sample imbalance, and robustness in varied environments. Moreover, the lack of standardized evaluation metrics impedes the establishment of a benchmark for UOD models. Addressing these challenges is crucial for advancing UOD capabilities and facilitating the deployment of reliable and efficient underwater detection systems in real-world scenarios.

III. PROPOSED UOD BASED HYBRID DEEP LEARNING METHOD

The automated methodology presented in this study is designed to identify multiple occurrences of fish presence in

underwater images. Frames extracted from underwater videos often encounter adverse effects such as blurring, illumination diffraction, occlusions, and various forms of deterioration, presenting challenges for accurate object identification. The proposed scheme for optimal detection of submerged objects comprises three distinct modules. Figure 1 illustrates the overview of the research model. The primary objective of the initial data pre-processing module is to address colour degradation and geometric distortions within the input frames. The results of the CLAHE are utilized by the transfer learning approach using HLAST-ACNet to produce bounding boxes for the targeted object.

A. INPUT IMAGE COLLECTION

This portion of the manuscript endeavors to establish an extensive underwater dataset to facilitate the examination about the proposed approach impacts the phenomenon of UOD for researchers. The URPC2018 dataset comprises 2,901 pictures designated to train the model and 800 images to test the model. The images presented in the dataset bear varying resolutions, namely 586 × 480, 704 × 576, 720 × 405, and 1, 920 × 1.080. The annotations about the test set are currently unavailable. Furthermore, several images were gathered from an artificial underwater habitat. Figure 2 depicts the proposed UOD based on HLAST-ACNet.

The OUC-VISION dataset offers annotated images of aquatic environments along with bounding box annotations for further analysis and research purposes. The present dataset encompasses a collection of 4,400 photographic images of submerged environments, which have been obtained under varying illuminative conditions through the utilization of a bespoke lighting apparatus. Furthermore, the present study has simulated three levels of turbidity variability, namely, clarity, moderate and high turbidity, through the incorporation of soil particles into the water sample. Consequently, the aquatic imagery captured by OUC-VISION exhibits a multitude of fluctuating illumination levels and turbidity variations. The resolution of the images is 486 × 648 pixels. The URPC 2018 and OUC dataset's raw underwater images extracted from the OUC-VISION dataset are presented through examples depicted in Figure 3 and Figure 4.

FIGURE 3. The input images of the URPC 2018 dataset.

B. IMAGE ENHANCEMENT USING CLAHE

The image histogram serves as a graphical representation of the intensity values present throughout an image. The principal purpose of a histogram is to provide statistical data

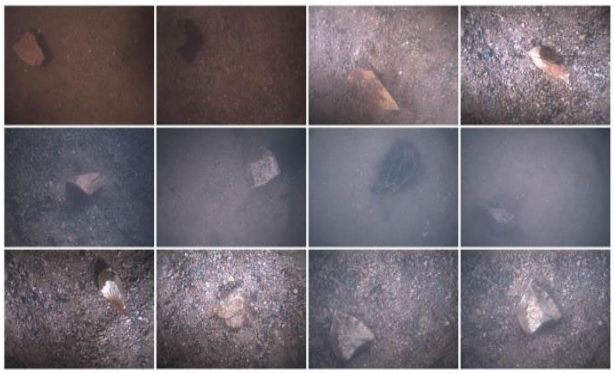


FIGURE 4. Samples of the raw images in the OUC dataset.

regarding an image. For this reason, the manipulation of the histogram can be employed as a means of conducting image enhancement. The histogram equalization technique is widely used in image enhancement owing to its simplicity and low computational burden. The present investigation utilizes the CLAHE technique to enhance colour retinal imagery. The present technique of enhancement finds its widespread application in the field of UOD, where the notable characteristics include, inter alia, object contrast. The contrast of an image is determined by the interplay between the extent of intensity values present and the distinction between the highest and lowest pixel values.

The primary aim of utilizing histogram manipulation for image enhancement involves achieving a consistent distribution of intensity throughout the image. The visual representation featuring limited contrast possesses a constrained effective range of intensity. Histogram equalization is an image processing technique that facilitates the spread of intensity distribution across an image and helps to adjust the intensity values of the original image. The schematic representation of the method for enhancing colour retinal images, as postulated in this study, is depicted in Figure 5.

The methodology utilized in this study involves using retinal images input in RGB color format. The initial step involves partitioning the colour image into separate channels, thereby generating three distinct images, each representing the green (G), red (R), and blue (B) colour channel. The image enhancement procedure utilizing CLAHE is exclusively performed on the G channel, on account of the significant blood vessel structural data present within that channel when compared to the others. The image in the G channel with improved clarity and definition. The subsequent step entails the conflation of the three-image channels, specifically the red channel, the enhanced green channel, and another red channel. Upon completion of the procedure, the improved subaquatic imagery for both the URPC 2018 and OUC dataset was attained.

C. UOD USING HLASWIN-T-ACOAT-NET METHOD

A hybrid framework comprising three steps is recommended for UOD, including the implementation of CLAHE for

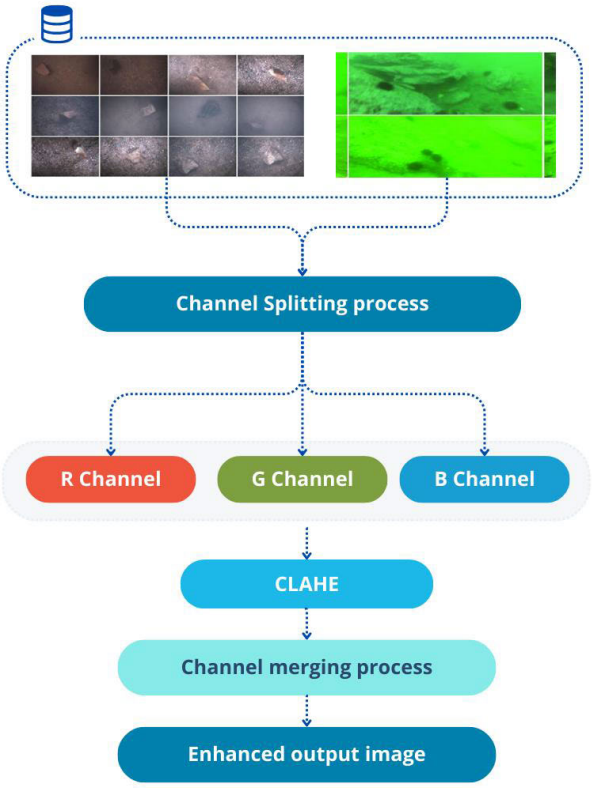


FIGURE 5. Underwater image enhancement flowchart using CLAHE.

pre-processing, training of HLAST-ACNet, and the use of the predictive HLAST-ACNet model for prediction.

LASwin-T Model: The Swin transformer has four different versions of the model: Swin-T, Swin-S, Swin-B, and Swin-L [27]. Given the distinctive features and intricate computational nature of remote sensing images, the present study presents Swin-T as a solution. Smartly paraphrased: The patch partition layer divides an initial RGB image into separate, non-overlapping sections, with each section containing a varying number of blocks per stage - 2, 2, 6, and 2. Every segment is considered a "token" and its attributes are established as a merging of the natural RGB values of the pixels. The Swin transformer comprises four phases designed to generate varying quantities of tokens.

If an underwater image, X, is available with a dimension of $\hbar \times \mathfrak{w}$, a single token comprises a concatenated raw pixel vector of an RGB image patch of 4×4. This token is subjected to a linear embedding technique, which transforms it into a vector of dimension c. Tokens are generated in different quantities at stages 1, 2, 3, and 4, which are $\frac{\hbar}{4} \times \frac{\mathfrak{w}}{4}$, $\frac{\hbar}{8} \times \frac{\mathfrak{w}}{8}$, $\frac{\hbar}{16} \times \frac{\mathfrak{w}}{16}$, and $\frac{\hbar}{32} \times \frac{\mathfrak{w}}{32}$, respectively. Every phase includes a merging segment for patches, comprised of a layer partitioning patches and a layer embedding linearly, a block for local perception, and several transformer blocks of Swin.

