TENCON 2023 - 2023 IEEE Region 10 Conference (TENCON)
31 Oct - 3 Nov 2023. Chiang Mai, Thailand

# Attention-fused Shallow Network for Underwater Object Detection

Amul Chourasia[1] Veronica Naosekpam[1] and Nilkanta Sahu[1]

*Abstract*—**Although generic object-detection methods have achieved a great deal, it needs more exploration for underwater images. Underwater object detection (UOD) is associated with challenges such as degraded image quality, low visibility, low contrast, colour shift, and limited computational capacity availability on the deployment environment. Moreover, most previous studies on deep learning-based underwater object detection have generally concentrated on increasing detection accuracy by utilizing huge networks. This work proposes a two-stage method where in the first stage, underwater images are enhanced based on PCA-fusion method. This step involves mutiple image enhancement steps such as color correction process, followed by the White Patch Retinex Algorithm for white balancing. On the colour-corrected image, we apply global histogram equalization, unsharp masking, and median smoothing separately to improve the contrast, sharpen the image, and reduce the white patch noise. Then, we create a single enhanced image by combining the results of the three methods using the principal component analysis (PCA) based fusion method. Finally, in the second stage, the enhanced image is passed as an input to the attention-fused lightweight single-stage object detection model for localization and classification. Experimental results show that the proposed method outperforms the state-of-the-art algorithms based on various image enhancement and object detection evaluation metrics on the URPC2019 dataset.**

*Index Terms*—**underwater object detection, lightweight network, attention mechanism, YOLOv5s, image enhancement**

## I. Introduction

Underwater object detection (UOD) is a prominent research area in the field of computer vision, which aims to develop reliable algorithms and systems capable of localizing and identifying objects in underwater environments. This domain presents particular challenges due to the inherent complexities of underwater imaging, including poor visibility, color distortion, light absorption, and scattering phenomena. The ability to accurately detect and classify underwater objects has numerous applications in various fields, such as marine biology, underwater exploration, environmental monitoring, and underwater robotics. Due to this, generic object detection methods [1]–[3] that have achieved remarkable success in above-water scenarios have limited usage in underwater environments. Therefore, specialized methods and models are required to overcome these obstacles and achieve accurate and robust object detection performance.

In recent years, significant progress has been made in underwater object detection through the adoption of deep learning techniques [4], [5], producing encouraging results. These deep learning models can extract discriminative features and capture detailed spatial information, improving object detection performance in underwater scenes. However, the majority of previous studies have mainly concentrated on enhancing detection accuracy through the use of expansive and intricate network designs [6]. There is still a need for further exploration and research specifically tailored to underwater image processing and object detection.

With the development of computer vision and image processing technology, research on the application of image processing methods to improve the underwater image quality to satisfy the requirements of the human visual system and machine recognition has gradually started along with the techniques of UOD such as the works in [7], [8]. As the ocean environment is complex, many unfavorable factors, such as the scattering and absorption of light by water and the underwater suspended particles, have serious interference with image quality. Due to this, using a fusion of various image enhancement techniques is a promising approach.

This paper presents a novel approach for underwater object detection that blends image-enhancing methods based on fusion technology with a shallow object detection network. Before feeding the underwater images into the object detection model, our methodology involves several image processing steps to improve the underwater images. The image-enhancement methods we have outlined here have been optimized for a particular kind of underwater image dataset that involves underwater living organisms, considering factors such as the prevalent lighting conditions, water turbidity, color aberrations, and the specific types of objects to be detected. This optimization is an essential step as different underwater environments can significantly vary in their properties and therefore the effect of each preprocessing step can vary accordingly. We begin by applying a color correction process to address color aberrations in underwater images. Subsequently, we employ the White Patch Retinex Algorithm for white balancing. We further enhance the contrast, sharpness, and noise reduction on the color-corrected image through the application of global histogram equalization, unsharp masking, and median smoothing techniques, respectively. To create a single enhanced image that preserves the essential information from the individual enhancement methods, we employ the principal component analysis (PCA) based fusion method. Finally, the enhanced image is fed into the attention-fused YOLOv5s model, a lightweight object detection network. The attention mechanism in the network enables the model to concentrate on the most informative regions of the enhanced image, leading to accurate and efficient detection of underwater objects.

