

Shape-Guided Detection: A joint network combining object detection and underwater image enhancement together

Chao Yang, Longyu Jiang, Zhicheng Li, Jie Wu

Abstract—Due to the poor quality of underwater images, they are usually not accurate when used for object detection. Most of the currently available solutions involve pre-processing, such as underwater image enhancement, to improve object detection accuracy. However, the pre-processing approach is to improve the subjective perception of the human eye, which does not necessarily improve the object detection performance and also consumes a lot of computational resources. Therefore, in this paper, we innovatively combine these two tasks and propose Shape-Guided Detection Network (SG Detection). We fuse underwater image enhancement and object detection into one network, where object detection can share clean features through joint learning, thus facilitating underwater object detection, and the object detection results will also guide the direction of underwater image enhancement. Meanwhile, in the SG Detection network, we embed the object shape prior features as a learnable module, and design the Shape Prior Enhancement module to achieve the full fusion and interaction between the prior features and local features. In addition, we innovatively adopt a combination of explicit and implicit constraint approach, which ultimately achieves high object detection accuracy. We have conducted extensive experiments on public available datasets, and we find that our approach achieves up to 10% mAP improvement in object detection performance and significantly increases the processing speed.

Index Terms—Underwater image enhancement, object detection, deep learning, priori information enhancement

I. INTRODUCTION

DEEP learning based models have been successfully used in object detection over the past few years. The ability of detecting underwater objects has also significantly increased. Due to their unique imaging technique, which produces low image contrast and ambiguous object edge information, underwater images yet suffer from a generally poor level of object detection accuracy.

There are two main research directions that can be taken to increase the accuracy of underwater image object detection: (1) designing underwater image enhancement algorithms, which can improve the quality of underwater image object detection by enhancing the quality of underwater images; and (2) designing object detection algorithms for underwater object features,

which are primarily used to address issues with blurred edges and dense distribution of underwater object.

The first type of methods, by designing image pre-processing algorithms to improve the underwater quality and enhance the contrast of underwater objects, thus significantly improving the underwater object detection accuracy. The main research methods are [1], [2], [3], [4]. Although these techniques increase the accuracy of underwater object detection to some level, there are still significant issues: 1) The underwater environment is unpredictable, there is a certain domain bias between the data of the training enhancement algorithm and the object detection data, and the underwater image enhancement algorithm sometimes does not significantly improve the image quality of the underwater object detection dataset; 2) These algorithms, which usually need to apply the underwater image enhancement algorithm first and then run the object detection algorithm, consume a lot of time and resources.

The second type of research approaches, which mostly rely on detection algorithms existing in use, focuses on the deliberate design of modules to address the issue of underwater object detection. For example, Chen [5] used high-resolution and semantically rich feature maps to build a network to significantly improve the accuracy of underwater small object detection; Wang [6] proposed a weighted plus ghost-CSPDarknet and a simplified PANet to improve the underwater object detection performance; Liu [7] proposed domain invariant modules and invariant risk minimization penalties to improve the generalization performance of underwater object detection. Fu [8] searches for transferable prior knowledge from detector-friendly images and uses prior knowledge that can guide the detector to eliminate interfering factors. In general, these algorithms improve the performance of underwater object detection to a certain extent, but there are still some problems: 1) most of the methods focus on a class of targets with poor generalization ability; and 2) they depend on image quality, and in the case of poor image quality, they may fail.

We observe that in underwater object detection, the objects are mostly rigid bodies whose shape and structure information remain unchanged in most cases, and that those shape and structure information can help us to detect the objects faster and more accurately. In other words, the introduction of shape priors can help the object detection model to better consider and exploit the shape information of the target object in the recognition and localisation process, thus improving the object detection performance.

Therefore, the approach in this paper combines the advan-

The authors are with School of Computer Science and Engineering, Southeast University, NanJing 210096, China(e-mail: chuckyang@seu.edu.cn, JLY@seu.edu.cn, 230219182@seu.edu.cn, jie_wu@seu.edu.cn).

The work of Chao Yang was supported by the National Natural Science Foundation of China(No: 61871124 and 61876037), The national defense Pre-Research foundation of China, by the fund of Science and Technology on Sonar Laboratory (No: 6142109KF201806), by the Stable Supporting Fund of Acoustic Science and Technology Laboratory (No:JCKYS2019604SSJSSO12).

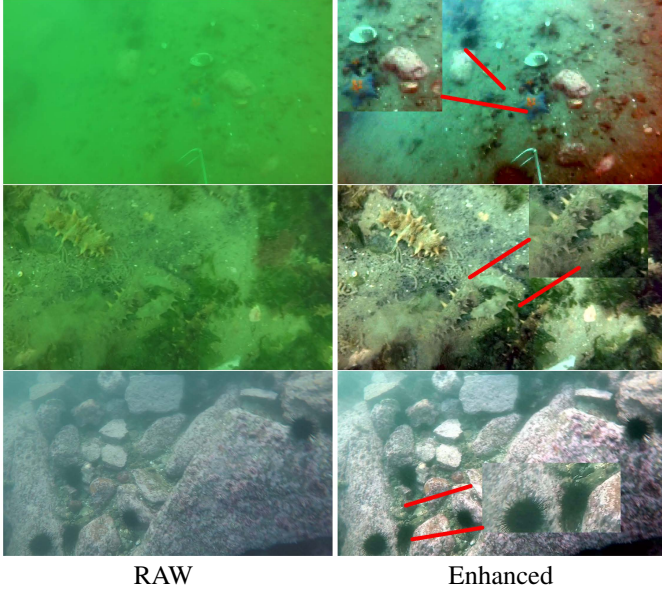


Fig. 1. An ideal example of underwater image enhancement algorithms. The underwater image enhancement algorithm is not only able to effectively correct colours and remove noise, fog and other disturbances, but more importantly, it can also reconstruct the information about the object's shape, colour and texture very well, which are the keys to improve the performance of high level vision tasks.

tages of the above two types of methods by fusing underwater image enhancement and object detection into one network. By sharing the feature extraction layer, object detection can obtain more potential features from image enhancement, a generative model; and the object detection results can also guide the image enhancement direction. Our contribution can be summarized as follows:

- 1) First, we propose a combined network called Shape-Guided detection (SG Detection), which combines the tasks of object detection and underwater image enhancement and significantly speeds up the preprocessing of data and inference time.
- 2) Next, we propose a shape prior enhancement module. Using the interaction between the prior information and the original skip features, we effectively fuse the global and local shape information, which leads to a significant improvement in object detection and underwater image enhancement performance.
- 3) Then, we use a combination of explicit and implicit prior constraints in the model to force the model to focus more on the utilization of shape prior information, which will guide the network for better object detection and image enhancement.
- 4) Finally, we perform extensive comparison experiments and ablation experiments on the public dataset. Compared with the state-of-the-art algorithms, our proposed method achieves good performance improvement in both object detection performance and underwater image enhancement. Meanwhile, our proposed method significantly improves the pre-processing speed of underwater images.

II. RELATED WORK

A. Underwater Image Enhancement

The existing underwater image enhancement algorithms are mainly divided into two categories. One type is pixel processing to improve the visual experience of human eyes, and typical methods include histogram equalization, white balance, multi-scale retinex algorithm, contrast-limited adaptive histogram equalization, and so on. Due to the complexity of the underwater environment, it is difficult for such methods to recover high-quality underwater images relying only on observational information. In addition, these methods are only evaluated according to people's subjective preferences and lack objective evaluation metrics.

The other category mainly focuses on underwater image recovery algorithms. Pairs of underwater images are trained to achieve the recovery of degraded underwater images to natural images. For example, [9] designed WaterGAN network to train degraded underwater images end-to-end, which consists of a depth estimation module and a color correction module. Based on the original image imaging process, [10] builds the basic degradation model and then introduces CNN or lightweight residual learning framework to estimate the degradation parameters. These classes of methods have shown some improvement in underwater image quality and some correction for chromatic aberration, but in practice, many algorithms do not adapt well to different scenes. Moreover, many algorithms do not provide much improvement for higher-order vision tasks (e.g., object detection, segmentation, etc.).