The fundamental component of the Swin transformer algorithm is the Swin transformer block. The block comprises three components namely, SMSA for space multi-head self-attention, LSMSA for lifted spaces multi-head self-attention, and MLP for multilayer perceptron. Integrating a layer normalization (LN) layer within the module enhances training stability while employing residual connections after each one. Equations (1) to (4) represent the LASwin-T model.

$$\hat{X}^l = SMSA\left(\mathcal{LN}\left(X^{l-1}\right)\right) + X^{l-1} \tag{1}$$

$$X^l = MLP\left(\mathcal{LN}\left(X^l\right)\right) + \hat{X}^l \tag{2}$$

$$\hat{X}^{l+1} = LSMSA\left(\mathcal{LN}\left(X^l\right)\right)\left(\mathcal{LN}\left(X^l\right)\right) + X^l \tag{3}$$

$$X^{l+1} = MLP\left(\mathcal{LN}\left(\hat{X}^{l+1}\right)\right) + \hat{X}^{l+1} \tag{4}$$

Local Acuity Block (LAB): Detecting local correlation and structural information of an image can be a difficult task for position encoding in a transformer. Even though the Swin transformer's hierarchical structure includes sequential layers with a shift window scheme, it does not effectively encode a significant amount of spatial context information. To address this issue, the team put forth a suggestion for the introduction of a local perception block (LAB) to be positioned at the start of the Swin transformer block. The Swin transformer utilizes vector-based data flow as opposed to traditional CNNs that operate with feature maps. As a result, in LPB, a group of vector features is initially restructured to form a spatial feature map. One way to rephrase this text could be: "To illustrate, the token ($\mathfrak{b}$, $\hbar * \mathfrak{w}$, c) is transformed into a feature map with dimensions ($\mathfrak{b}$, $\hbar * \mathfrak{w}$, c)."

To enhance the extraction of localized spatial characteristics while maintaining a sufficiently broad receptive field, a residual connection is employed alongside 3×3 dilated convolution layers with a dilation of 2 and a GELU activation function. Afterwards, the feature map is transformed into a (b,c,h,w) shape and forwarded to the Swin transformer block. The use of dilated convolution enhances the receptive field of a spatial image, enabling effective coding of contextual information across various scales. As a result, a broad range of contextual details can be accurately captured. In [28] introduced the concept of dilated convolution that facilitates the extension of the receptive field when juxtaposed with the conventional convolution operation. One important point to consider is that the conventional 3×3 convolutions have a field of 3×3. When the kernel size remains the same for the dilated convolutions with dilation factors of 2, the receptive fields are increased to 7×7. Thus, the usage of dilated convolution enables the expansion of the relevant area without compromising the feature details.

ACoat-Net method: Translation equivariance is a notable attribute of convolution layers, providing a valuable advantage by imparting a robust inductive bias. This quality becomes particularly crucial for enhancing model generalization when dealing with limited training datasets and proves beneficial for handling unseen data sets. The convolution operation for an input (x) at position (i) was expressed

mathematically through Equation (5).

$$y_i \sum_{j \in \mathcal{LR}(i)} w \odot xi \tag{5}$$

In this context, y_i represents the convolution output, with $\mathcal{LR}(i)$ denoting the local receptive fields. Practically designed for datasets in sequential, transformers, such as the Vision Transformers (ViT), have demonstrated superior capability compared to CNN approaches. Transformers-based DL models leverage the self-attention (SA) layer characterized by the global receptive (GR) fields. A key distinction among self-attention and convolution layers lies in the receptive field's size, as self-attention layers boast the global receptive fields, offering more extensive contextual data. Additionally, self-attention layers incorporate the input-adaptive weighting strategy, contributing to the higher model capacity of transformers-based models, particularly advantageous for large data sets. It is important to acknowledge the trade-offs between receptive field size and computational complexities. The self-attention strategy was formally described using Equation (6).

$$y_i = \sum_{j \in \mathfrak{gs}} \frac{DAW}{\sum_{k \in \mathfrak{gs}} DAW} \tag{6}$$

where $\mathfrak{gs}$ represents the global spatial space and $DAW = \sum_{k \in g} e_i^{x^T} x_k$ represents the dynamic attention weight. The ACoat-Net draws inspiration from the Coat-Net approach, which integrates both transformer and convolution layers. The primary objective of ACoat-Net was to enhance the performances of Bounding Box Detection (BDD) by leveraging the strengths of both self-attention and convolution layers. Consequently, ACoat-Net seeks to utilize the convolution blocks to improve generalizations while employing the self-attention layers to augment the model capability. Straightforward approaches for combining convolutions and SA layers involve the addition of the global static convolution kernels to the adaptive attention matrix, as illustrated in Equation (7).

$$y_i = \sum_{j \in \mathfrak{gs}} \frac{DAW * e^{(w_{i-j})}}{DAW * e^{(w_{i-k})}} X_j \tag{7}$$

One important point to highlight is that this form is a variation of the self-attention mechanism referred to as relative self-attention that solely concentrates on relative distance or position. Integrating attention and convolution layers directly results in a notable rise in computational complexity. The ACoat-Net framework consists of a duo of convolutional units, along with three two-dimensional-relative attention sections and three FNN modules. Additionally, the classification head is composed of a FC, global average pooling, and the SoftMax layers. Notably, two key enhancements distinguish this framework from the original Coat-Net:

- The incorporation of asymmetric convolution structures with varying sizes of kernel, and
- The utilization of depth-wise separable convolution to enhance network effectiveness and reduce model parameters.

Convolution Blocks: Serving as the initial segment of the proposed framework, the convolution block is designed for in-depth feature extraction. Within this block, convolution layers play a crucial role in extracting useful higher-level attributes from the input data set. The employed convolution layers encompass three types: the standard convolutional layers with the 3×3 kernel size, the asymmetric convolution structures with (1×3 and 3×1) kernel sizes, and the depth-wise separable convolutions (DWSC). The DWSC layer consists of the depth-wise convolutional layers and a point-wise kernel convolutions with kernel size.

FIGURE 6. RFFN Structure.

Residual Multi-Head Self-Attention Model: A crucial element of the transformer widely applied in various signals, images, or language processing tasks is the SA layer. This model incorporates two-dimensional relative positions encryption. Initially, the input feature map undergoes layer normalization. Subsequently, the pooling layer was employed for reducing the spatial dimensions of the feature maps. The processed feature maps are then directed to the relative SA model. At last, the convolutional layers are applied for additional explorations prior to the resulting feature maps are added to the outcome of the attention model. In the case of the feature maps with dimensions ($\mathfrak{h} \times \mathfrak{w}$), the relative position encryption involves the learnable parameters (LP) with dimensions $(2\mathfrak{h} - 1) \times (2\mathfrak{w} - 1)$.

Residual FFN (RFFN) Model: The design of the RFFN module mirrors the MLP head block in the ViT model. As illustrated in Figure 6, this block comprises like: layer normalizations, convolutional layer (1×1), Gaussian Error Linear Units (GELU) activation, convolutional layer (1×1), GELU, and summation with the inputs.

Model Training: The iterative acquisition of model parameters employs an optimizer, adjusting parameters through back-propagation at each step to minimize discrepancies

between model output and actual values. This involves training the model with sample data, evaluating performance by assessing the loss functions on a validation data set, and ultimately gauging model effectiveness on a testing dataset. Swin, an acronym for Shifted window (SwinT) [30], generates hierarchical feature map by integrating image patches in deep layer. ViT, a novel transformer-based framework, incorporates spatial dimension conversion within its structure. The CCT employs the convolution tokenizer to generate more intricate token and retain local data.

D. ACoat-Net MODEL EXPLAIN ABILITY

DL models often yield promising outcomes, yet their internal mechanisms are frequently perceived as opaque, earning them the label of "black box" methods. Grad-Cam represents a prevalent visual explanation technique employed with deep learning models to enhance the interpretability of their predictions. Its purpose is to visually illustrate the model's predictions. The initial step involves computing the gradient of the scores y and c for all classes, with respect to the feature map (f_k) of the specific layer.

$$g_c\left(f^k\right) = \frac{\partial y^c}{\partial f^k} \tag{8}$$

In this context, k denotes the channel indexes. Subsequently, the gradients undergo global averages pooling to assess significance of the weights $(a_k) f^k$ for class c in each channel.

$$a_k^c = \frac{1}{w_f \times h_f} \sum\nolimits_{i=1}^{w_f} \sum\nolimits_{j=1}^{h_f} \frac{\partial y^c}{\partial f_{i,j}^k} \tag{9}$$

Here, h_f and w_f represent the height and weight of feature map. The conclusive Grad-Cam heat maps $H_{Grad-cam}^c$ was the weighted summation of the feature map, trailed by the application of a rectified linear unit (ReLU) function $H_{Grad-cam}^c = ReLU\left(\sum_k a_k^c f_k\right)$.