The main contributions of the paper are threefold:

- We propose a comprehensive underwater image enhancement pipeline that addresses various challenges related to underwater images. A PCA-based fusion technique is used as a post-processing step to preserve essential information while effectively creating a single enhanced image.
- We propose a novel lightweight attention-fused UOD framework using YOLOv5s. A global attention module is inserted between the neck and head of the YOLOv5s network to extract useful feature information effectively.
- We conduct extensive experiments on the URPC2019

[1]CSE Department, Indian Institute of Information Technology Guwahati
veronica.naosekpam@iiitg.ac.in

979-8-3503-0219-6/23/$31.00 ©2023 IEEE

dataset to demonstrate the effectiveness of the proposed methodology using object detection and image enhancement evaluation metrics.

The remainder of this paper is organized as follows: Section II provides an overview of related works in underwater object detection and image enhancement techniques. Section III describes the proposed methodology in detail. We describe the experimental setup and analysis in Section IV. Finally, in Section V, we draw a conclusion and potential future directions.

## II. Related Work

### A. Generic object detection

One of the most fundamental and challenging issues in computer vision is object detection [9], which has recently attracted much attention. Viola-Jones detector [10], histograms of oriented gradients, and deformable part-based model [11] are a few examples of traditional object recognition frameworks based on handcrafted feature engineering.

With the rapid development in deep learning techniques, state-of-the-art generic object detection is categorized into: one-stage methods and two-stage methods. The one-stage detectors mostly followed the works include You Only Look Once (YOLO) [2] and Single Shot Detector (SSD) [1]. The two-stage object detection algorithms such as R-CNN [12], Fast-RCNN [13] and Faster-RCNN [14]. The convolutional neural network (CNN) is the primary component of deep model-based object detection frameworks that automatically performs feature learning (or representation learning), integrated with classification or regression tasks.

### B. Underwater object detection

Vision-based underwater object detection has garnered research attention considering its wide applications in marine engineering, ocean environment monitoring, and underwater robotics. One of the first studies on UOD focussed on detecting and identifying fish species [4] using a quick R-CNN-based technique. Multiscale features and complementary context data were used in [5] to create a single-shot feature aggregation network for UOD. From the data augmentation perspective, Lin et al. [15] proposed an augmentation technique to conduct proposal-level fusion among multiple images. An underwater detection framework with feature augmentation and anchor refinement was proposed by Fan et al. [6]. DG-YOLO [16] performed domain generalization to address the UOD task's domain shift. By interpolating paired images on the feature level, Chen et al. [17] established the training paradigm known as DMCL to sample new domains on the domain manifold.

Some methods combine image enhancement with underwater object detection as an end-to-end process methods. To solve the weakly illuminated problem, Fenglei et al. [7] combined the max-RGB method, shades of gray method, and a CNN to obtain an illumination map. Yeh et al. [18] proposed to jointly train the color conversion and object recognition for underwater object detection address the underwater environment's low contrast. Wang et al. [19] also proposed an end-to-end CNN-based underwater image enhancement framework for color correction and haze removal. Although the aforementioned methods considerably increased detection accuracy, majority of the approaches utilized deep learning techniques which further increases the resource requirement to train the model. In our approach, we use simple statistical image features extraction techniques which require low computational resources.

## III. Proposed Methodology

Based on the above analysis, we argue that utilizing only the object detection method without enhancing the input image cannot handle the challenging underwater environment well. Therefore, we choose to take advantage of both image enhancement methods and single-stage object detection methods to solve the underwater target detection task.

### A. Fusion-based Image Enhancement

We illustrate the proposed framework for image enhancement in Fig. 1. The first step involves color correction, followed by parallel processing of contrast, sharpening, and smoothing enhancement on the color-corrected image. Finally, the results are fused using a PCA-based fusion technique to obtain the final enhanced image.