B. Underwater Image Object Detection

Existing methods for underwater object detection are mainly derived from General Object Detection (GOD) and Underwater Image Enhancement (UIE) methods. The former aims to better localize and identify targets, while the latter aims to improve underwater object detection performance by improving the quality of degraded underwater images.

To address the problem of poor localization and recognition accuracy, [11] combined Faster R-CNN [74] with three classification networks to achieve automatic detection and recognition of fish species in underwater videos. [12] proposed a data enhancement method based on candidate region fusion to generate simulated overlapping, occluded, and blurred training samples, thus improving the accuracy and robustness of the model. [13] proposed a two-stage network of underwater small object detection methods using a deformable convolutional pyramid structure to solve the object deformation problem. [14] proposed a sea urchin detection network based on the SSD algorithm. A multi-directional edge detection algorithm was proposed to enhance the spiky edge feature of sea urchins. [5] proposed a neural network structure for underwater small object detection based on DSSD, which was called SWIPENet. These methods mentioned above have improved the efficiency of underwater object detection to some extent, but due to the low contrast of underwater images, color bias, and difficulty in extracting rich details, these problems pose a great challenge to the task of underwater object detection.

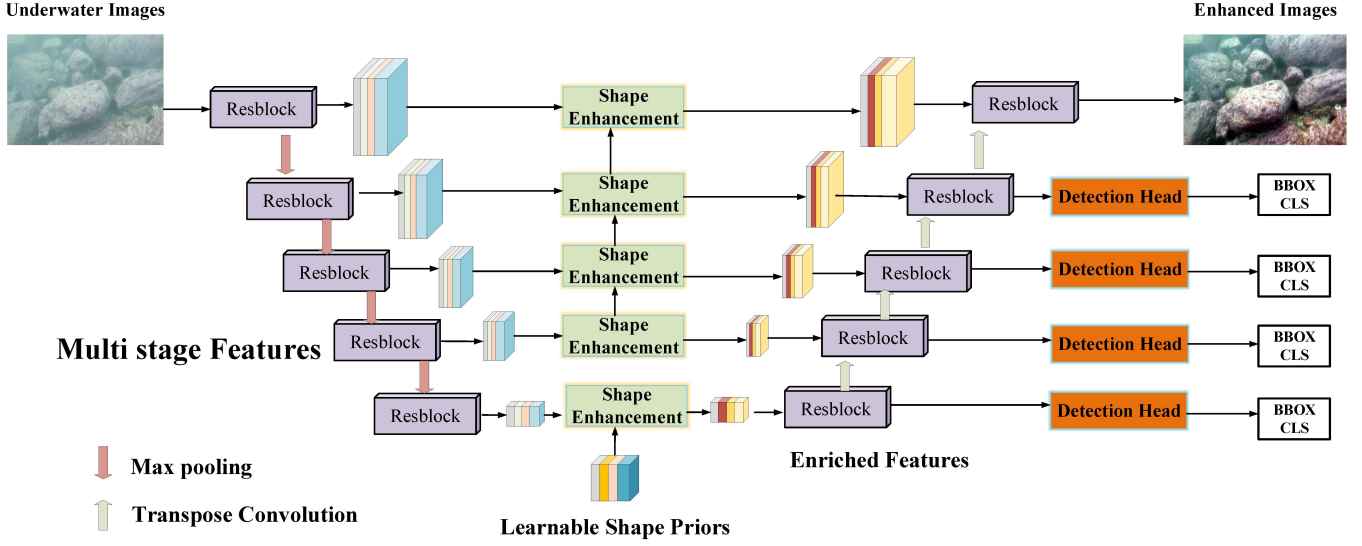


Fig. 2. Architecture of Shape-guided Detection Network. Our network is trained directly from underwater images, after a common feature extraction layer, to obtain the original features at different layers. Then, the shape prior information is embedded in the network in a learnable feature ground manner, which is then enhanced for each layer of features after a shape enhancement module. The enhanced feature maps are connected to two different decoders for image decoding and detection decoding, respectively, to obtain clean underwater images and target detection results.

In addition, researchers have tried to combine underwater image enhancement with object detection. [15] proposed a Cycle-GAN based underwater image color correction method to improve underwater object detection. [16] proposed an improved MSRCP and YOLOv3 based underwater target detection method to solve the underwater blurring and low contrast problems of images. [2] proposed a depth model for joint learning of color conversion and object detection of underwater images. The image color conversion module aims to convert color images to corresponding grayscale images to solve the problem of underwater color absorption, thus improving object detection performance with lower computational complexity. Most of these methods mentioned above are still based on natural image enhancement methods when it comes to image enhancement with the goal of improving visual quality. However, according to the available literature comparison results, there is no strict positive correlation between detection accuracy and visual quality, and the problems of noise and domain bias caused by such methods are still serious. In conclusion, the optimization goal of underwater image enhancement algorithms should consider the optimization goal of object detection rather than simply improving the visual quality to satisfy the human eye.

Therefore, the motivation for this paper comes from the fact that we combine the two tasks into a combined network where clean features produced by the enhancement network can be shared in the detection network for better object detection rather than using underwater image enhancement as a pre-processing method for object detection. At the same time, the results of object detection can be used to guide the direction of image enhancement, allowing our image enhancement to focus on improving the performance of object detection. To further improve the network performance, this paper proposes a Shape Prior Enhancement module that embeds shape prior

information as a learnable module and constrains it using a combination of explicit and implicit methods. This allows the model to focus on object regions in the image and obtain richer information about object features to improve the performance of object detection and image enhancement.

III. PROPOSED METHOD

The SG Detection network proposed in this paper is a joint network that incorporates the object detection and underwater image enhancement tasks. In this section, we firstly describe the overall network structure. Then we give a detailed description for the designed shape enhancement module. Finally a specific procedure is given for our proposed combined explicit and implicit edge constraint method.

A. Overall network structure

The SG Detection is shown in Fig.2. Our proposed SG Detection consists of three components, feature encoder network (backbone), feature processing network (Shape Enhancement) and decoder network (restoration Decoder and Detection Decoder). Specifically, we first use ResNet-50 [17] as the feature extraction network (backbone) to obtain features at different stages.

Then, in the feature processing stage, we embed the shape information as a learnable feature map into the SG detection network and interact with the different stages feature maps through the Shape Prior Enhancement module. We embed these prior features from the highest level, updating them layer by layer and passing them down. This allows each layer features to fully interact with the shape prior features to capture richer information about the object area.