ROI Pooling: The term ROI pertains to the candidate blocks within the features architecture. In the FR-CNN approach, candidate blocks were produced through the Region Proposal Networks, and these are subsequently mapped onto the feature architecture to obtain the ROI. ROI pooling is a process for extracting tiny feature maps from the ROI, involving the following operations:

- Mapping the ROI to the related region positions on the feature maps.
- Ensuring that ROIs of varying size are uniformly resized to the fixed dimension of the N × N.
- Dividing the ROI evenly into the N × N region.
- Extracting the higher pixel values from all divided regions, essentially performing the max pooling operations on all regions to serve as the 'representatives.' This ensures that all ROIs attains a consistent size of N × N.

Nevertheless, the employment of this technique will result in a decrement in precision because of the presence of rounding errors. To clarify, will provide an example in the following manner. Let us consider the backbone networks of the model has the stride of 16, which means that the extracted image is only 1/16 the size of the actual images. If the actual image was 400 × 400 and the last layer's feature maps is 25 × 25, then ROI pooling will result in a fixed 5 × 5 feature map. The original image contains a proposal area, measuring 200 × 200, which translates to a feature map size of 12.5 × 12.5 (200 divided by 16 multiplied by 200.16). After executing the process of rounding, the dimensions of the features mapped are now 12 × 12, referred to as the initial quantization function.

To achieve a consistent 5 × 5 size for the feature map, the 12 × 12 regions proposal obtained earlier is subdivided into 25 equally-sized smaller regions, each measuring 2.4 × 2.4 (12 divided by 5 and 12.5, respectively). During this stage, a mathematical operation known as the second quantization is carried out, resulting in the reduction of the tiny region to 2 × 2. Following these procedures, the resulting candidate box exhibits the certain level of displacement from its initial placement as determined by RPN. This alteration has a detrimental impact on the detection precision, particularly for smaller objects. In this endeavor, the utilization of ROI align was opted as a solution to prevent the issue, in contrast to implementing rough ROI pooling. The technique utilized for pooling ROI is distinct as it involves converting it into a constant process using regional characteristic accumulation instead of quantification and pooling. Eliminate the two quantization processes involved in ROI pooling and replace them with direct floating-point computations. The initial output size measures 12.5 × 12.5 while the second output size measures 2.4 × 2.4. Simultaneously, a hyperparameter is established that denotes the number of sampling points within each region. The number of points extracted from each region to compute the area's 'representative' value is typically set to four. The region of the candidate is partitioned into square cells measuring 2 by 2, and each cell remains unquantified. Identify four locations of a specimen within every individual cell. The method of bilinear interpolation is utilized to determine the floating-point coordinate of the points sampled through the calculation of their values in four different positions. A fixed dimension can provide the output for ROI. They select the highest value among the four central pixels in each partition and utilize it as the 'representative' value, which alters the ROI to a 5 × 5 dimension.

The implementation of ROI pooling yields a significant enhancement in the precision of the detected candidate regions. Furthermore, it can augment the network's efficacy in the identification of minute flaws in objects. The convolutional layers possess the property of being translationally equivariant. Having a strong inductive bias is beneficial because it enhances the model's capability to generalize to new datasets that are not used during training, especially when the dataset used for training is limited. The process of applying a convolution to the input (x) at a specific position (i) can be expressed in equation (10).

$$y_i = \sum\nolimits_{j \in \mathcal{L}R(i)} w \odot xi \tag{10}$$

The value of y_i resulting from the convolution can be represented as $\mathcal{LR}(i)$ which denotes the local receptive field. The original purpose of developing transformers was for a dataset that is processed in a step-by-step manner. According to research, transformer-based models like the ViT have a greater model capacity compared to CNN models. The DL model based on transformers uses a self-attention layer with a broad sensory scope. A significant dissimilarity among self-attention and convolution layers lies in the magnitude of the receptive fields. The self-attention layer possesses the comprehensive scope of perception, which imparts increased contextual knowledge. Moreover, the self-attention layer employs the mechanism that adjusts the weightings based on the input. Therefore, transformer-based models possess substantial model capacity when it comes to extensive datasets. It is important to consider that the computational complexity increases as the receptive field size grows. As per reference [29] the self-attention mechanism can be described in the equation (11).

$$y_i = \sum_{j \in \mathfrak{gs}} \frac{DAW}{\sum_{k \in \mathfrak{gs}} DAW} \tag{11}$$

The dynamic attention weights, represented by DAW $= \sum_{k \in g} e_i^{x^T} x_k$, for each k element within the broader spatial area denoted as gs. The design of the ACoat-Net takes inspiration from the successful Coat-Net algorithm, which fuses convolutional and transformer layers. The core concept behind ACoat-Net involves enhancing BDD's efficacy by effectively utilizing both convolutional and self-attention layers. ACoat-Net has the goal of utilizing the convolution blocks to enhance standardization capabilities and incorporating the self-attention layers to bolster model capability. One straightforward way to merge self-attention and convolution layers is to add an adaptive attention matrix with a global static convolution kernel.

The initial component of the research concept is the convolution block, which is designed to enable extensive characteristic extraction. The high-level features with significance are extracted from the input dataset using convolution layers. In this study, three distinct convolutional layers comprising a conventional 3×3 kernel size layer, the asymmetric convolution architecture comprising 1×3, and 3×1 kernel sizes, and the depth-wise separable convolutions were used. The layer known as depth-wise separable convolutions involve the 3×3 sized kernel for depth-wise convolution, as well as convolution performed by a 1×1 sized point-wise kernel.

The RFFN module bears a resemblance to the MLP head found in the ViT algorithm. As illustrated in Figure 6, this block comprises six distinct layers, namely:

(1) normalization of the layer,
(2) single pixel convolution,
(3) implementation of the Gaussian Error Linear Units activation functions,
(4) another single-pixel convolutions,
(5) a repeat of the GELU activation function, and
(6) addition of the input.

The optimizer is used iteratively to derive the model parameters during model training. The model parameters undergo continuous adjustment through back-propagation to reduce errors that arise once comparing the model outcome with the authentic value. To achieve this objective, the model is first trained on a set of sample data and then its performance is assessed by computing the loss functions on the data set for validation. The testing data set was utilized for assessing the efficacy of the module. The model's efficacy was measured against other advanced transformer-based models such as SwinT, CCT, ViT, and conventional Coat-Net. The SwinT can generate layered feature maps by combining patches of image in its deep layers. The ViT utilizes the modern architecture based on transformers and considers how the spatial dimension is transformed. To ensure the retention of local data, the CCT incorporates a convolutional tokenizer that generates intricate tokens. A model named Coat-Net can explain itself using deep learning and is expected to produce impressive outcomes. Although these models have been deemed as ''black box'' approaches due to a lack of understanding regarding their internal mechanisms. Grad-Cam ranks among the frequently employed techniques for visually justifying deep learning models. This is employed as a means of depicting the visual representation of the models' forecasts. Firstly, the gradient of the score g_c is computed concerning the feature maps (f_k) of a specific layer according to the formula presented in [30]. This calculation is performed before the SoftMax operation.

$$g_c\left(f^k\right) = \frac{\partial y^c}{\partial f^k} \tag{12}$$

''K represents the channel identifier.'' Afterwards, the significance of the weight $(a_k) f^k$ for class c in each channel is calculated by globally averaging the obtained gradients.

$$a_k^c = \frac{1}{\mathfrak{w}_f \times h_f} \sum_{i=1}^{w_f} \sum_{j=1}^{h_f} \frac{\partial y^c}{\partial f_{i,j}^k} \tag{13}$$

The values $\mathfrak{w}_f$ and $\hbar_f$ indicate the dimensions of the feature maps, with the former representing width and the latter representing height. The last heat map generated by Grad-Cam (GC), denoted as $H_{GC}^c = ReLU\left(\sum_k a_k^c . f_k\right)$, is produced by taking a combination of the feature maps, which are multiplied by corresponding weights, trailed by applying a ReLU function.

ROI pooling involves selecting the specific section of the feature map that is of interest, also known as the candidate block. The process of generating candidate blocks in the Faster R-CNN algorithm involves utilizing the RPN technique, which enables mapping these blocks onto the feature architecture to obtain ROI. This pooling involves extracting tiny feature maps from ROIs [21]. The sequence of steps it undergoes during processing are:

- The position of the region on the feature map is correlated with its ROI.
- To standardize the ROI size to N × N, regardless of its initial dimensions, it is divided evenly into N × N

sections. One way to make each ROI the same size is to conduct a max pooling operation on each section by selecting the highest pixel value as a representative. This results in a square ROI of N × N dimensions.