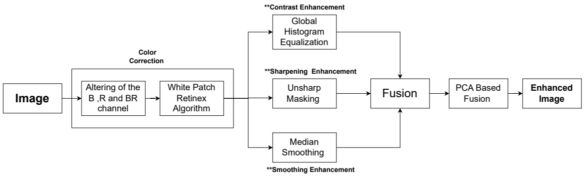

Fig. 1. Proposed image enhancement method

*1) Color Correction:* Color correction in underwater images involves altering the color balance to compensate for the loss of colors due to light absorption in water. It is observed that in underwater images, the red and yellow predominate after the blue and green wavelengths are initially absorbed. For color correction, we must first compensate the red (R) channel's loss of quality. If images have a greenish look, we may also need to compensate for the blue (B) channel's loss of quality. Since the green channel is the least damaged, adding a portion to the red and blue channels when necessary serve as compensation. We use a color correction technique [20] in this investigation that entailed modifying the B, R, and BR channels. Then, we utilize a white balance algorithm to help correct any color casts that may have been present in the image and make sure the colors look genuine and natural.

Given an input image $I$, where $I_r$, $I_g$, and $I_b$ represent the red, green, and blue color channels, respectively, we want to compensate for the loss of quality in the red and blue channels. The formula for the compensated red channel $I_{rc}$ and blue channel $I_{bc}$ at every pixel location $(x)$ are :

$$I_{rc}(x) = I_r(x) + (\overline{I}_g - \overline{I}_r) \cdot (1 - I_r(x)) \cdot I_g(x)$$

$$I_{bc}(x) = I_b(x) + (\overline{I}_g - \overline{I}_b) \cdot (1 - I_b(x)) \cdot I_g(x)$$

The formulas compensate for the loss of quality in the red and blue channels by incorporating the information from the relatively less damaged green channel, thereby improving color balance in the image.

The White Patch Retinex Algorithm is applied after color balancing an image to rectify the color shifts caused by light absorption and scattering. It is based on the Retinex theory [21], which states that the colors perceived by the human visual system are determined by the ratio of an object's reflectance to that of its surroundings. In underwater images, the colors of objects often undergo alterations due to light absorption and scattering. To rectify these color distortions, the White Patch Retinex algorithm estimates the reflectance of a "White Patch" in the image and utilizes it to modify the colors of other objects. The algorithm follows these steps:

1. Convert the input image from the RGB color space to the logarithmic domain:

$$\begin{aligned}\log_R &= \log(R+1)\\ \log_G &= \log(G+1)\\ \log_B &= \log(B+1)\end{aligned}$$

2. Calculate the illumination map $I$ using the following formula:

$$I = \max(\log_R, \log_G, \log_B)$$

3. Find the maximum value $I_{\text{max}}$ of the illumination map within a small region around the brightest pixel. This region is determined using a localized search. Calculate the reflectance map $R$ using the following formula:

$$R = \log_R - \log(I_{\text{max}})$$

4. Convert the reflectance map back to the RGB color space:

$$R = \exp(R) - 1$$

5. Apply a gamma correction to the output image to improve its visual quality: $R_{\text{out}} = R^{\gamma}$, where $\gamma$ is a user-defined parameter controlling contrast.

6. Normalize the output image to ensure that pixel values fall within the desired range:

$$R_{\text{out}} = R_{\text{out}} \times \frac{255}{\max(R_{\text{out}})}$$

By estimating the reflectance of a reference "White Patch"and adjusting the colors of other objects based on this reference, the White Patch Retinex Algorithm can effectively rectify color distortions in underwater images.

*2) Global Histogram Equalization:* It is comprised of the following steps :

1. Convert the input RGB image to the HSV color space. The HSV color space represents colors based on their hue, saturation, and value/intensity.

2. Apply global histogram equalization to the value component (V) of the HSV image, while keeping the hue (H) and saturation (S) components unchanged.

3. The equalised HSV image is created by combining the equalised value component with the original hue and saturation components.

4. The contrast-enhanced image is produced by converting the equalised HSV image back to RGB colour space.

*3) Image Sharpening:* We perform image sharpening using unsharp masking technique. The process for image sharpening can be represented as follows:

1. We convert the RGB input image to grayscale and create a blurred version of the grayscale image using a smoothing filter that is, Guassian blur:

$$I_{\text{gray}}(x,y) = 0.299 \cdot R(x,y) + 0.587 \cdot G(x,y) + 0.114 \cdot B(x,y)$$

$$I_{\text{blurred}}(x,y) = \text{gBlur}(I_{\text{gray}}(x,y))$$

2. We then calculate the high-pass filtered image by subtracting the blurred grayscale image from the original grayscale image:

$$I_{\text{highpass}}(x,y) = I_{\text{gray}}(x,y) - I_{\text{blurred}}(x,y)$$

3. After this, we perform sharpening by adding the scaled high-pass filtered image multiplied by a scaling factor called "amount" to the original grayscale image:

$$I_{\text{sharp}}(x,y) = I_{\text{gray}}(x,y) + \text{amount} \times I_{\text{highpass}}(x,y)$$

4. Lastly, we convert the sharpened grayscale image back to the original color space by replicating the sharpened grayscale image across all three color channels (R, G, B).