Finally, the different stage features processed by the shape prior enhancement module are up-sampled layer by layer and constructed with lateral connections. The bottom layer features

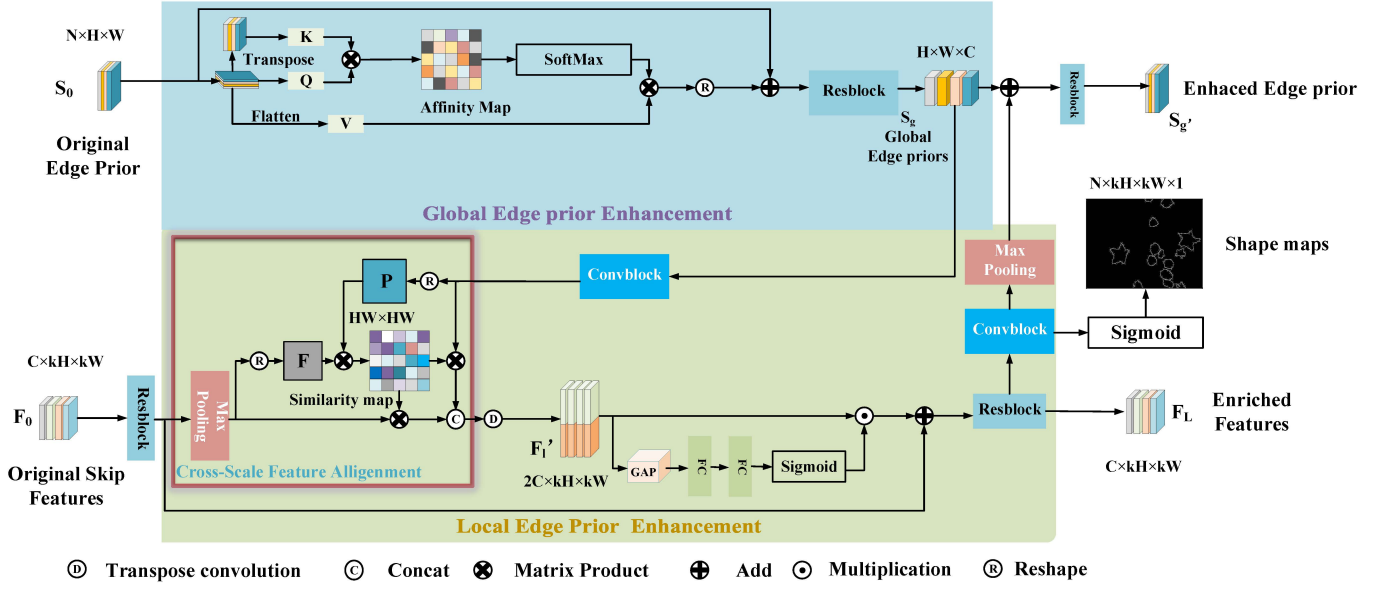


Fig. 3. The structure of Shape Prior Enhancement module. After the self-attention mechanism, the global shape prior is obtained and the better local features are acquired by interacting with the original skip features at each level. The full fusion of global and local features allows the network to get more accurate object detection results.

is used for image decoding to achieve clean underwater images, while the middle four layers features are connected to the detection head for object category and location.

B. Shape Prior Enhancement

a) Overview: Shape information refers to information such as edges, contours and sizes of objects. When we perform underwater object detection, these specific classes of objects usually have specific shapes and structures, and this shape and structure information can help us distinguish objects significantly; similarly, for underwater image enhancement, our mission is to better detect objects, and these shapes in turn provide a priori knowledge for image enhancement, which allows us to focus our enhancement on object regions. For this reason, we propose the Shape Prior Enhancement module to help our model better account for and exploit shape information for more accurate object detection.

The proposed Shape Prior Enhancement module in our model architecture that focuses on improving the shape edge features of the object region in the image, as shown in Fig.3. Global Shape Enhancement and Local Shape Enhancement are the two components of our Shape Prior Enhancement module. Global Shape Prior Enhancement (GSPE) is used to model long-range contextual dependencies to capture rich texture information associated with global regions, while Local Shape Prior Enhancement (LSPE) is used to capture local visual structural elements at multiple scales and to localise objects at different scales.

b) Global Shape Prior Enhancement: To explicitly introduce image shape prior information into our SG detection network, we define an $N \times H \times W$ shape prior feature S_o , where N is the number of categories, with the same resolution as the top-level feature map. Since a fixed receptive field of the

convolution will lead to an asymmetry in spatial information, we propose a Global Shape Prior Enhancement module that aims to capture the long-range dependency of the shape prior. Inspired by the ViT structure, we first generate a global shape prior with N channel interactions by using self-attention [18], and then perform a residual connection with the original shape prior features. The global shape prior between the N categories is constructed by the following formula, which describes the similarity and dependence between the different shape prior channels.

$$S_g = \text{Softmax}\left(\frac{Q_s(S_o) \times K_s(S_o)^T}{\sqrt{N}}\right) \times V_s(S_o) + S_o \quad (1)$$

Where, Softmax represents the normalization operator, Q_s and K_s denote the convolutional transform that projects S_o into the query and key vectors, T is the transpose operator, V_s denotes the convolutional transform that projects S_o into the value vector, $\times$ denotes the process of matrix multiplication, and S_g is the final enhanced global shape prior feature.

After the above operations, the S_g can establish long-term dependencies within S_o , including rough shape and position representations, as well as sufficient texture information of the global regions.

c) Local Shape prior Enhancement: The shape prior S_o , after GSPE processing, contains rich contextual information, but lacks local visual structure features and does not have accurate shape and contour information. Therefore, we design the Local Shape Prior Enhancement module to interact the original skip feature F_o with the enhanced shape prior feature S_g so to obtain more detailed local shape information.

First, we perform a convblock operation on the extended shape prior to expand the channel dimension to be consistent with F_o . Then, the Cross-Scale Feature Alignment(CSFA)

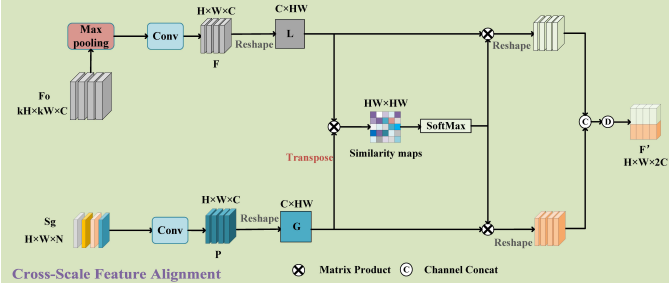


Fig. 4. The detail of Cross-Scale Alignment module.

module is designed to achieve deep interaction between local features and the shape prior, as shown in Fig.4. Finally, the SE Net [19] structure is used to assign different channel weights, which allows some features to be highlighted to achieve a better representation of the object region.

Cross-Scale Feature Alignment. Among them, the CSFA is the core of the interaction between local and global features, and its design is motivated by the fact that there is a semantic gap between features of different levels, which can lead to inefficient feature fusion. Therefore, in our cross-scale feature alignment module, we calculate the correlation between pixels in different feature maps by matrix multiplication, and then use this correlation as the weight vector of the feature maps at each level to bridge the semantic gap and achieve the effective fusion of features at different scales. The specific computational process is as follows.

First, the local feature F_o is operated by maxpooling and the convolution block module to obtain a feature map F of the size $H \times W \times C$. Similarly, we use a convolutional block operation on S_g to obtain the same size feature P . P and F are expressed as follows:

$$\begin{aligned} F &= SeqConv_{3 \times 3}^{2D}(Maxpooling(F_o)) \\ P &= SeqConv_{3 \times 3}^{2D}(S_g) \end{aligned} \quad (2)$$

Where, $SeqConv$ refers to sequentially performs the 2D 3×3 convolution with Leaky ReLU activation.

Then, reshaping operations are performed at the same time on F and P in order to scale their dimensions to G and L of $C \times HW$. Based on G and L , we define the similarity matrix S . The similarity matrix S is defined as follows:

$$s_{ji} = \frac{\exp(P_i \cdot Q_j)}{\sum_{i=1}^M \exp(P_i \cdot Q_j)}, \quad (3)$$

where s_{ji} measures the correlation between the i th position in the lower feature map and the j th position in the higher feature map. M denotes the number of pixels. The greater the similarity between the feature representations of pixels at two locations, the higher the correlation between them.