Our proposed approach, with its advanced capabilities in underwater object detection, contributes significantly to the protection and monitoring of these critical communication infrastructures. By accurately detecting and monitoring underwater cables, we enhance the resilience of global telecommunications, ensuring uninterrupted connectivity and communication. Furthermore, oil and gas industry underscore the versatility of our proposed approach. Identifying and monitoring subsea structures, pipelines, and equipment are vital aspects of ensuring the integrity of infrastructure in this industry. The ability of HLASwin-T-ACoat-Net to assess underwater objects not only supports maintenance and inspection efforts but also plays a crucial role in preventing potential environmental hazards.

Pseudocode for HLASwin-T-ACoat-Net

Input:
input_frame: The original input frame to be processed.
rois: Regions of interest for the object detection.
epochs: The number of training epochs for the HLASwin-T-ACoat-Net.
Output:
binary_segmented_frame: The final binary segmented frame representing the output of the object detection using HLASwin-T-ACoat-Net.

1. *HLASwin-T-ACoat-Net (input_frame, rois, epochs):*
   for k in range(epochs):
   a. *frame = capture_frame(input_frame)*
   b. *enhanced_image = apply_clahe(frame)*
   c. *swin_transformed_features=local_acuity_swin _transformer(enhanced_image)*
   d. *coat_net_features=adapted_coat_net(swin_transformed_features)*
   e. *aggregated_features=path_aggregation_network (swin_transformed_features, coat_net_features)*
   f. *aligned_features = roi_alignment (aggregated_features, rois)*
   g. *predictions = object_detection_head(aligned_features)*
   h. *update_weights_and_biases(predictions)*
   return binary_segmented_frame
   # Image Processing
2. *def apply_clahe(image):*
   enhanced_image = clahe(image)
   return enhanced_image
   # Local Acuity Swin Transformer
3. *def local_acuity_swin_transformer(features):*
   transformed_features = swin_transformer(features)
   return transformed_features
   # Adapted Coat-Net
4. *def adapted_coat_net(features):*
   processed_features = coat_net(features)
   return processed_features
   # Path Aggregation Network
5. *def path_aggregation_network(deep_features, shallow_features):*
   aggregated_features = aggregate features(deep_features, shallow_features)
   return aggregated_features
6. *def roi_alignment(features, rois):*
   aligned_features = align_rois(features, rois)
   return aligned_features
   # Object Detection Head
7. *def object_detection_head(features):*
   predictions = detect_objects(features)
   return predictions
   # Combined Algorithm for Object Detection using HLASwin-T-ACoat-Net

IV. EXPERIMENTAL ANALYSIS

A. EXPERIMENTAL SETUP

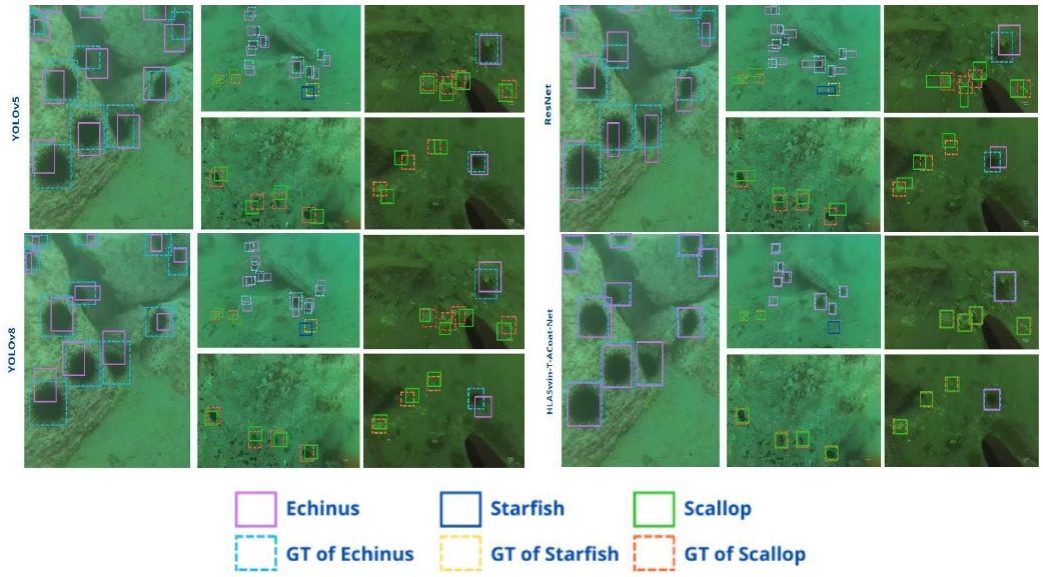
This section presents a detailed analysis of the experimental results of the proposed UOD model. The proposed technique is implemented on an NVIDIA Inferno-CUDA GPU with 256 GB of RAM. The implementation is carried out using the Python programming language and the PyTorch framework, operating within the windows 10 operating system. Various metrics such as mAP, recall, and precision are used to evaluate the research model performance. To compare performance, the URPC 2018 [32] and OUC datasets [5] were applied for this study. The URPC2018 dataset is a broadly used benchmark data set in the domain of UOD.

It consists of a large collection of high-resolution underwater images captured in various underwater environments, including different water conditions, depths, and lighting conditions. The dataset covers a diverse range of underwater objects, such as fish, corals, rocks, and man-made structures. Each image in the dataset is annotated with bounding box labels to indicate the presence and location of objects of interest. The dataset provides a valuable resource for training, testing, and evaluating UOD algorithms. It enables researchers to develop and compare different methods for detecting and recognizing objects in challenging underwater scenarios.

The OUC dataset is a comprehensive dataset specifically designed for UOD research. It was created by the Ocean University of China (OUC) and contains many high-quality underwater images. The dataset covers various underwater scenes and environments, including different water types, depths, and lighting conditions. It includes a wide range of underwater objects, such as marine life, submerged structures, and underwater vegetation. The images in the dataset are annotated with accurate bounding box labels to facilitate object detection and evaluation. The OUC dataset provides a valuable resource for training and testing UOD algorithms. It allows researchers to assess the performance of different algorithms under different underwater conditions and compare their results.

The URPC2018 dataset comprises 3,701 images, in which 2,901 images were applied for training and 800 images for testing. The OUC-VISION dataset contains 4,400 images, in which 3,520 images were applied for training and 880 for testing. The datasets are divided into training, validation, and testing sets in an 80:10:10 ratio. The proposed approach was utilized to classify the datasets into four distinct groups of submerged entities. To assess the effectiveness of the system, evaluation criteria such as precision, recall, and mean Average Precision (mAP) were employed.

Precision is a parameter c that measures the accuracy of the positive predictions made by the model. In the context of UOD, precision is the ratio of correctly predicted positive instances (correct detections) to the total number of instances predicted as positive by the model (both correct and incorrect). A higher precision indicates a lower rate of false



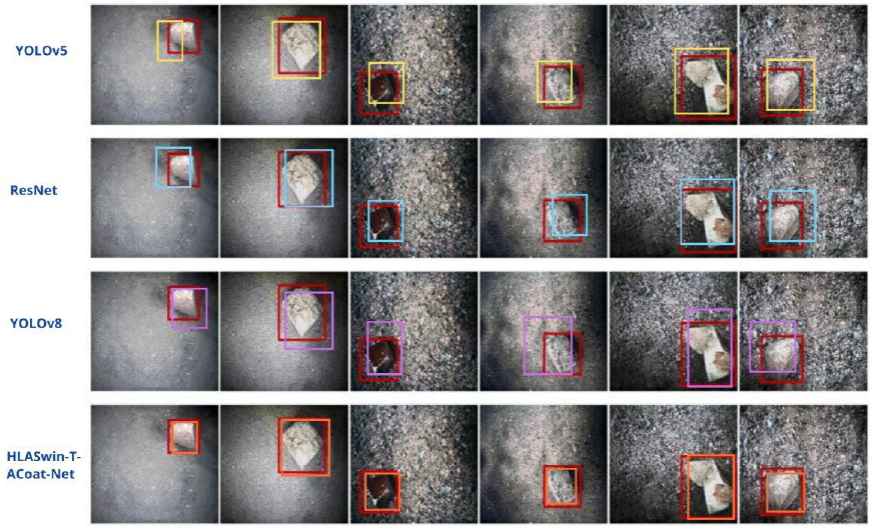


FIGURE 7. Visualization object detection results of different methods on URPC 2018 and OUC.

positives, which is crucial in applications where the cost of misidentifying objects is high. Recall evaluates the capability of the model to capture all relevant instances of a class. It is the ratio of correctly predicted positive instances to the total number of actual positive instances in the dataset. A higher recall indicates that the model can successfully identify a larger proportion of the actual positive instances.

mAP is a comprehensive metric that considers precision at different levels of recall across multiple thresholds. It involves calculating the average precision for each class and then taking the mean across all classes. mAP provides a more nuanced evaluation of the model's performance, especially in scenarios with imbalanced class distributions or varying object sizes. The HLAST-ACNet method is compared to YOLOv5 [14], ResNet [18], and YOLOv8 [19] using comparative tests to highlight its superiority. Figure 7 depicts the object detection results of different methods on Underwater Robot Picking Control 2018 (URPC 2018).

