國立政治大學理學院資訊科學系
碩士學位論文
Department of Computer Science
College of Science
National Chengchi University (NCCU)
Master Thesis

水下顯著物目標檢測
Underwater Salient Object Detection

指導教授：彭彥璁 博士
Advisor: Yan-Tsung Peng, Ph.D.
研究生: 林祐丞
Yu-Cheng Lin

中華民國 110 年 9 月
September, 2021

DOI:10.6814/NCCU202101682

摘要

顯著物偵測 (SOD) 在深度學習架構下已達到相當先進的成果。然而既有的研究大部分都專注在陸上場景，水下場景的顯著物偵測仍有待發展。在這篇論文中，我們蒐集並標註一水下顯著物資料集，用以驗證我們提出的模型方法。本論文中提出二種方法提昇顯著物偵測準確度。第一，我們先嘗試利用了水下影像模糊特性，幫助深度網路學習顯著物偵測。首先，我們會從原圖計算生成模糊圖，並與原圖一起輸入模型抽取特徵並融合，藉以提昇顯著物偵測準確度。第二，我們提出基於模糊圖對原圖增益作調整的一種資料擴增的方法。實驗結果顯示在最新顯著物偵測模型上，使用這兩種方法，皆可有效提昇效能。而提出的資料擴增方法的成效，比第一種方法更為有效。

關鍵詞：水下顯著物偵測、資料擴增、深度學習。

DOI:10.6814/NCCU202101682

Abstract

Salient object detection (SOD) has achieved state-of-the-art performance with the help of deep networks. However, most of the works focus on terrestrial scenes, and underwater scenes for SOD are still unexplored. In this work, we propose two practical approaches to boost the performance of underwater SOD. First, we utilize image blurriness to enable a more accurate SOD prediction. The blurriness map is calculated based on the input image, fed into the model with the input, and fused with the input image to produce the saliency map. Next, we propose a data augmentation method called FocusAugment for underwater SOD, which adjusts the image intensity based on the blurriness map. We can modify images by highlighting less blurred regions or enlarging the difference of pixels based on the blurriness maps. We test underwater SOD by the proposed dataset collected and annotated by ourselves for evaluation. The experimental results show that both of our approaches work; moreover, the presented FocusAugment works better than the blurriness-guided SOD model.

Keywords: Underwater salient object detection, data augmentation, deep learning.