We transpose G for matrix multiplication and apply a softmax layer to compute the spatial attention map $S \in HW \times HW$.

$$S = softmax(\frac{G^T \times L}{\sqrt{N}}) \quad (4)$$

Finally, we perform matrix multiplication with G and L , based on S . The results of the two are channel concat to obtain the enhanced feature F'_l .

$$F'_l = U \cdot Concat(reshape(G \times S) + reshape(L \times S)) \quad (5)$$

Where $U \cdot$ refers to the transpose convolution following by a 1×1 convolution. $F' \in H \times W \times 2C$. The concat features contain local and global information, and the importance of the features of each channel is not the same, so it is necessary to refine their features and highlight the important ones.

$$F_L = Resconv_{3 \times 3}^{2D}(\sigma(W_2 \delta(W_1(GAP(F'_l))) \times F'_l) + F_o) \quad (6)$$

Where, σ represents for sigmoid function, W_1, W_2 refers to two Full Connections, $Resconv$ means sequentially conducts the 2D 3×3 convolution with Resblock and Leaky ReLU activation function. δ represents the Relu function, and F'_l represents the feature after CSA.

After global and local enhancement, we also update the shape prior features. We denote the output shape prior feature as S'_g , which is updated by follows:

$$S'_g = SeqConv_{3 \times 3}^{2D}(D \cdot (SeqConv(F_L)) + S_g) \quad (7)$$

Where, $D \cdot$ means Maxpooling operation. Finally, we add local information to the global shape prior features. This provides us with more detailed local information.

In summary, our proposed shape prior enhancement module first performs a global shape prior module to obtain a coarse shape prior with adequate global texture information. It is also enhanced by local shape prior to obtain more refined shape information. In addition, the original skipped features are further enhanced by interacting with the global shape prior, which will facilitate the generation of features with a discriminative shape representation and global context, and ultimately obtain more accurate object detection performance.

C. Shape Prior Constraints

Currently, there are different ways to integrate prior information with models, such as atlas-based models [20], statistical-based models [21], and U-NET-based models [22]. The first two approaches have poor generalization performance, while U-NET based models have better generalization performance.

In addition, based on the embedding method, it can be divided into explicit constraints and implicit constraints. Explicit constraints refer to the direct combination of interpretable information, such as shape and structure, with the model. It has the advantage of strong interpretability, but is not easy to obtain. Implicit constraints, on the other hand, are achieved by adding specific prior information, such as a loss function or regularisation term, to the model. For example, in the U-Net image segmentation model, the DICE [23] loss function is often used to force the model to focus on contour information. The disadvantage of implicit constraints is poor interpretability.

Therefore, we propose a joint explicit and implicit shape a priori constraint method based on the SG detection network

structure, Shape Prior Constraint (SPC), which can help us to better consider and exploit the shape information for more accurate object detection. It explicitly takes shape prior information as input and uses this information to guide the model to better detect objects. However, this constraint is not entirely explicit, as the prior information is not obtained directly from the image object, but from training through the design of a loss function.

In our model, as shown in Fig.3, we add a new sigmoid output after each layer of the feature map to classify each pixel point in the image into two categories (edge or non-edge), and then calculate the edge pixel cross-entropy loss to make our model pay more attention to pixels at the edge of the object. The calculation process is as follows

$$\ell = - \sum_{i,j} (Y_{i,j} \alpha \log(E_{i,j})) + (1 - Y_{i,j})(1 - \alpha) \log(1 - E_{i,j}) \quad (8)$$

Where E, Y is the generated shape map and the corresponding ground-truth, $E_{i,j}, Y_{i,j}$ are the element of E, Y ; $\alpha = |Y^-| / (|Y^-| + |Y^+|)$ indicates the percentage of negative pixel samples (i.e., non-edge points). In addition, we obtain multiple shape map outputs $S_g^1, S_g^2, \dots, S_g^4$, obtained by feeding the original skip features at different scales into the shape prior enhancement module. The overall Shape loss function is as follows:

$$L_s = \frac{1}{4} \sum_{k=1}^4 \ell(S_g^k, Y) \quad (9)$$

When explicit and implicit constraints are used jointly, our proposed SG Detection Net can focus more on the processing of object shape information, and thus to achieve better object detection results.

D. The Overall Loss Function

Our model implements two tasks, object detection and underwater image enhancement, so the overall loss of the model is as follows:

$$L_{\text{Total}} = L_{\text{de}} + L_{\text{re}} + L_s, \quad (10)$$

Where, L_{de} represents the detection loss, L_{re} refers to the image reconstruct loss and L_s stands for the Shape loss which is detailed introduced in Section.III-C.

The L_{de} is expressed as follows:

$$L_{\text{de}} = L_{\text{Cls}} + \lambda L_{\text{IoU}} + L_{\text{FOCAL}}, \quad (11)$$

Where λ is the hyperparameter, which is usually set to 5 in object detection training stages. L_{Cls} is the classification loss, L_{IoU} is the regression loss, and L_{FOCAL} stands for Focal loss [24].

The L_{re} loss is used to evaluate the similarity between the restored image and the Ground-Truth. It is formulated as follows:

$$L_{\text{re}} = \frac{1}{n} \sum_{i=1}^n \left(\hat{\mathbf{X}}_i - \mathbf{X}_i \right)^2 \quad (12)$$

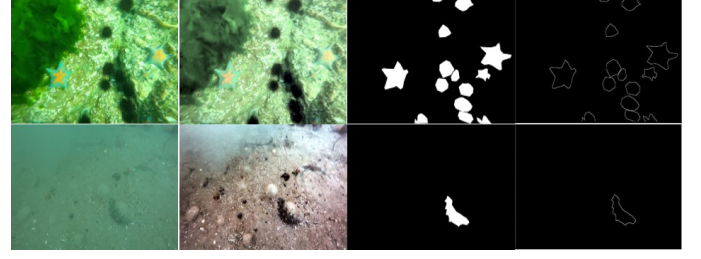


Fig. 5. The sample of dataset we used, along with the corresponding shape mask and shape edge Ground-Truth.

TABLE I
SUMMARY OF DATASETS USED IN OUR EXPERIMENT

Dataset	Train	Test
UTDAC	5168	1293
URPC2021	5320	2813
Total	10488	4106

Where, $\hat{\mathbf{X}}_i$ represents the enhanced image, $\mathbf{X}_i$ refers to the Ground-truth image, n stands for the batch size. We train the image enhancement network by computing the MSE loss.

In order to make our model finally obtain the best object detection performance, we weight the three losses in L_{Total} by assigning different weights λ_1, λ_2 and λ_3 , respectively. So, the overall loss function is formulaed as follows:

$$L_{\text{Total}} = \lambda_1 L_{\text{de}} + \lambda_2 L_{\text{re}} + \lambda_3 L_s, \quad (13)$$

Eventually, we are surprised to find that when the three weights are configured as 0.2, 0.8 and 0.2, the enhancing effect on object detection is most obvious. The detailed experimental validation process is shown in Section.IV-D.

IV. EXPERIMENTS

A. Datasets and Evaluation Metrics

We have conducted a large number of experiments to verify the effectiveness of our method. For the dataset, we used the URPC2021 dataset. This is an underwater optical image data for underwater object detection. The dataset has a total of 7600 underwater object images divided into four categories: scallop, echinus, holothurian, and starfish. UTDAC2020 is an underwater dataset, which stands for Underwater Target Detection Algorithm Competition 2020. There are 5168 training images and 1293 validation images. It has the same categories as the URPC dataset. The image data are collected through the field and some of the objects are similar to the environment, which makes the overall object detection difficult. The samples of URPC are shown in Fig.5, and the dataset division strategy is shown in Table I.