Research suggests that when it comes to identifying and locating small or camouflaged objects in complicated underwater settings, current methods often fall short due to their inadequate feature representation. The detection capability of small objects is improved with HLAST-ACNet as compared to other available methods. Despite the availability of numerous techniques, there remain certain items that are similar in appearance to their surroundings, making it challenging to make accurate assessments using current methods. When it comes to identifying 'starfish', which bear a striking

resemblance to the surroundings, there tends to be a comparatively high incidence of failure to detect them accurately. The bidirectional feature pyramid generates a robust feature representation, resulting in significant enhancements for the detection of 'starfish' in HLAST-ACNet. The approach exhibits superior results compared to the previously mentioned methods. The approach's capability to effectively differentiate and classify objects with significant meaning is believed to enhance the process of detecting underwater objects.

The HLAST-AC-Net technique is superior to other approaches in terms of detecting a larger number of items with more precise pinpointing, particularly for smaller objects that may blend into the surroundings. The expected boundary areas are shown by the solid rectangles, while the ground truth boundary areas are depicted by the dashed rectangles. An image or illustration that displays information clearly and understandably. The diagram illustrates the distribution of the leading incorrect positives across every category of the UOD assessment collections. HLAST-ACNet can precisely identify objects during cluttered information, even in instances where those objects are situated against intricate backgrounds. The successful use of the CLAHE method, a potent technique showcased in Figure 8, has made this achievable. Table 2 represents the outcomes of the OUC experimentations. MPCNN elevates the ability to detect objects underwater with its performance of 91.25 mAP on OUC. The increase in input size directly impacts the profit outcome. MPCNN outperforms other existing methods including YOLOv5, ResNet, and YOLOv8 in terms of performance.

Table 3 and Figure 9 present a clear representation of the findings derived from the URPC 2018 experiments. HLAST-ACNet significantly improves the detection of objects underwater by achieving a 91.25 mAP score on UOD. The increase in the input size leads to an increment in profits. HLAST-ACNet's superiority in detecting small objects is attributed to its capability to detect objects of various sizes across multiple layers. The HLAST-ACNet stands out in its ability to detect small entities, outperforming all other cutting-edge methods by a considerable degree on UOD. Figure 10 showcases the Precision/Recall graphs for different methods used in URPC 2018.

TABLE 2. Comparisons with the state-of-the-arts on the OUC data set.

Models	Precision (%)	Success rate (before denoising)	Success rate (after denoising)	mean of average precision (mAP) %
YOLOv5 [14]	0.363	0.48	0.54	70.8
ResNet [18]	0.372667	0.46	0.565	81.36
YOLOv8[19]	0.382	0.47	0.586	85.63
HLAST-ACNet	0.479333	0.51	0.599	91.25

The HLAST-ACNet model, depicted as a blue curve, demonstrates superior performance in the sea urchin and scallop categories, while the Scallops approach surpasses its predecessors. This implies that the HLAST-ACNet model is unlikely to improve performance for either a single or multiple categories of objects. One can observe a steady and significant enhancement in the accuracy of all trackers when assessed against the updated dataset. This is substantiated by the quantitative results displayed in Table 3 and figure 10.

According to the graph in Figure 11, which demonstrates a comparison of results, utilizing the recommended method of HLAST-ACNet leads to a more effective enhancement of data compared to using pre-enhanced data with the same technique. The graphs depicting the benchmark results demonstrate that the tracking devices show exceptional efficiency with the use of the suggested CLAHE enhancement technique to refine the underwater data. The HLAST-ACNet outperforms all other methods with its exceptional performance.

The results from the evaluation of object detection models for the identification of Holothurian, Echinus, Scallops, and Starfish reveal intriguing insights into the performance of the considered models. Precision/Recall percentages serve as critical indicators of a model's ability to balance accuracy and completeness in its predictions. YOLOv5, the initial model in the comparison, demonstrated a Precision/Recall of 0.52, indicating a moderate level of accuracy. However, both ResNet and YOLOv8 surpassed YOLOv5, achieving Precision/Recall percentages of 0.6925 and 0.8175, respectively. The standout performer in this regard was HLAST-ACNet, boasting an impressive Precision/Recall of 0.8475, suggesting superior accuracy in identifying the target marine organisms. Success rates before and after denoising are crucial factors in assessing the robustness of models in real-world scenarios. YOLOv8 exhibited significant improvements in success rates, reaching 0.7875 after denoising, surpassing both YOLOv5 and ResNet. HLAST-ACNet outperformed all other models in both success rates before and after denoising, showcasing its effectiveness in accurately detecting marine organisms even in noisy conditions. This emphasizes the importance of denoising techniques in enhancing the practical applicability of object detection models.

The mean of average precision (mAP) provides a comprehensive measure of overall detection accuracy. HLAST-ACNet once again emerged as the top performer with a mAP of 92.36, underscoring its superiority in accurately identifying Holothurian, Echinus, Scallops, and Starfish. While YOLOv8 demonstrated a competitive mAP of 87.63, indicating strong overall performance, YOLOv5 and ResNet trailed behind, suggesting potential for improvement in their detection capabilities. In summary, the results highlight the diverse performance levels among the models, with HLAST-ACNet standing out as the most accurate and reliable in detecting marine organisms.

To make the combination of various techniques clear, the mAP scores shall be depicted in a visual format. Figure 12 displays the number of loop executions. After analyzing and contrasting the outcomes, HLAST-ACNet boasts better recognition precision than alternative techniques, although the difference is not pronounced. The proposed approach

TABLE 3. Comparisons with the state-of-the-arts on URPC 2018 data set.

Models	Precision/Recall (%)				Success rate (before denoising)	Success rate (after denoising)	mean of average precision (mAP) %
	Holothurian	Echinus	Scallops	Starfish			
YOLOv5[14]	0.52	0.5325	0.59	0.7275	0.46	0.56	75.8
ResNet [18]	0.6925	0.6125	0.6775	0.7825	0.47	0.58	86.36
YOLOv8[19]	0.8175	0.68	0.7875	0.8175	0.49	0.59	87.63
HLAST-ACNet	0.8475	0.7375	0.8225	0.86	0.53	0.61	92.36