DOI:10.6814/NCCU202101682

Table of Contents

DOI:10.6814/NCCU202101682

DOI:10.6814/NCCU202101682

List of Figures

DOI:10.6814/NCCU202101682

List of Tables

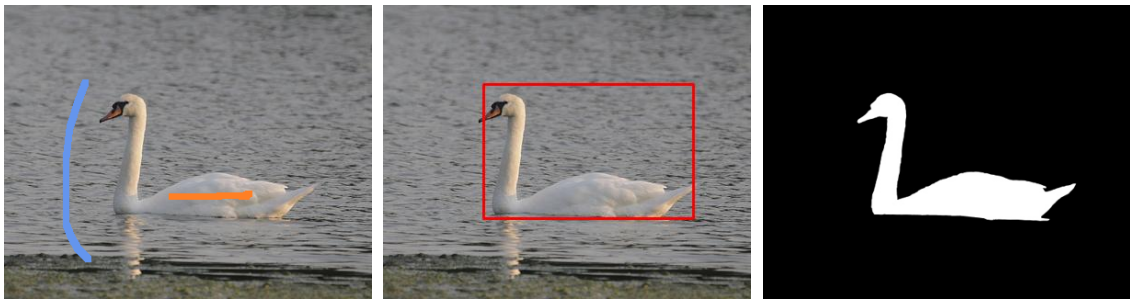
DOI:10.6814/NCCU202101682

1 INTRODUCTION

Image saliency aims at finding the most attractive object or area in the image. One research field is Eye Fixation Prediction, which attempts to model where human visual fixation may be located. Another branch is salient object detection (SOD), which we concentrate on in this paper. By detecting and segmenting pixels in the image, SOD precisely simulates the human visual mechanism that focuses on the most informative part at first sight.SOD has been widely discussed and researched in the field of pattern recognition and computational vision since SOD can be applied to various applications and cooperate with other tasks, such as object detection and recognition [3], and [4], object tracking [5], and [6], image captioning [7], video compression [8], video abstraction [9], user interface optimization [10]. Zhang *et al*. [3] simulated visual attention mechanism as a prior to help weakly supervised object detection find objects in scenes. In object tracking, Lee *et al*. [6] separated the background and object in the bounding box according to salient region prediction. And the extracted salient object region is further adopted as color and shape models to estimate the bounding box in the next frame. Hadizadeh *et al*. [8] proposed a saliency-aware video compression to keep viewers' attention in regions of interest (ROI) by reducing coding artifacts in non-ROI areas. It compresses less for the ROI region and more for the non-ROI region for better overall visual quality. Ji *et al*. [9] adopted SOD to extract ROI so that less informative frames are discarded in the video abstraction. Gupta *et al*. [10] stated that in the user interface (UI) design, the iterative process of feedback and

DOI:10.6814/NCCU202101682

improvement is time-consuming, so they proposed a feedback tool based on the saliency of different elements of a UI.



(a) GT(scribble) (b) GT(Bbx) (c) GT(per-pixel)

Figure 1.1: An example of different weak annotations. (a) Scribble annotations. (b) Bounding box annotation. (c) Pixel-wise annotation. Figure is taken from Fig. 1 of [11].

In addition, other relevant SOD research includes weakly supervised SOD, co-salient object detection (CoSOD), and RGB-D SOD. Weakly supervised SOD does not need time-consuming pixel-wise labels to train a model. With weak annotations, we could obtain and adopt different extra information, including bounding boxes [12], image-level labels (class category) [13], and scribble labels [11], as illustrated in Fig. 1.1. CoSOD aims to find the common salient object inside an image or inter images, shown in Fig. 1.2. Researchers consider depth information a significant complementary for RGB images. Hence RGB-D is another active research topic in the SOD field.

There has been much work studying SOD for terrestrial scenes. In contrast, little work explicitly has been done for underwater scenes, and most of the existing popular datasets only include seldom numbers of underwater scenes. With object-level distinguishing, it is helpful to apply SOD to underwater exploration missions [15] and benefit other underwater tasks. As shown in Fig. 1.3, before masked with saliency maps, the image captioning model misstate the image because of the complex scene and blurry effect under these. Besides, Islam *et al*. [16] trained a simultaneous enhancement and super-resolution model guided by predicted saliency maps. In [17] and [18], saliency detection was also intro-

DOI:10.6814/NCCU202101682

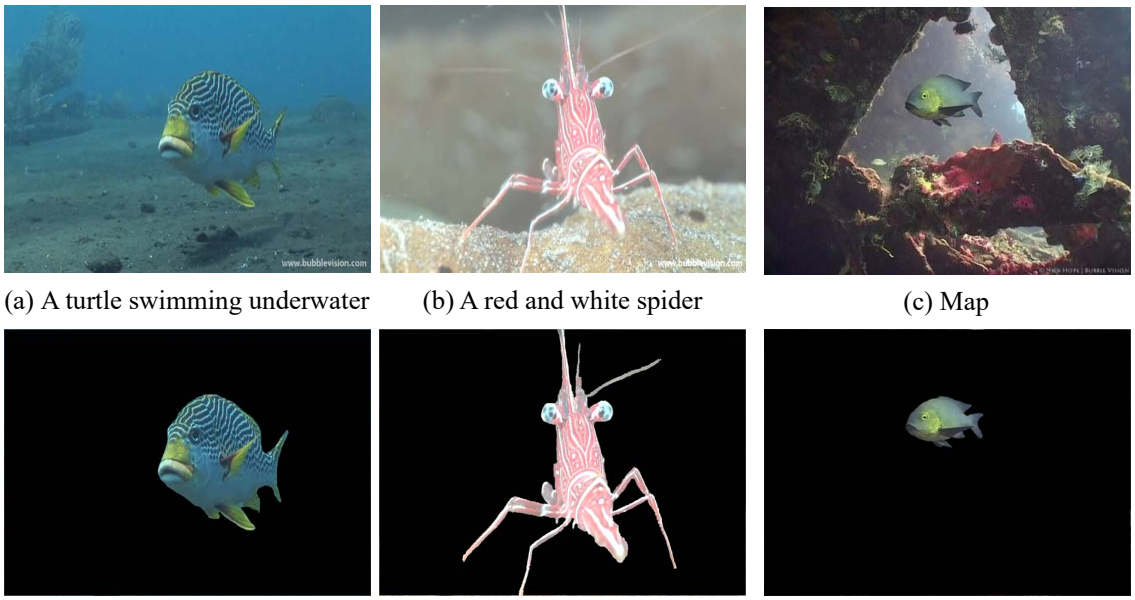
Figure 1.2: An example of CoSOD. The first row lists the input images, and the second row has their corresponding ground truth. (a) Vanilla SOD. (b) Intra-image CoSOD. (c) Inter-images CoSOD. Figure is taken from Fig. 1 of [14].