Since our model is ultimately to achieve high accuracy underwater target detection, we mainly use mAP as the main model evaluation metric. The mAP we use is mainly computed as COCO. In the evaluation process, COCO evaluates the access points under a total of 10 IoUs with different IoU ratios (IoU) [0.5:0.05:0.95], and finally takes the average of

TABLE II

COMPARISON OF MAP METRICS FOR UNDERWATER IMAGE DATA OBJECT DETECTION. THE BEST RESULTS ARE SHOWN IN BOLD. THE EXPERIMENTAL RESULTS SHOW THAT THE SG DETECTION NETWORK DESIGNED IN THIS PAPER COMBINED WITH DIFFERENT DETECTORS CAN ACHIEVE IMPROVED DETECTION PERFORMANCE AND INFERENCE SPEED WITH SIGNIFICANT IMPROVEMENT COMPARED TO EXISTING UNDERWATER OBJECT DETECTORS AND PRE-PROCESSING OBJECT DETECTION METHODS.

method	backbone	mAP (URPC/UTDAC)	Parameters(M)	FPS
SOTA Detectors:				
YOLO_V5 [28]	DarkNet	0.356/0.384	54.2	58.4
YOLOX [29]	CSPDarkNet	0.371/0.432	60.3	54.8
Faster R-CNN [30]	ResNet-50	0.401/0.428	122.14	29.4
Faster R-CNN [30]	ResNet-101	0.414/0.445	131.6	28.5
Libra R-CNN [31]	ResNet-50	0.420/0.458	146.60	24.4
Libra R-CNN [31]	ResNet-101	0.435/0.454	164.9	22.5
Sparse R-CNN [32]	ResNet-50	0.425/0.443	212.5	18.6
DiffusionDet [33]	ResNet-50	0.428/0.433	314.4	12.2
FCOS [34]	ResNet-50	0.394/0.428	117.4	27.8
VFNet [35]	ResNet-101	0.416/0.440	141.4	23.1
Underwater Object Detector:				
Boosting R-CNN [36]	ResNet-50	0.433/0.462	143.55	23.0
Boosting R-CNN [36]	ResNet-101	0.443/0.456	160.55	20.3
RoiAtt [37]	ResNet-50	0.462/0.460	183.45	18.1
RoiAtt [37]	ResNet-101	0.473/0.467	192.45	17.8
SWIPENet [5]	ResNet-50	0.451/0.471	166.7	19.3
SWIPENet [5]	ResNet-101	0.466/0.478	179.4	18.9
Ours:				
SG Detection(YOLOX Body)	CSPDarknet	0.471/0.483	124.4	31.8
SG Detection(Faster R-CNN Body)	ResNet-50	0.476/0.488	227.7	17.4
SG Detection(Faster R-CNN Body)	ResNet-101	0.482/0.491	244.6	14.1

the access points under these thresholds as the result, denoted as $mAP@[.5, .95]$.

To further validate the enhancement performance of our model for underwater images, we use three non-reference metrics, UCIQE [25], UIQM [26] and FDUM [27]. UCIQE is a linear combination of chroma, saturation and contrast, designed to quantify non-uniform colour cast, blur and contrast. UIQM and FDUM measure colourimetry, sharpness and contrast from different perspectives.

B. Implement Details

Our approach requires paired underwater images (original underwater image and corresponding clear image) and precise information about the shape of the object. To obtain the paired images, this paper relies on the latest algorithms such as CWR, UGAN, etc., together with subjective image enhancement methods such as white balance, contrast enhancement, smoothing and denoising to process the images and select the "clean" images that favor human eyes as the corresponding clear images of the ground truth. In order to obtain the prior information about the shape of the object more accurately, the object region is first extracted based on the annotated file, and then the canny operator is run to obtain the shape contour of the object. After completing the above steps, the

paired underwater images and the shape prior information are obtained and fed into our proposed SG Detection network for training. The data used for our training and the corresponding clean and shape images are shown in Fig.5.

As for the experimental network part, our SG Detection network is mainly combined with YOLOX and Faster R-CNN to share the backbone network. And all the backbone networks are pre-trained on the ImageNet dataset. Specifically, our network was trained with stochastic gradient descent (SGD) for 50k iterations with an initial learning rate of 0.0001 and mini-batches of 4 images. At 20k and 40k iterations, the learning rate decreases by a factor of 10, respectively. And the weight decay and momentum were set to 0.0001 and 0.9, respectively. all models were trained based on NVIDIA RTX 3090 with 24GB memory.

C. Comparision Study

1) *Comparison of Underwater Image Object Detection:*
We carry out experiments comparing our own model, with various state of the art(SOTA) object detector : Faster R-CNN [30] , YOLOX [29], Libra R-CNN [31], Spares R-CNN [32], FCOS [34], DiffusionDet [33], VF [35], pre-processing object detection algorithm like MRSCR [38], GUDCP [39], UWGAN [9], UGAN [40],CWR [41], U-Transformer [42]

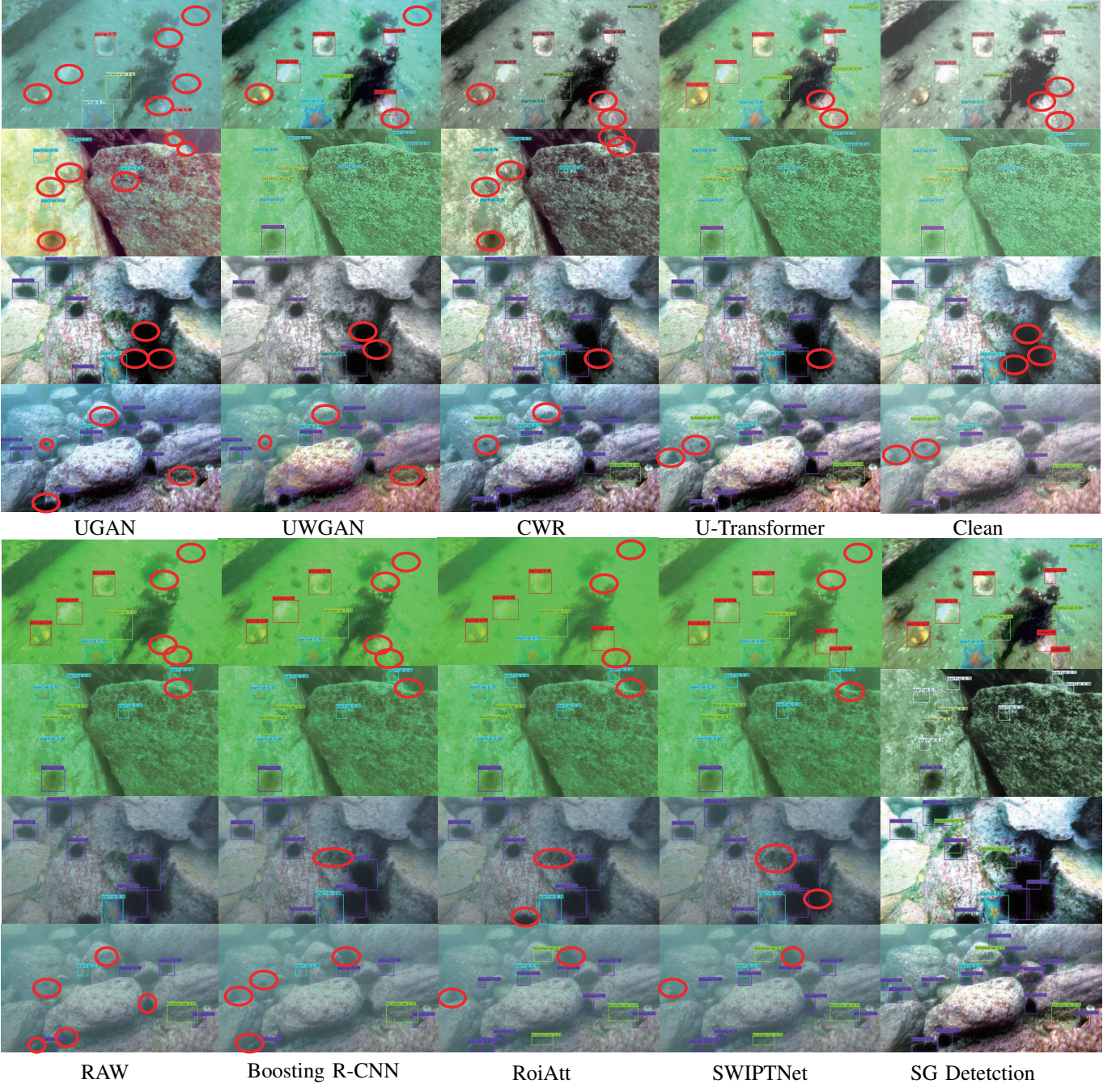


Fig. 6. Test samples of different methods. In terms of results, the method in this paper has the most accurate object detection results and a good visual experience. For the other methods, although the visual experience is improved, the accuracy of object detection is not improved.

and underwater object detector: RoiAtt [37], Boosting R-CNN [36], SWIPENet [5], and we use mAP as an evaluation metric for object detection performance.