*4) Median Smoothing:* It is a method for smoothing out noise while maintaining an image's edges. It functions by substituting the median value of each pixel's surrounding pixels for that pixel's value.

*5) PCA-based Fusion:* After applying the respective enhancement techniques, we have the following enhanced images: Global histogram equalized Image: $I_{\text{ghe}}$, Sharpened Image: $I_{\text{sharp}}$, Smoothed Image: $I_{\text{smoothed}}$. We convert these enhanced images into matrices by rearranging the pixel values for each color channel: Equalized Matrix: $M_{\text{ghe}}(c,n)$ where $c$ represents the color channel and $n$ denotes the total number of pixels in the image. Sharpened Matrix: $M_{\text{sharpen}}(c,n)$ Smoothed Matrix: $M_{\text{smoothed}}(c,n)$. Next, we concatenate these matrices into a single matrix: $M_{\text{concat}} = \begin{bmatrix} M_{\text{ghe}} \\ M_{\text{sharpen}} \\ M_{\text{smoothed}} \end{bmatrix}$ Next, we perform PCA on the concatenated matrix to yield the principal components (PCs) and their corresponding eigenvalues. Let $P$ be the matrix of principal components and $\Lambda$ be a diagonal matrix containing the eigenvalues. Then, we have, $M_{\text{concat}} = P\Lambda$. Assuming the original image has dimensions $M \times N$ and there are three enhanced images (smoothened, sharpened, and histogram-equalized), the concatenated matrix $C$ will have dimensions $M \times (3N)$. Each row of $C$ represents a pixel, and each column corresponds to a different enhanced image or processing step.

The next step involves selecting a subset of the principal components that capture the most significant information. Let $l$ be the number of principal components to select. We then can create a new matrix, $P_l$, containing only the $l$ leading columns of $P$, and a corresponding diagonal matrix, $\Lambda_l$, containing the $l$ leading eigenvalues. $M_{\text{concat}} \approx P_k \Lambda_k$

Finally, the fused image is reconstructed by multiplying the selected principal components with their corresponding eigenvectors and summing the results as: $M_{\text{fused}} = P_l \Lambda_l$. The PCA-based image fusion technique assists in creating a fused image that integrate the improved features from multiple image enhancement methods.

### *B. Attention-fused Shallow Underwater Object Detection Network*

We modified the YOLOv5 detection framework for underwater object detection by introducing a global attention mechanism (GAM) to the overall network. Fig. 2 illustrates the overall framework of the detection network. As our focus is to make our scheme feasible for embedded systems, we choose the shallow version, that is, YOLOv5s which have a small depth and width while ensuring high accuracy. To amplify cross-dimensional interactions among the features at different scales, we add GAM between the neck and the head.