duced to assist feature extraction and visual tracking under the water. Therefore, it is crucial to develop a useful SOD solution explicitly for underwater scenes.

With the development of the deep learning method, researchers have made significant progress in the SOD task. Zhao *et al*. [20] considered global context and local context by a model consisting of fully connected (FC) layers and convolutional neural network (CNN). Li *et al*. [21] also used three branches of CNN layers and FC layers to extract multi-scale features and generate saliency maps. Moreover, the emergence of a fully convolutional network[22, 23] improved the saliency map from coarse result to pixel-wise prediction. Following those efforts, learning-based saliency detection models have been developed comprehensively.

For conventional SOD methods, there are various hand-crafted cues used to estimate saliency for image pixels. SOD methods using local contrast as a cue [24] mostly focus on object boundaries but miss interior parts of the objects. Cheng *et al*. introduced global contrast [25] to address the issue by considering spatial correlations across local regions the global information. Cues like background information [26] and foreground object [27]

DOI:10.6814/NCCU202101682



(a) A turtle swimming underwater (b) A red and white spider (c) Map

(d) A fish swimming in water (e) A red shrimp with white spots (f) A white fish in the dark

Figure 1.3: Example of image captioning [19] on underwater images. Top row: images and their captions. Second row: images masked by saliency maps and their captions.

are also helpful for non-deep learning SOD. Recently, a similar work is RGB-D SOD, where RGB data are fused with depth cues to distinguish saliency objects. Peng *et al*. [28] concatenated both RGB images and their depth maps as the model input. Han *et al*. [29] first separately extracted the RGB and depth features and then fused them into a fully connected layer to get a predicted map.

This thesis focuses on underwater SOD tasks. Although SOD for terrestrial scenes is very common, underwater SOD has less attention. Hence, there is few public benchmark dataset for underwater SOD. We have collected and annotated a dataset to evaluate our work. This thesis presents two practical approaches to boost the performance of underwater SOD. First, we present a blurriness-guided SOD model by adding blurriness features and RGB features since underwater scene depth can be estimated using its image blurriness [30]. Compared to the RGB-D SOD, our approach can be considered using pseudo depth estimated based on blurriness. Thus, our design is more applicable to various models. Moreover, we can use an attention-based fusion to combine features from both RGB images and their blurriness maps.

DOI:10.6814/NCCU202101682

Data augmentation is a practical technique to boost model performance. There have been several data augmentation methods, such as random flipping, rotating, cropping, color jittering, and data augmentation methods specific to SOD [1, 31]. The second approach is our proposed data augmentation method, called FocusAugment, specifically for underwater SOD. FocusAugment adjusts the image gain based on the blurriness map since salient objects in an image are usually less blurred. We can modify images by highlighting less blurred regions or enlarging the difference of pixels based on the blurriness maps.

The contributions of my work are three-fold: :

1. An underwater SOD dataset is constructed to train and evaluate our method.
2. A blurriness-guided method is proposed to estimate underwater SOD, exploiting an inherent blur characteristic of underwater images.
3. A task-specific data augmentation method is introduced to boost model performance.

The rest of this thesis is organized as follows. Chapter 2 describes related works. Chapter 3 depicts the collected dataset for underwater SOD. Chapter 4 details the two proposed approaches. Experimental results are demonstrated and discussed in Chapter 5. At last, Chapter 6 concludes the thesis.

DOI:10.6814/NCCU202101682

2 RELATED WORKS

In this section, we will review some SOD and augmentation methods. In Sec. 1, some traditional non-deep-learning methods have already been mentioned, and we will focus on the deep-learning-based model here.