The comparison results of mAP are shown in Table.II. From the results we can conclude that:

(a) Our method has significant performance advantages. Both with the state-of-the-art object detectors and underwater object detectors, our method has a performance gain.

(b) Our method also shows generality. When our method is combined with different detectors such as YOLOX, Faster R-CNN, etc., the mAP performance metrics can be improved

by up to 10%, which proves that our proposed structure is applicable to a wide range of detectors.

The visual results of object detection between our method and various methods are shown in Fig.6. It can also be seen from the figure that the miss and false detection rates of our method are significantly reduced, and its object detection performance is greatly improved.

2) *Comparison of Underwater Image Enhancement:* Our SG detection network implements object detection and underwater image enhancement simultaneously, so we also compare it with various SOTA underwater image enhancement methods

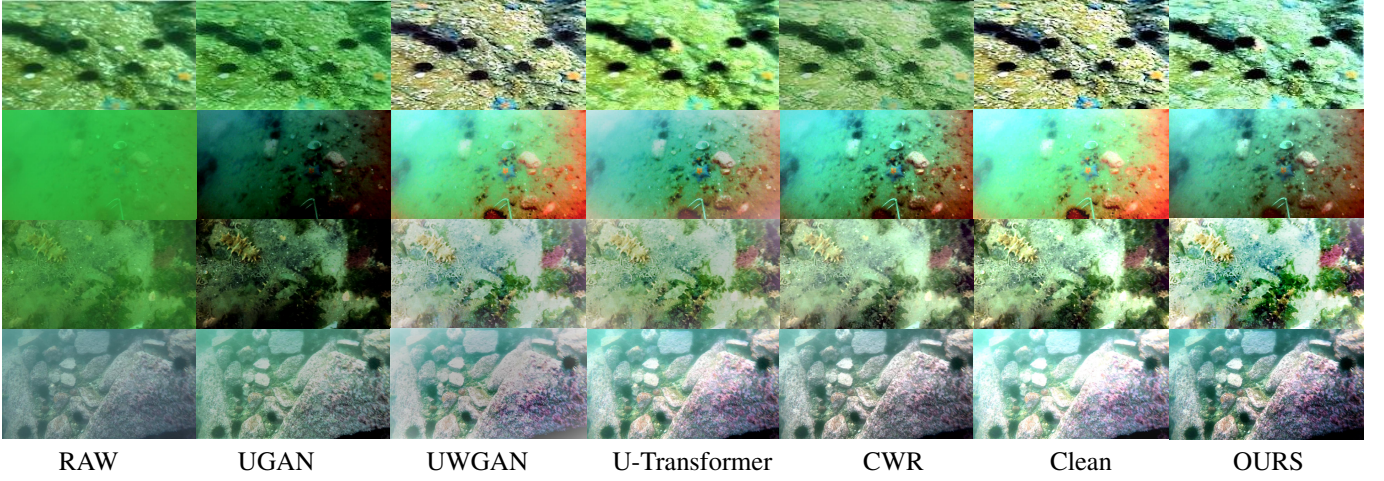


Fig. 7. The visual comparison results of several deep learning based image enhancement algorithms. Compared to several state-of-the-art underwater image enhancement algorithms, our approach demonstrates notable performance in terms of visual effects.

TABLE III
THE COMPARISON RESULTS OF IMAGE ENHANCEMENT PERFORMANCE ON URPC AND UTDAC DATASET.

Datasets	URPC			UTDAC		
	UCIQE↑	UIQM↑	FDUM↑	UCIQE↑	UIQM↑	FDUM↑
RAW	0.41	0.43	0.33	0.42	0.47	0.35
Clean	0.64	0.86	0.59	0.64	0.79	0.59
MRSCR [38]	0.37	0.42	0.29	0.45	0.49	0.40
GUDCP [39]	0.43	0.44	0.29	0.54	0.47	0.36
UGAN [40]	0.49	0.59	0.41	0.56	0.67	0.46
UWGAN [9]	0.53	0.63	0.47	0.55	0.74	0.41
CWR [41]	0.66	0.79	0.51	0.66	0.78	0.56
U-Transformer [42]	0.68	0.84	0.54	0.66	0.79	0.59
Ours	0.65	0.88	0.57	0.71	0.75	0.56

TABLE IV
THE MAP COMPARISON OF PRE-PROCESSING BASED METHODS.

method	mAP (URPC/UTDAC)	Parameters(M)	FPS
RAW	0.380/0.384	122.4	29.4
Clean	0.465/0.478	122.4	29.4
MSCR [38]	0.375/0.389	122.4	11.6
GUDCP [39]	0.376/0.401	122.4	18.1
UGAN [40]	0.428/0.441	146.6+122.4	11.2
UWGAN [9]	0.431/0.450	186.2+122.4	9.8
CWR [41]	0.461/0.475	221.4 + 122.4	8.1
U-Transformer [42]	0.467/0.477	383.4 + 122.4	5.1
SG Detection	0.482/0.491	244.6	14.1

to validate the underwater image enhancement performance of our method. Since the URPC and UTDAC datasets we use do not have corresponding reference images, we use UCIQE, UIQM and FDUM as evaluation metrics to evaluate the enhancement effect from a subjective perspective.

The results of the comparison experiments are shown in the Table.III. From the comparison results, we can conclude that our method still shows competitive performance compared to the SOTA underwater image enhancement algorithms, like CWR and U-Transformer, even though our method focuses more on the object detection performance.

The visual comparison results are shown in Fig.7. From the figure, it can be observed that our method is still better in terms of visual results, both in terms of colour saturation, brightness and sharpness, compared to the most advanced deep learning based methods such as UGAN, UWGAN, U-Transformer and CWR. More importantly, in the object regions, our method recovers better, with more detail and texture in the object.

3) *Analysis of Underwater Image Enhancement and Object Detection:* We compare the object detection and underwater image enhancement performance separately. On this basis, we compare the object detection performance of images processed by different underwater image enhancement algorithms, so as to further explore the connection between underwater image

detection and underwater image enhancement. The comparison results are shown in Table.IV.

From this we can analyse that, firstly, there is no strict positive correlation between underwater image enhancement and object detection. For example, the original image has a better object detection performance compared to methods such as MRSCR and GUDCP.

Another point is that our proposed method is significantly improved in terms of both model size and inference speed. Existing preprocessing-based methods consume a lot of computational time and resources, while our method significantly improves the inference time while improving the performance.

D. Ablation Study

Effect on Underwater image object detection. We also performed detailed ablation experiments to validate the effectiveness of our proposed method. First, the object detection performance is compared. We used YOLOX and Faster R-CNN as the detection head, combined with our proposed SG Detection network, and gradually added various improvements to evaluate the effectiveness of our method, the results of which are shown in Table.V.