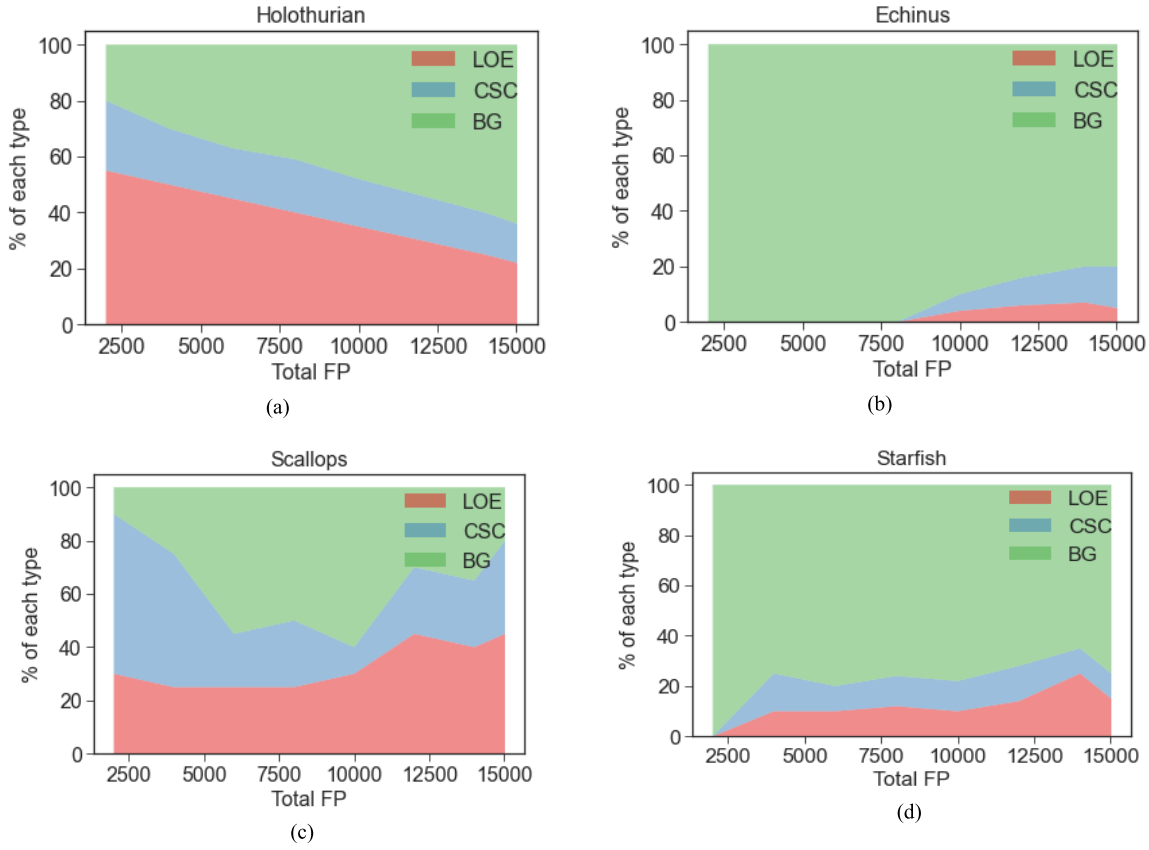


FIGURE 8. The distribution of top-ranked false positive (FP) types for each category and all categories on UOD (a) sea cucumbers, (b) sea urchins, (c) scallops and (d) all categories. The false positives include localization errors (LOE), confusion with similar categories (CSC), with background (BG).

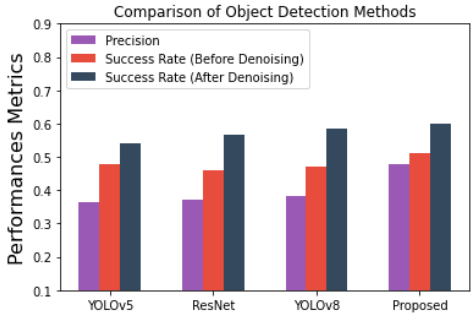


FIGURE 9. Comparison of Models on the OUC data set.

is significantly more effective and has the potential to achieve superior detection outcomes when compared with existing methods. HDCNN-UODT achieved convergence after approximately 2500 iterations, which is a relatively early stage compared to other approaches. By incorporating the updated strategies, the level of stability and mAP can be improved. The proposed methods can produce a more accurate result compared to traditional methods.

TABLE 4. Quantitative performance evaluation results of proposed and existing methods.

Method	YOLOv5 [14]	ResNet [18]	YOLOv8 [19]	HLAST-ACNet
GFLOPs	42.13	38.56	22.65	5.06
Time cost (ms)	94	83	42	20
FPS(s)	98.58	82	65.52	2.28

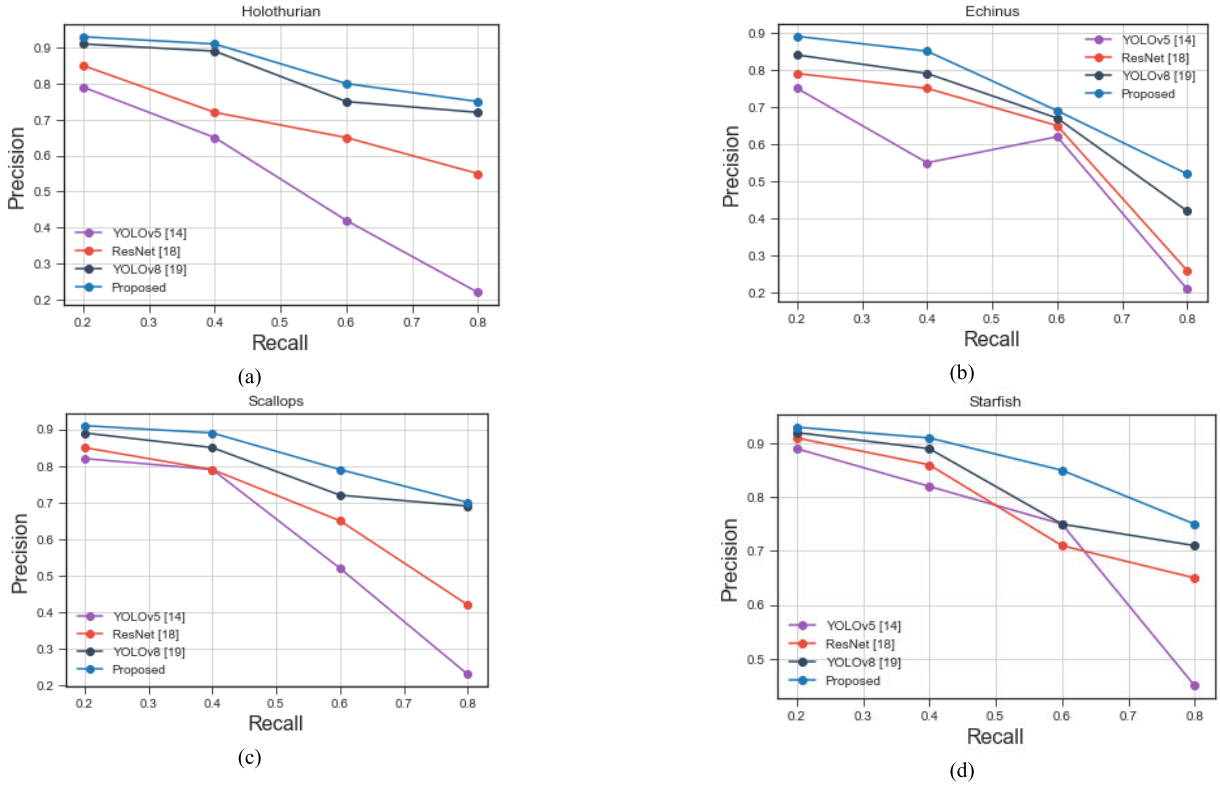


FIGURE 10. Precision/Recall curves of different methods on UOD.

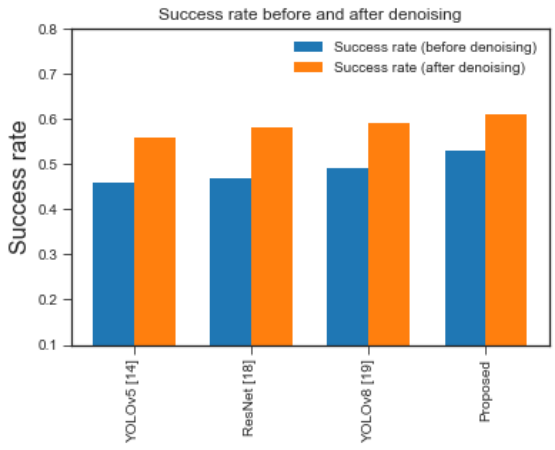


FIGURE 11. Success rate plots on the URPC 2018 dataset-after and before the denoising process in HLAST-ACNet.

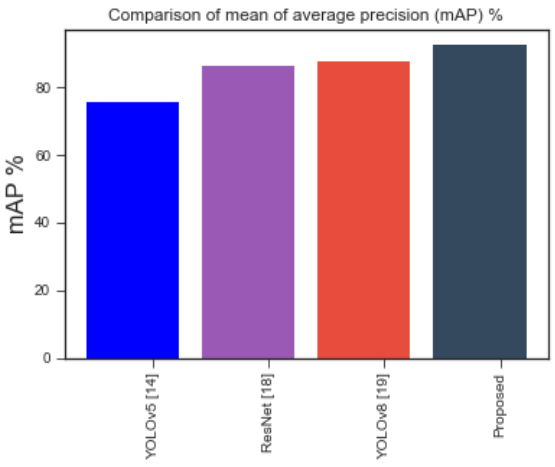


FIGURE 12. mAP results and comparison with other methods (%).

All the techniques demonstrate accuracy that is adequate for detecting objects and can therefore be effectively used in real-time scenarios. Table 4 illustrates the importance of identifying velocity, as expressed in measurable variables including time cost, GFLOPs, and training epochs. The proposed framework underwent an ablation study, resulting in the attainment of said accomplishment. According to the data displayed in Table 4, the advanced deep model suggested demonstrates better excellence in detecting objects, although it requires slightly higher GFLOPs and fewer training epochs than the proposed method that lacks colour conversion and the suggested method with traditional colour conversion.