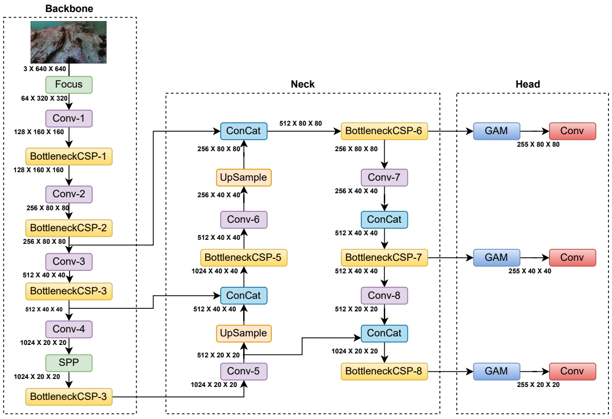

Fig. 2. Architecture of attention-fused YOLOv5s

The YOLOv5s network has three parts: backbone, neck, and head. The backbone is a modified version of the CSP-Darknet53 network. It consists of multiple convolutional layers that extract hierarchical features from the input image. These features capture low-level and high-level information about the objects in the image at three different scales. The Focus module, the Conv module, the C3 module, and the spatial pyramid pooling module are all a part of it. The neck is a component that connects the backbone network to the detection heads. In YOLOv5s, a PANet (Path Aggregation Network) and Feature Pyramid Network (FPN) are used to fuse features from different scales to enhance the representation of objects of various sizes. The three detection heads in the YOLOv5s architecture work at various spatial scales of the feature map. Each detection head is responsible for predicting bounding boxes, object classes, and confidence scores for objects at its particular scale.

The attention mechanism, as outlined by Guo et al. [22], prioritizes regions of the image where distinctive features are present over those with less discriminative information. By doing so, this global attention mechanism (GAM) augments the network's capacity to learn and extract meaningful patterns, thereby enhancing its overall performance. The structure of GAM is shown in Fig. 3.

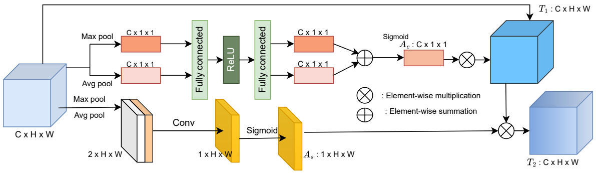

Fig. 3. Structure of the global attention mechanism

Given an input feature map $T \in \mathbb{R}^{C\times H\times W}$, it is handled in the channel as well as in the spatial dimension to obtain channel attention map $A_c \in \mathbb{R}^{C\times 1\times 1}$ and spatial attention map $A_s \in \mathbb{R}^{1\times H\times W}$. The final output is obtain by the following steps:

$$T_1 = A_c(T) \otimes T$$

$$T_2 = A_s(T_1) \otimes T_1$$

where $\otimes$ refers to the element-wise multiplication, $T_1$ and $T_2$ are intermediate and final output.

We compress the spatial dimensions of the feature map using average pooling and maximum pooling to generate two features, send the two features to a fully connected layer and a ReLU layer, and then combine the two features using element summation to calculate the channel attention effectively. Finally, a sigmoid function standardizes the channel attention map. The spatial link between the features creates the spatial attention map. Two features are obtained by compressing the channel dimensions of the feature map using average pooling and maximum pooling to calculate the spatial attention effectively. We use a 3 × 3 kernel for convolution to link them. Finally, we standardize all elements using the sigmoid function to create the spatial attention map.

The loss functions include localization loss $L_{loc}$, which measures the discrepancy between the predicted bounding box coordinates and the ground truth coordinates calculated using mean squared error; confidence loss $L_{conf}$, which penalizes the confidence scores of predicted bounding boxes based on their overlap with ground truth boxes and is given by binary cross-entropy; and class loss $L_{cls}$ which measures the discrepancy between the predicted class probabilities and the ground truth class labels typically calculated using categorical cross-entropy. The overall loss function used for training YOLOv5s is a linear combination of the localization, confidence, and class losses, with appropriate weights assigned to each loss:

$$L_{\text{total}} = \alpha \cdot L_{\text{loc}} + \beta \cdot L_{\text{conf}} + \gamma \cdot L_{\text{cls}}$$

Here, $\alpha$, $\beta$, and $\gamma$ are hyperparameters that control the relative importance of each loss term.

## IV. Experiment and Analysis

We choose URPC2019 dataset for experimental analysis. The dataset contains 4707 images, randomly divided into a training set and validation set according to the ratio of 8:2. The images are of size 640 × 640. It comprises four object categories: Echinus, Holothurian, Scallop and Starfish.

We adopt the mean average precision (mAP), precision, recall, and F1-score for object detection and entropy, average gradient, and UCIQE to asses the image enhancement method. Our network is initialized using a pre-trained CSP-Darknet53 model. We conducted experiments on NVIDIA GeForce RTX 3090 graphics processor (40 G RAM). We optimize the model using the SGD (stochastic gradient descent) method. The training epochs are fixed to 400, with a batch size of 16, an initial learning rate of 0.01, a weight decay of 0.0005, and the SGD momentum set to 0.9. In addition, we also used the default data augmentation technique used in YOLOv5. The hyperparameters settings of data enhancement as : Scale =0.25, Mosaic = 0.75, Mix up = 0.50, Fliplr = 0.50 , Flipud =0.10, Translate = 0.10, hsv_h = 0.01, hsv_s = 0.70, and hsv_v = 0.40.