TABLE V

DETECTION RESULTS OF ABLATION STUDY. UNET STANDS FOR DETECTION NETWORK BASE WITH ANOTHER IMAGE GENERATION DECODER; SPE STANDS FOR SHAPE PRIOR ENHANCEMENT MODULE; SPC STANDS FOR SHAPE PRIOR CONSTRAINT.

method	backbone	mAP(URPC/UTDAC)	Parameters	FPS
Faster R-CNN	ResNet-101	0.414/0.445	131.6	24.5
Faster R-CNN +UNET	ResNet-101	0.431/0.450	146.6	22.5
Faster R-CNN +SPE	ResNet-101	0.461/0.464	144.6	22.8
Faster R-CNN +SPC	ResNet-101	0.443/0.459	148.6	22.3
Faster R-CNN +SG Detection	ResNet-101	0.482/0.491	164.6	20.1
YOLOX	CSPDarknet	0.371/0.432	60.3	54.8
YOLOX +UNET	CSPDarknet	0.422/0.446	70.4	51.1
YOLOX +SPE	CSPDarknet	0.451/0.469	71.6	50.4
YOLOX +SPC	CSPDarknet	0.432/0.440	64.6	52.4
YOLOX +SG Detection	CSPDarknet	0.471/0.483	84.4	46.8

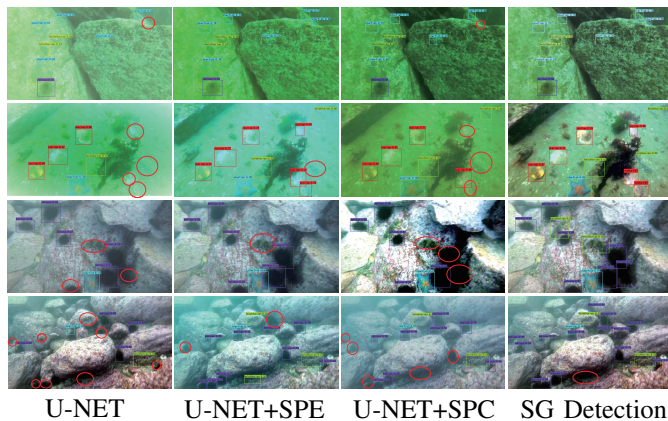


Fig. 8. Some examples of our object detection ablation experiments.

From the results, we can conclude that our model can achieve up to 10% performance improvement in mAP values (with YOLOX as the detector) and significant improvement on different frameworks, demonstrating the effectiveness and generality of our model. Furthermore, we can conclude that each of our proposed modules provides different levels of performance improvement for object detection. The two object detection networks, YOLOX and Faster R-CNN, show the greatest performance improvement after adding the shape prior enhancement module proposed in this paper. This may also indicate that for the object detection task, the performance improvement is greater when more a priori information can be added. The visual ablation results of object detection are shown in Fig.8.

Effect on Underwater image Enhancement. In addition, we also perform ablation experiments on the underwater image

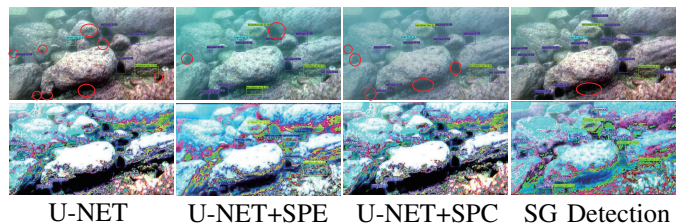


Fig. 9. Feature map visualisation results. Brighter colours indicate higher feature correspondence. The red circle represents the location of the missed target.

TABLE VI

ABLATION EXPERIMENTS RESULTS ON IMAGE ENHANCEMENT.

Datasets	URPC			UTDAC		
	UCIQE↑	UIQM↑	FDUM↑	UCIQE↑	UIQM↑	FDUM↑
U-NET	0.44	0.51	0.36	0.45	0.49	0.39
U-NET+SPE	0.56	0.69	0.51	0.59	0.70	0.51
U-NET+SPC	0.51	0.64	0.44	0.49	0.62	0.48
U-NET+SPE+SPC	0.65	0.88	0.57	0.71	0.75	0.56

enhancement performance to progressively validate the effect of each module on performance enhancement. The results are presented in Table.VI.

From the results, we can conclude that all the different modules we propose have an improvement effect on the visual results, with the SPE module being the most obvious for visual enhancement. Although our use of the shape prior is mainly to improve object detection, the shape prior also provides a prior for underwater image enhancement, so the focus of our model is on the target region. With the combined use of SPE+SPC, we obtain the highest performance improvement, confirming the superiority of the combination of explicit and

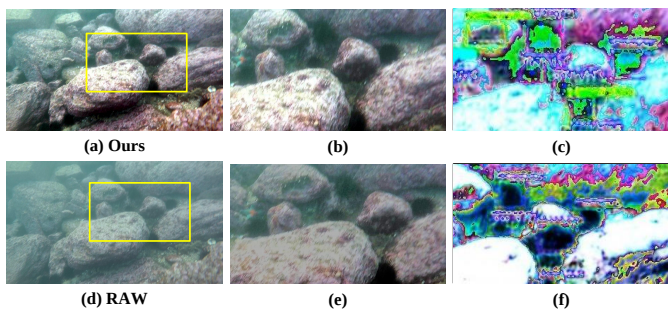


Fig. 10. The comparison result between our method and raw image. The yellow box represents the object region, and (c) and (f) represent the feature response at the object location. It can be seen that our method is more effective in recovering the object region.

TABLE VII
THE ABLATION EXPERIMENT MAP RESULTS OF LOSS WEIGHTS λ_1
AND λ_2 . WE USED YOLOX AS THE BASE DETECTOR.

$\lambda_1 \& \lambda_2 \& \lambda_3$	1 & 1 & 1	0.7 & 0.3 & 0.3	0.5 & 0.5 & 0.5	0.2 & 0.8 & 0.2	0.4 & 0.6 & 0.4	0.1 & 0.9 & 0.1
URPC	0.454	0.468	0.466	0.471	0.443	0.441
UTDAC	0.460	0.477	0.455	0.483	0.454	0.446

implicit constraints together.

In addition, we also perform an experimental comparison of the ablation of feature map responses, the results of which are shown in Fig.9. The brightest colour represents the highest feature response. From the feature response results, our method has the highest feature response in the object region, which also verifies the ability of our model to locate the target region and extract richer information about the target region.

Effect of weights in Loss Function. To improve the detection performance of SG detection, we use an effective unified loss function that includes the object detection loss L_{de} and the underwater image enhancement loss L_s . Accordingly, two loss weights (λ_1 , λ_2 and λ_3) are used to balance the performance of these loss functions. For and , detailed experiments are performed on the dataset to ensure their optimal values, as shown in Table.VII. As observed, the image restoration loss is very helpful in improving the detection ability of the proposed method. Therefore, when $\lambda_1=0.2$ and $\lambda_2=0.8$ are set in our experiments, the performance of SG detection is the best.

E. Discussion

We have conducted detailed experiments to compare the performance of our method with existing SOTA methods in both object detection and underwater image enhancement. From the comparison results, our method achieves the best performance in object detection and is also very effective in visual enhancement of underwater images.

In addition, from the table and the visual comparison results, both MSCR and GUDCP reduce the performance of object detection to a certain extent, which also proves that there is not necessarily a positive correlation between image enhancement and object detection. This suggests that rather than enhancing the feature information required for object detection, these enhancement methods introduce noise.

Therefore, this positively illustrates the need to combine the two efforts of underwater object detection and image

enhancement: with underwater image enhancement, we obtain the "clean" features needed for object detection; on the other hand, the object detection results also guide the direction of the underwater image enhancement, rather than just improving the visual experience. The experimental metrics as well as the visual experience results show that our approach significantly improves the object detection performance while enhancing the visual experience.