The main reason for this is that the proposed method, which employs LASwin-T and CLAHE for efficient feature extraction, can eliminate interference and enhance the image, all while retaining its edges. The approach suggested has proven to have a higher computational complexity compared to four other existing methods, as evidenced by its GFLOPs,

per-image processing time with a speed of 20ms, and the FPS measures for average processed frames per second reaching 2.28s. Furthermore, the suggested network has a limited computational power of only 5.05 GFLOPs. The proposed advanced model is beneficial in detecting underwater objects by enabling the removal of noise and facilitating object detection.

TABLE 5. Computational and space complexity performance evaluation results of proposed and existing methods.

METHOD	COMPUTATIONAL COMPLEXITY	SPACE COMPLEXITY
ResNet	O(n2d2HWC)	O(dmax Hmax Wmax)
YOLOv5	O (backbone complexity + neck complexity + head complexity + n log n)	O(dmax Hmax Wmax)
YOLOv8	O(d'k'2 H'W' + n log n)	O(dmax Hmax Wmax)
Proposed Method	O(H2 W2 d_in d_out)+ O(Adapted Coat-Net Complexity) + O(Path Aggregation Complexity)	O (H2 W2 d)

Time and Space complexity: The complexity of the Neck component can be expressed in terms of the variables n, d, H, W, C, and k. The use of PANet suggests that there are multiple operations, such as convolutions and element-wise operations, contributing to the overall time complexity. A more detailed breakdown may involve (n) number of input channel, number of output channels (d), input height (H), input width (W), number of channels in the current layer (C), and the kernel size (k). In table 5 shows the comparative analysis time and space complexity.

ResNet is noted for its efficacy in addressing the vanishing gradient problem, enabling the training of deep networks and providing expressive feature representations. However, its quadratic dependency on input size may lead to increased computational demands. YOLOv5 stands out for real-time object detection with a balanced one-stage architecture, yet the logarithmic term in its computational complexity raises concerns about scalability with larger datasets. YOLOv8 introduces optimizations in bounding box processing efficiency, making it suitable for real-time applications, but the logarithmic term may pose challenges with increased dataset sizes.

The proposed method combines an efficient feature representation with the introduction of adapted Coat-Net and path aggregation complexities. While these enhancements aim to improve overall performance, the method's multiple components may increase computational demands, potentially impacting real-time processing.

V. CONCLUSION

To enhance the precision of underwater targets identification in the complicated underwater environments, a solution called HLAST-ACNet is proposed. In the conducted experiments, the performance of the ACoat-Net model with Swin Transformers of various sizes as the backbone network was evaluated. Ablation experiments are conducted to assess the effectiveness of various improvement techniques. The proposed algorithm is compared with alternative algorithms, successfully demonstrating its advanced nature.

Through experimentation, the HLAST-ACNet model is enhanced to achieve improved detection results in intricate underwater settings. Notably, mAP rates of 91.25% and 92.36% are achieved on the URPC 2018 and OUC datasets, respectively. While the model may have a slower detection speed compared to single-stage detection algorithms, it is important to consider its larger size. The URPC dataset comprises numerous indistinct images that were not specifically designed to enhance detection performance. The research encompasses several objectives, including model compression, detection acceleration, dataset expansion through the collection of additional underwater target data, and enhancing robustness through data augmentation techniques. In future, the proposed work can be extended for detecting very tiny small objects in the underwater and system integration, robustness evaluation in real-world scenarios, and addressing operational challenges such as underwater object tracking.

Future works for the proposed HLAST-ACNet include exploring multi-modal data fusion to further enhance object detection accuracy in complex underwater environments and integrating real-time adaptive learning mechanisms for improved model adaptability. Additionally, investigating the model's performance across diverse datasets and extending its applicability to three-dimensional object detection are avenues for future research. However, limitations include the need for robustness testing in dynamic underwater scenarios, potential challenges in scalability for large-scale deployments, and the necessity for continuous refinement to accommodate evolving underwater conditions and object diversity. Addressing these aspects will strengthen the model's utility and advance its potential in real-world applications.

CONFLICT OF INTEREST

The authors have no financial or other conflicts of interest.

REFERENCES

[1] M. Li, A. Mathai, S. L. H. Lau, J. W. Yam, X. Xu, and X. Wang, "Underwater object detection and reconstruction based on active single-pixel imaging and super-resolution convolutional neural network," *Sensors*, vol. 21, no. 1, p. 313, Jan. 2021.

[2] J. S. Jaffe, "Underwater optical imaging: The past, the present, and the prospects," *IEEE J. Ocean. Eng.*, vol. 40, no. 3, pp. 683–700, Jul. 2015.

[3] N. Karpel and Y. Y. Schechner, "Portable polarimetric underwater imaging system with a linear response," *Proc. SPIE*, vol. 5432, pp. 106–115, Jul. 2004.

[4] H. Lu, Y. Li, Y. Zhang, M. Chen, S. Serikawa, and H. Kim, "Underwater optical image processing: A comprehensive review," *Mobile Netw. Appl.*, vol. 22, no. 6, pp. 1204–1211, Dec. 2017.

[5] L. Chen, L. Tong, F. Zhou, Z. Jiang, Z. Li, J. Lv, J. Dong, and H. Zhou, "A benchmark dataset for both underwater image enhancement and underwater object detection," 2020, *arXiv:2006.15789*.

[6] J. Zhang, L. Zhu, L. Xu, and Q. Xie, "Research on the correlation between image enhancement and underwater object detection," in *Proc. Chin. Autom. Congr. (CAC)*, 2020, pp. 5928–5933.

[7] M. Yang and A. Sowmya, "An underwater color image quality evaluation metric," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 6062–6071, Dec. 2015.
[8] J. Zhou, T. Yang, and W. Zhang, "Underwater vision enhancement technologies: A comprehensive review, challenges, and recent trends," *Int. J. Speech Technol.*, vol. 53, no. 3, pp. 3594–3621, Feb. 2023.
[9] A. Vats and T. Patnaik, "A systematic review on underwater image enhancement and object detection methods," in *Proceedings of Emerging Trends and Technologies on Intelligent Systems*. Singapore: Springer, 2022, pp. 359–372.
[10] J. D. Kelleher, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2019.
[11] D. J. Hemanth and V. V. Estrela, *Deep Learning for Image Processing Applications*, vol. 31. IEEE, TX, USA: IOS Press, 2017.
[12] A. Abu and R. Diamant, "A statistically-based method for the detection of underwater objects in sonar imagery," *IEEE Sensors J.*, vol. 19, no. 16, pp. 6858–6871, Aug. 2019.
[13] W. Ji, J. Peng, B. Xu, and T. Zhang, "Real-time detection of underwater river crab based on multi-scale pyramid fusion image enhancement and MobileCenterNet model," *Comput. Electron. Agricult.*, vol. 204, Jan. 2023, Art. no. 107522.
[14] X. Hua, X. Cui, X. Xu, S. Qiu, Y. Liang, X. Bao, and Z. Li, "Underwater object detection algorithm based on feature enhancement and progressive dynamic aggregation strategy," *Pattern Recognit.*, vol. 139, Jul. 2023, Art. no. 109511.
[15] S. Panda and P. K. Nanda, "Kernel density estimation and correntropy based background modeling and camera model parameter estimation for underwater video object detection," *Soft Comput.*, vol. 25, no. 15, pp. 10477–10496, Aug. 2021.
[16] C.-H. Yeh, C.-H. Lin, L.-W. Kang, C.-H. Huang, M.-H. Lin, C.-Y. Chang, and C.-C. Wang, "Lightweight deep neural network for joint learning of underwater object detection and color conversion," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 11, pp. 6129–6143, Nov. 2022.
[17] X. Sun, J. Shi, L. Liu, J. Dong, C. Plant, X. Wang, and H. Zhou, "Transferring deep knowledge for object recognition in low-quality underwater videos," *Neurocomputing*, vol. 275, pp. 897–908, Jan. 2018.
[18] T.-S. Pan, H.-C. Huang, J.-C. Lee, and C.-H. Chen, "Multi-scale ResNet for real-time underwater object detection," *Signal, Image Video Process.*, vol. 15, no. 5, pp. 941–949, Jul. 2021.
[19] E. Li, Q. Wang, J. Zhang, W. Zhang, H. Mo, and Y. Wu, "Fish detection under occlusion using modified you only look once v8 integrating real-time detection transformer features," *Appl. Sci.*, vol. 13, no. 23, p. 12645, Nov. 2023.
[20] Y. Shi, "An underwater target wake detection in multi-source images based on improved YOLOv5," *IEEE Access*, vol. 11, pp. 31990–31996, 2023.
[21] H. Wang and N. Xiao, "Underwater object detection method based on improved faster RCNN," *Appl. Sci.*, vol. 13, no. 4, p. 2746, Feb. 2023.
[22] I. S. Isa, M. S. A. Rosli, U. K. Yusof, M. I. F. Maruzuki, and S. N. Sulaiman, "Optimizing the hyperparameter tuning of YOLOv5 for underwater detection," *IEEE Access*, vol. 10, pp. 52818–52831, 2022.
[23] J. Jia, M. Fu, X. Liu, and B. Zheng, "Underwater object detection based on improved EfficientDet," *Remote Sens.*, vol. 14, no. 18, p. 4487, Sep. 2022.
[24] W. Yi and B. Wang, "Research on underwater small target detection algorithm based on improved YOLOv7," *IEEE Access*, vol. 11, pp. 66818–66827, 2023.
[25] J. Liu, S. Liu, S. Xu, and C. Zhou, "Two-stage underwater object detection network using swin transformer," *IEEE Access*, vol. 10, pp. 117235–117247, 2022.
[26] J. Muwei, Q. Qi, J. Dong, Y. Yin, W. Zhang, and K. M. Lam, "The OUC-vision large-scale underwater image database," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 1297–1302.
[27] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.
[28] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*.
[29] M. M. Naseer, K. Ranasinghe, S. H. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Intriguing properties of vision transformers," in *Proc. NIPS*, Dec. 2021, pp. 23296–23308.
[30] D. Kvak, "Visualizing CoAtNet predictions for aiding melanoma detection," 2022, *arXiv:2205.10515*.
[31] P. E. Pope, S. Kolouri, M. Rostami, C. E. Martin, and H. Hoffmann, "Explainability methods for graph convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10764–10773.
[32] J. Yuan, Y. Hu, Y. Sun, and B. Yin, "A multi-scale feature representation and interaction network for underwater object detection," *IET Comput. Vis.*, vol. 17, no. 3, pp. 265–281, Apr. 2023.