Table I compares different schemes or methods based on three quantitative measurements: entropy, average gradient, and Universal Image Quality Index for Enhancement (UCIQE). The proposed method yields the highest entropy and average gradient based on these results. At the same time, Kamul et al. [23] scheme achieves the highest UCIQE value, suggesting superior image quality. Here, original refers to the original images without any enhancement method.

TABLE I
QUANTITATIVE MEASUREMENT OF IMAGE ENHANCEMENT METHOD ON URPC2019.

| Scheme | Entropy | Average Gradient | UCIQE |
|---|---|---|---|
| Our method | **7.47** | **38.40** | 0.43 |
| Original | 6.11 | 12.31 | 0.43 |
| Kamil et al. [23] | 7.07 | 22.60 | **0.44** |

Fig. 4 illustrates some of the results from the proposed enhancement method as a step-by-step visuals. We notice that the PCA-based fusion, the retinex based algorithm and the histogram equalization significantly removes the haze effect and the color channel removal clearly performs the balancing of the RGB channels. The sharpening step also removes the blurriness of the image.

We show the overall results obtained via the proposed UOD method on the original and the enhanced URPC2019 image datasets in Table II. The proposed method performs reasonably well in terms of precision, especially when the images are enhanced. However, the method without enhancement has a relatively low recall, indicating that it misses a significant number of positive instances. The enhancement improves recall, resulting in a higher F1-measure, indicating better overall performance in precision and recall.

TABLE II
OVERALL SCORES FOR PROPOSED UOD. OD REFERS TO THE RESULT WITHOUT THE ENHANCEMENT PART AND ED REFERS TO THE IMAGE WITH ENHANCEMENT.

| Scheme | Precision | Recall | F1-measure |
|---|---|---|---|
| Our method (OD) | 0.86 | 0.52 | 0.65 |
| Our method (ED) | **0.89** | **0.72** | **0.80** |

Table III presents a comparison of the performance of various state-of-the-art schemes on the URPC2019 dataset. The evaluation metrics used include average precision (AP) and class-wise precision for different marine species categories: Echinus, Holothurian, Scallop, and Starfish. The first row of the table corresponds to proposed method which incorporates the proposed image enhancement techniques and attention-stage object detection method. It achieves an AP of 0.76, demonstrating its superior performance compared to the other schemes. In particular, it achieves the highest precision for the Echinus and Starfish categories, with values of 0.88 and 0.73, respectively. The second row corresponds to the YOLOv5s scheme, which achieves an AP of 0.57. It demonstrates relatively lower performance compared to our method across all categories except for the Scallop category, where it achieves a precision of 0.75. The subsequent rows correspond to other state-of-the-art schemes such as [8], [16], [24]–[26]. Notably, Paul et al. [25] achieves the highest precision for the Scallop category (0.75), while Zhang et al. [8] achieves the highest precision for the Holothurian category (0.62).

TABLE III
COMPARISON WITH THE STATE-OF-THE-ART ON THE URPC2019 DATASET.

| Scheme | AP | Ehinus | Holothurian | Scallop | Starfish |
|---|---|---|---|---|---|
| Our method | **0.76** | **0.88** | 0.60 | 0.69 | **0.73** |
| YOLOv5s | 0.57 | 0.66 | 0.50 | 0.75 | 0.28 |
| Paul et al. [24] | 0.65 | 0.66 | 0.50 | **0.75** | 0.71 |
| Weo et al. [25] | 0.59 | 0.66 | 0.50 | 0.50 | 0.71 |
| Zhang et al. [8] | 0.65 | 0.72 | **0.62** | 0.61 | 0.68 |
| Liu et al. [16] | 0.55 | 0.63 | 0.35 | 0.36 | 0.27 |
| Dai et al. [26] | 0.48 | 0.48 | 0.36 | 0.44 | 0.51 |

Fig. 5 shows some qualitative results on URPC2019 UOD before (left) and after (right) the enhancement method. We observe that after the enhancement process, more number of objects are detected and overall as well as class-wise performance is also improved.