V. CONCLUSION

In this paper, from the current research status and challenges of existing underwater image object detection, we point out that the existing underwater object detection mostly relies on pre-processing to improve the image quality and then achieve the purpose of improving the object detection accuracy. However, this consumes more computational resources and the pre-processing methods are not universally applicable. Therefore, we innovatively combine two tasks, object detection and underwater image enhancement, and propose the SG detection network, which shares clean features through the underwater image enhancement process, thus facilitating object detection. Meanwhile, the object shape prior is embedded in our network as a learnable module, and the global and local shape feature modelling capabilities are enhanced by the shape prior enhancement module, which interacts with the original features. To further enhance the network's ability to use the shape prior, we constrain it both explicitly and implicitly in a way that allows object recognition accuracy to be further improved. We conduct experiments on both URPC and UTDAC datasets, and our method achieves large performance gains in combination with different detectors, demonstrating the effectiveness and generality of our model and showing a competitive speed advantage over existing methods.

However, in general, object detection in underwater images is still a challenging problem. Our method improves the shortcomings of existing methods to some extent, but the overall accuracy still lags far behind natural images. In addition, the existing underwater object dataset is still relatively sparse, the detection categories are relatively discrete, and the development of a general underwater object detection model is still some way off.

REFERENCES

- [1] C. Liu, Z. Wang, S. Wang, T. Tang, Y. Tao, C. Yang, H. Li, X. Liu, and X. Fan, "A new dataset, poisson gan and aquanet for underwater object grabbing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 2831–2844, 2021.
- [2] 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.
- [3] L. Jiang, Y. Wang, Q. Jia, S. Xu, Y. Liu, X. Fan, H. Li, R. Liu, X. Xue, and R. Wang, "Underwater species detection using channel sharpening attention," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4259–4267.
- [4] K. Li, L. Wu, Q. Qi, W. Liu, X. Gao, L. Zhou, and D. Song, "Beyond single reference for training: underwater image enhancement via comparative learning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

- [5] L. Chen, Z. Liu, L. Tong, Z. Jiang, S. Wang, J. Dong, and H. Zhou, "Underwater object detection using invert multi-class adaboost with deep learning," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.
- [6] Z. Wang, H. Chen, H. Qin, and Q. Chen, "Self-supervised pre-training joint framework: Assisting lightweight detection network for underwater object detection," *Journal of Marine Science and Engineering*, vol. 11, no. 3, p. 604, 2023.
- [7] 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.
- [8] C. Fu, X. Fan, J. Xiao, W. Yuan, R. Liu, and Z. Luo, "Learning heavily-degraded prior for underwater object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.
- [9] J. Li, K. A. Skinner, R. M. Eustice, and M. Johnson-Roberson, "Watergan: Unsupervised generative network to enable real-time color correction of monocular underwater images," *IEEE Robotics and Automation letters*, vol. 3, no. 1, pp. 387–394, 2017.
- [10] M. Hou, R. Liu, X. Fan, and Z. Luo, "Joint residual learning for underwater image enhancement," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 4043–4047.
- [11] R. Mandal, R. M. Connolly, T. A. Schlacher, and B. Stantic, "Assessing fish abundance from underwater video using deep neural networks," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–6.
- [12] 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.
- [13] S. Qi, J. Du, M. Wu, H. Yi, L. Tang, T. Qian, and X. Wang, "Underwater small target detection based on deformable convolutional pyramid," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 2784–2788.
- [14] K. Hu, F. Lu, M. Lu, Z. Deng, and Y. Liu, "A marine object detection algorithm based on ssd and feature enhancement," *Complexity*, vol. 2020, pp. 1–14, 2020.
- [15] T. Katayama, T. Song, T. Shimamoto, and X. Jiang, "Gan-based color correction for underwater object detection," in *OCEANS 2019 MTS/IEEE SEATTLE*. IEEE, 2019, pp. 1–4.
- [16] T. Guo, Y. Wei, H. Shao, and B. Ma, "Research on underwater target detection method based on improved msrnp and yolov3," in *2021 IEEE International Conference on Mechatronics and Automation (ICMA)*. IEEE, 2021, pp. 1158–1163.
- [17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [18] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," *arXiv preprint arXiv:1803.02155*, 2018.
- [19] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
- [20] M. Cabezas, A. Oliver, X. Lladó, J. Freixenet, and M. B. Cuadra, "A review of atlas-based segmentation for magnetic resonance brain images," *Computer methods and programs in biomedicine*, vol. 104, no. 3, pp. e158–e177, 2011.
- [21] D. A. Reynolds *et al.*, "Gaussian mixture models," *Encyclopedia of biometrics*, vol. 741, no. 659–663, 2009.
- [22] X. You, J. He, J. Yang, and Y. Gu, "Learning with explicit shape priors for medical image segmentation," *arXiv preprint arXiv:2303.17967*, 2023.
- [23] T. A. Soomro, A. J. Afifi, J. Gao, O. Hellwich, M. Paul, and L. Zheng, "Strided u-net model: Retinal vessels segmentation using dice loss," in *2018 Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 2018, pp. 1–8.
- [24] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.
- [25] M. Yang and A. Sowmya, "An underwater color image quality evaluation metric," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 6062–6071, 2015.
- [26] K. Panetta, C. Gao, and S. Agaian, "Human-visual-system-inspired underwater image quality measures," *IEEE Journal of Oceanic Engineering*, vol. 41, no. 3, pp. 541–551, 2015.
- [27] N. Yang, Q. Zhong, K. Li, R. Cong, Y. Zhao, and S. Kwong, "A reference-free underwater image quality assessment metric in frequency domain," *Signal Processing: Image Communication*, vol. 94, p. 116218, 2021.
- [28] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "Tph-yolov5: Improved yolov5 based on transformer prediction head for object detection on drone-captured scenarios," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 2778–2788.
- [29] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.
- [30] 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.
- [31] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra r-cnn: Towards balanced learning for object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 821–830.
- [32] P. Sun, R. Zhang, Y. Jiang, T. Kong, C. Xu, W. Zhan, M. Tomizuka, L. Li, Z. Yuan, C. Wang *et al.*, "Sparse r-cnn: End-to-end object detection with learnable proposals," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 454–14 463.
- [33] S. Chen, P. Sun, Y. Song, and P. Luo, "Diffusiondet: Diffusion model for object detection," *arXiv preprint arXiv:2211.09788*, 2022.
- [34] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9627–9636.
- [35] A. Ahmed, P. Tangri, A. Panda, D. Ramani, and S. Karmakar, "Vfnet: A convolutional architecture for accent classification," in *2019 IEEE 16th India Council International Conference (INDICON)*. IEEE, 2019, pp. 1–4.
- [36] P. Song, P. Li, L. Dai, T. Wang, and Z. Chen, "Boosting r-cnn: Reweighting r-cnn samples by rpn's error for underwater object detection," *Neurocomputing*, vol. 530, pp. 150–164, 2023.
- [37] X. Liang and P. Song, "Excavating roi attention for underwater object detection," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 2651–2655.
- [38] D. J. Jobson, Z.-u. Rahman, and G. A. Woodell, "A multiscale retinex for bridging the gap between color images and the human observation of scenes," *IEEE Transactions on Image processing*, vol. 6, no. 7, pp. 965–976, 1997.
- [39] Z. Liang, X. Ding, Y. Wang, X. Yan, and X. Fu, "Gudcp: Generalization of underwater dark channel prior for underwater image restoration," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 7, pp. 4879–4884, 2021.
- [40] C. Fabbri, M. J. Islam, and J. Sattar, "Enhancing underwater imagery using generative adversarial networks," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 7159–7165.
- [41] J. Han, M. Shoeiby, T. Malthus, E. Botha, J. Anstee, S. Anwar, R. Wei, M. A. Armin, H. Li, and L. Petersson, "Underwater image restoration via contrastive learning and a real-world dataset," *Remote Sensing*, vol. 14, no. 17, p. 4297, 2022.
- [42] L. Peng, C. Zhu, and L. Bian, "U-shape transformer for underwater image enhancement," *IEEE Transactions on Image Processing*, 2023.