S. MANIMURUGAN (Senior Member, IEEE) received the Bachelor of Engineering degree in computer science and engineering from Anna University, India, in 2005, the Master of Engineering degree in computer science and engineering from Karunya University, India, in 2007, and the Ph.D. degree from Anna University, in 2012.

He is currently a Professor with the Faculty of Computers and Information Technology, University of Tabuk, Tabuk, Saudi Arabia. He has made significant contributions to these areas and has published numerous papers in various conferences and journals. His research interests include artificial intelligence, security, image processing, and the Internet of Things. He is a Life Member of the ISTE.

C. NARMATHA (Member, IEEE) received the bachelor's and master's degrees in electronics and communication engineering from Anna University, India, in 2005 and 2012, respectively, and the Ph.D. degree in electronics and communication engineering from Karpagam University, India, in December 2018.

She is currently an Associate Professor with the Faculty of Computers and Information Technology, University of Tabuk, Tabuk, Saudi Arabia. With over ten years of teaching and research experience, she specializes in the fields of image processing, AI, and security. She has made significant contributions to her field and has published numerous research articles in reputable international and national journals and international conferences indexed by ISI, Thomson Reuters/Clarivate Analytics, and Scopus. She has also served as a guest editor and a reviewer for various esteemed national and international journals, including Elsevier, Springer, and IEEE. Her expertise and research achievements make her a valuable member of the Faculty of Computers and Information Technology, University of Tabuk.

MAJED M. ABOROKBAH (Member, IEEE) received the B.Sc. degree from Taif University, Saudi Arabia, the M.Sc. degree from Bradford University, U.K., and the Ph.D. degree from De Montfort University, U.K.

He is currently the Dean of the Faculty of Computers and Information Technology, University of Tabuk, Saudi Arabia. His research interests include artificial intelligence, software engineering, context-aware systems, cybersecurity, and steganography. He has contributed significantly to these fields and has published numerous papers in international journals and conferences. Additionally, he has played an active role in organizing various workshops and conferences related to his research areas. He has also made notable contributions to the establishment of the robotics center with the University of Tabuk.

NAVEEN CHILAMKURTI (Senior Member, IEEE) received the Ph.D. degree from La Trobe University, Melbourne, VIC, Australia.

He is currently the Acting Head of the Department, Computer Science and Computer Engineering, La Trobe University. He has published about 165 journals and conference papers. His current research interests include intelligent transport systems (ITS), wireless multimedia, and wireless sensor networks. He also serves on the editorial boards for several international journals. He is also an Associate Editor of *IJCS* (Wiley), *SCN*, *JETWI* (Inderscience), and *IJIPT*. He is also the Inaugural Editor-in-Chief of *International Journal of Wireless Networks and Broadband Technologies*, in July 2011.

SUBRAMANIAM GANESAN (Life Senior Member, IEEE) received the master's and Ph.D. degree from Indian Institute of Science (IISc), Bengaluru, India.

He was the Chair of the CSE Department, from 1991 to 1998. He is currently a Professor in electrical and computer engineering with Oakland University, Rochester, MI, USA. He has over 35 years of teaching and research experience in digital systems and computer engineering. He has been with the Electrical and Computer Engineering Department, since 2008. He was with the National Aeronautical Laboratory (NAL), India, Ruhr University, Germany, Concordia University, Canada, and Western Michigan University, before joining Oakland University.

RAJENDRAN THAVASIMUTHU (Member, IEEE) received the bachelor's, master's, and Ph.D. degrees in electronics and communication engineering from the Karpagam Academy of Higher Education (Deemed to be University), Coimbatore, in 2013, 2015, and 2021, respectively.

Previously, he was the Managing Director of Makeit Technologies, Coimbatore, and the Vice Principal of the MM Polytechnic College, Trichy. He is holding the office of the Organizational Head of Makeit Technologies, Coimbatore, and an Adjunct Professor with the Saveetha School of Engineering, Chennai. He has published more than 30 indexed/peer-reviewed articles and one international (Australia) patent to his credit. His research interests include biomedical image/signal processing, soft computing, VLSI technology, the IoT, AI, and networks.

P. KARTHIKEYAN received the Bachelor of Engineering (B.E.) degree in computer science and engineering from Anna University, Chennai, Tamil Nadu, India, in 2005, the Master of Engineering (M.E.) degree in computer science and engineering from Anna University Coimbatore, India, in 2009, and the Ph.D. degree from Anna University, Chennai, in 2018.

He was a Ph.D. Research Fellow with the Department of Computer Science and Information Engineering, National Chung Cheng University, Taiwan. He has contributed to the national projects on using technology to promote human rights and sustainable development, supported by the National Science and Technology Council, Taiwan. He possesses a high level of proficiency in project development and research, specializing in the domains of cloud computing and the practical application of deep learning. He is well-versed in programming languages, including Java, Python, R, and C. His research endeavors have led to the publication of more than 30 articles in esteemed international journals, many of which have garnered commendable impact factors. Furthermore, he has shared his research insights through presentations at over 20 international conferences, solidifying his reputation within the academic community.

M AMMAD UDDIN holding the Doctorate and Post-Doctorate degrees from ENSTA Bretagne, France. He is a Distinguished Researcher in the field of wireless sensor networks. He currently serves as an Associate Researcher with ENSTA Bretagne, while previously holding a position as an Assistant Professor at the University of Tabuk, Saudi Arabia. Throughout his career, he has been instrumental in advancing the knowledge of computer science, teaching a multitude of graduate and bachelor's courses across prestigious universities in France, Saudi Arabia, and Pakistan. His academic expertise extends beyond the classroom, as he has spearheaded numerous research and development projects in vital areas such as wireless sensor networks, underwater sensor networks, and smart agriculture. Presently, He is deeply engaged in the Smart Case for Accidental Spill Monitoring Intervention (SAMI) initiative, a collaborative effort funded by Total Energies aimed at safeguarding seawater against the detrimental effects of oil spill incidents. His commitment to addressing pressing environmental concerns underscores his dedication to leveraging technology for the greater good. His scholarly contributions are evidenced by his authorship of multiple research papers, prominently featured in IEEE and other esteemed journals. Furthermore, his inventive contributions have earned him recognition as an inventor in a patent registered in the USA. His research interests primarily revolve around routing, clustering, and localization of sensor nodes within wireless sensor networks, reflecting his unwavering commitment to advancing the frontiers of sensor technology and its applications.