2.1 Deep SOD models

Convolution neural networks have recently been very successful in computer vision tasks. We first introduce various SOD methods based on CNN as follows.

Multi-stage learning: Integrating multi-level features [32, 33, 34, 35] have been widely used in salient object detection regions. For example, Hou *et al*. [32] proposed to fuse multi-level features with short connections as skip-layer structures. Wei *et al*. (F^3Net) [35] used cross-feature modules to fuse multi-level features of different stages, and their cascaded-decoder architecture refines multi-level features for image saliency. Qin *et al*. (BASNet) [36] designed a consecutive encoder-decoder structure model, where short connections in different corresponding stages connect encoder and decoder to predict and refine saliency maps. In addition, a hybrid loss is also proposed to supervise multi-scale outputs from different decoder blocks. Pang *et al*. (MINet) [2] developed a network focusing on the connection between encoder-decoder blocks and decoder units. The previous one interactively aggregated multi-level features from different encoder outputs and fed them to the decoder units after adding with the previous decoder-level. Each

DOI:10.6814/NCCU202101682

decoder unit also interactively learns features with different scales derived from the input. Zhao *et al*. (GateNet) [37] proposed a dual branch structure with multi-level gate unites to balance the contribution of each encoder block and suppress the irrelevant features feeding into decoder blocks. Gao *et al*. (CSNet) [38] proposed a generalized OctConv module to utilize both in-stage and cross-stages multi-scale features in a lightweight manner.

Attention mechanisms: With attention mechnisms [39, 40], CNN models can further emphasize feature maps by highlighting important regions and suppress less useful regions. Wu *et al*. [39] showed that partially discarding some low-level features wouldn't seriously influence the performance, but it will save much time on training. It also proposed a holistic attention module that can comprehensively get additional information. Li *et al*. (SKNet) [40] introduced a selective kernel unit that generates multiple branches from the input and selects the significant part by weights learned from each branch and a softmax operation. The gate units from GateNet mentioned earlier are also seen as a kind of attention mechanism.

Learning with other cues: Traditional hand-crafted cues would also be beneficial supplementary information to detection, such as using scene depth as prior information. These datasets are called RGB-D datasets. With fusing RGB texture information and depth cues, models can be modified to a depth-induced detection model. Liu *et al*. [41] uses self-mutual attention to fuse RGB and depth features. Jiang *et al*. [42] exploited conditional generative adversarial networks to handle cross-modality of RGB and depth for SOD. However, an obvious drawback of the RGB-D datasets is that we need to get salient ground truth maps and get depth ground truth maps simultaneously. It is impractical and difficult to collect data. Zhao *et al*. [43] proposed an edge guidance network taking into account multi-scale local salient edge to locate salient object regions and boundaries simultaneously. And, in their later work [38] (ITSD), an interactive two-stream decoder for saliency and contour respectively was proposed, and a correlation module fuses the

DOI:10.6814/NCCU202101682

Table 2.1: Data augmentation policy adopted by recent SOD models

SOD Methods	Random Horizontal Flipping	Random Cropping	Random Scaling	Random Rotating	Random Color Jittering	Random Vertical Flipping
KRN [47]	✓		✓	✓		
PFSNet [48]	✓	✓	✓			
U2Net [49]	✓	✓				
GCPANet [50]	✓	✓				
F3Net [35]	✓	✓	✓			
MINet [2]	✓			✓	✓	
ITSD [38]	✓	✓	✓			
LDF [51]	✓	✓	✓			
CSNet [52]	✓	✓				
GateNet [37]	✓			✓	✓	
CAGNet [53]	✓			✓		
BASNet [36]	✓	✓				
AFNet [54]	✓	✓				✓
PoolNet [55]	✓					

features from two cues. Wang *et al*. [44] proposed a salient edge detection module to better segment salient objects and refine object boundaries.

Underwater SOD: For underwater SOD, Feng *et al*. [45] adopted an improved spectral residual method and Fuzzy c-Means clustering method to segment underwater saliency maps. Chen *et al*. [46] proposed a biologically inspirited model by combining 2D features, i.e., the color and intensity, and 3D depth features extracted by the DCP-based method; however, it has been shown that the depth estimation from the DCP-based method is not reliable on underwater scenes [30].