## V. CONCLUSION

In this work, we proposed an attention-fused lightweight UOD network based on YOLOv5s. We also proposed an image enhancement method as a pre-requisite for underwater image to reduce the difficulty in detecting targets in underwater environment based on PCA-fusion. Therefore, the increase in the final detection accuracy plus its competitiveness with the state-of-art methods on the URPC2019 although the network is shallow is the result of the combination of enhancement step and the global attention mechanism. The future research prospects includes : end-to-end enhancement and detection framework that works for UOD including underwater text detection and recognition [27].

## REFERENCES

[1] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 21–37.

[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[3] V. Naosekpam, N. Kumar, and N. Sahu, "Multi-lingual indian text detector for mobile devices," in *International Conference on Computer Vision and Image Processing*. Springer, 2020, pp. 243–254.

[4] X. Li, M. Shang, H. Qin, and L. Chen, "Fast accurate fish detection and recognition of underwater images with fast r-cnn," in *OCEANS 2015-MTS/IEEE Washington*. IEEE, 2015, pp. 1–5.

[5] L. Zhang, X. Yang, Z. Liu, L. Qi, H. Zhou, and C. Chiu, "Single shot feature aggregation network for underwater object detection," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 1906–1911.

[6] B. Fan, W. Chen, Y. Cong, and J. Tian, "Dual refinement underwater object detection network," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*. Springer, 2020, pp. 275–291.

[7] F. Han, J. Yao, H. Zhu, and C. Wang, "Underwater image processing and object detection based on deep cnn method," *Journal of Sensors*, vol. 2020, 2020.

[8] J. Zhang, L. Zhu, L. Xu, and Q. Xie, "Research on the correlation between image enhancement and underwater object detection," in *2020 Chinese Automation Congress (CAC)*. IEEE, 2020, pp. 5928–5933.

[9] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proceedings of the IEEE*, 2023.

[10] P. Viola, "Hobust real-time face detec-tiorj," *Int. J. Comput*, vol. 57, pp. 137–154, 2004.

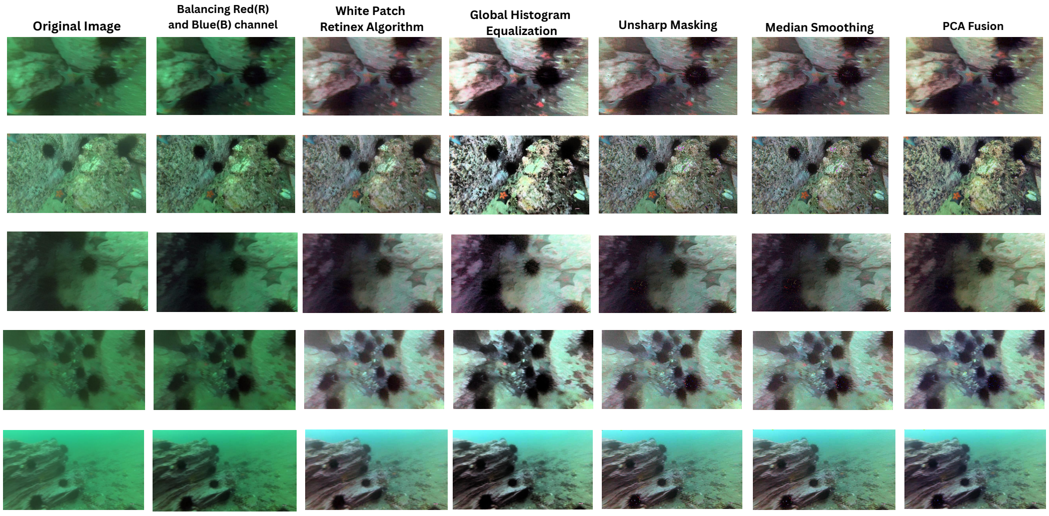
Original Image
Balancing Red(R) and Blue(B) channel
White Patch Retinex Algorithm
Global Histogram Equalization
Unsharp Masking
Median Smoothing
PCA Fusion

Fig. 4. Visuals from step-by-step results of the proposed enhancement method.

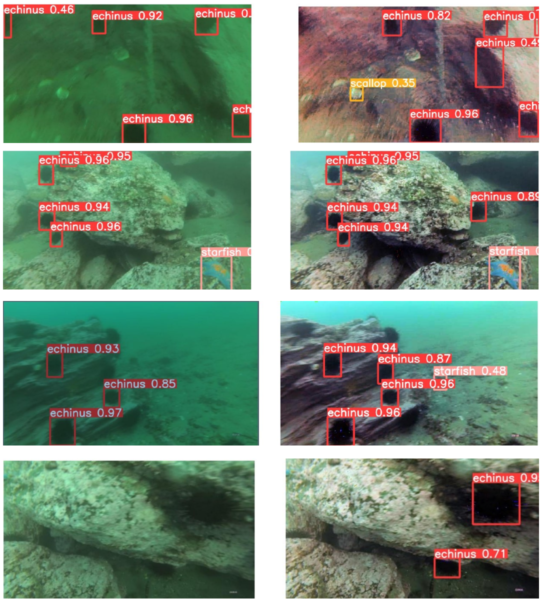

Fig. 5. Visual results of the proposed method on the original image (left) and enhanced image (right).

[11] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2009.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[13] R. Girshick, "Fast r-cnn," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1440–1448.

[14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[15] W.-H. Lin, J.-X. Zhong, S. Liu, T. Li, and G. Li, "Roimix: proposal-fusion among multiple images for underwater object detection," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 2588–2592.

[16] H. Liu, P. Song, and R. Ding, "Towards domain generalization in underwater object detection," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 1971–1975.

[17] L. Chen, F. Zhou, S. Wang, J. Dong, N. Li, H. Ma, X. Wang, and H. Zhou, "Swipenet: Object detection in noisy underwater scenes," *Pattern Recognition*, vol. 132, p. 108926, 2022.

[18] 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 Transactions on Neural Networks and Learning Systems*, vol. 33, no. 11, pp. 6129–6143, 2021.

[19] Y. Wang, J. Zhang, Y. Cao, and Z. Wang, "A deep cnn method for underwater image enhancement," in *2017 IEEE international conference on image processing (ICIP)*. IEEE, 2017, pp. 1382–1386.

[20] C. O. Ancuti, C. Ancuti, C. De Vleeschouwer, and P. Bekaert, "Color balance and fusion for underwater image enhancement," *IEEE Transactions on image processing*, vol. 27, no. 1, pp. 379–393, 2017.

[21] E. H. Land, "The retinex theory of color vision," *Scientific american*, vol. 237, no. 6, pp. 108–129, 1977.

[22] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R. R. Martin, M.-M. Cheng, and S.-M. Hu, "Attention mechanisms in computer vision: A survey," *Computational Visual Media*, vol. 8, no. 3, pp. 331–368, 2022.

[23] K. Z. M. Azmi, A. S. A. Ghani, Z. M. Yusof, and Z. Ibrahim, "Natural-based underwater image color enhancement through fusion of swarm-intelligence algorithm," *Applied Soft Computing*, vol. 85, p. 105810, 2020.

[24] P. Drews, E. Nascimento, F. Moraes, S. Botelho, and M. Campos, "Transmission estimation in underwater single images," in *Proceedings of the IEEE international conference on computer vision workshops*, 2013, pp. 825–830.

[25] W. Song, Y. Wang, D. Huang, and D. Tjondronegoro, "A rapid scene depth estimation model based on underwater light attenuation prior for underwater image restoration," in *Advances in Multimedia Information Processing–PCM 2018: 19th Pacific-Rim Conference on Multimedia, Hefei, China, September 21-22, 2018, Proceedings, Part I 19*. Springer, 2018, pp. 678–688.

[26] L. Dai, H. Liu, P. Song, H. Tang, R. Ding, and S. Li, "Edge-guided representation learning for underwater object detection," *arXiv preprint arXiv:2306.00440*, 2023.

[27] V. Naosekpam and N. Sahu, "Text detection, recognition, and script identification in natural scene images: A review," *International Journal of Multimedia Information Retrieval*, vol. 11, no. 3, pp. 291–314, 2022.