2.2 Data augmentation for SOD

Data augmentation is an important technique that generates more training examples and reduces overfitting by increasing data diversity. Conventionally, we could manipulate images from the perspective of geometric and photometric transforms. For instance, random flipping, rotation, scaling, cropping, and color jittering, etc., are all commonly used in training deep-learning-based computer-vision models. The common data augmentation approaches used in SOD are listed in Tab. 2.1.

To further utilize the aforementioned methods, the Autoaugment [56] family, *e.g*., Fast Autoaugment [57], Randaugment [58], applied meta-learning concepts and presented a

DOI:10.6814/NCCU202101682

method that automatically searches for optimal combination of data augmentation techniques. The Neural augmentation [59] utilized a style transfer network to generate an augmented image with two randomly selected images. Similarly, Smart augment [60] merged two or more images randomly chosen from the same class and outputted an augmented sample. Frid-Adar *et al*. [61] generated synthetic medical images using Generative Adversarial Networks (GANs) to enlarge the data size and its diversity. Mariani, *et al*. [62] attempted to restore balance in imbalanced datasets by their proposed balancing GAN.

IDA [1] and Anda [31] are the recent task-specific data augmentation methods designed for SOD. They could increase the diversity using background replacement for salient objects. They first generated new background images by removing salient objects and inpainted the removed regions. Next, the kNN algorithm is adopted to select a similar background from different images. Last, the work [31] randomly chose the overlay position of salient objects in the new background, while [1] determined the object position by intra-image optimization. With these synthetic images, more diverse training data is introduced to the SOD model. However, inpainting methods tend to add noise and distortion to the images, and the context between object and background is undermined by different kinds of color cast veiled on themselves. Our work proposed a label-invariant data augmentation by making the focused objects more prominent to boost the model performance.

DOI:10.6814/NCCU202101682

3 PROPOSED DATASET

3.1 Dataset for underwater SOD

As known, data is the most important key for deep-learning-based methods to succeed. Although many SOD datasets are available for terrestrial scenes, there is not much data for underwater SOD. That is, a generally acknowledged dataset for underwater SOD is not presented. Therefore, we have collected and labeled a total of 1,111 underwater images with their ground-truth saliency maps, where 800, 100 and 211 image-saliency pairs are randomly chosen to be the training, validation and testing sets. Our data comes from the images or videos downloaded from [63], and [64] and National Geographic footage [65]. These images have a wide variety of contents, watercolors, visibility degrees, and scales of objects. We excluded those without obvious objects. All collected images are then labeled using [66], an annotation tool for image segmentation. Fig. 3.1 shows some samples of our dataset. The first row demonstrates images with different object sizes, categories, and watercolors, and the second row shows their annotated saliency maps.

3.2 Dataset analysis and comparison

First, we analyze and compare our underwater SOD dataset, named U-SOD, with five public terrestrial datasets as:

DOI:10.6814/NCCU202101682

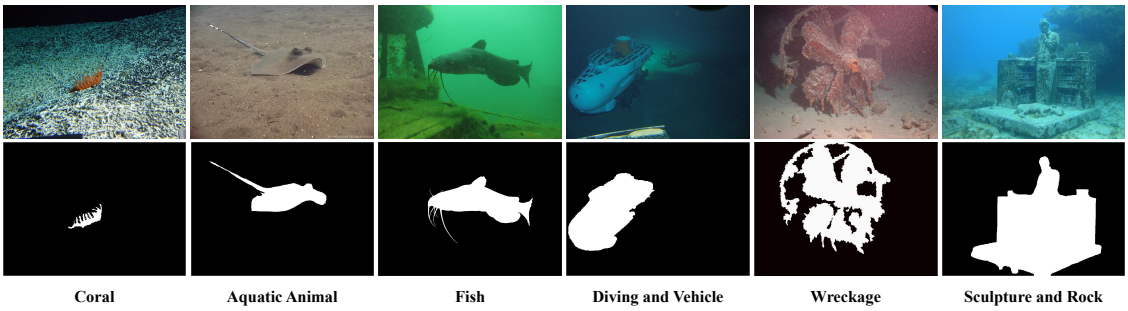


Figure 3.1: Sample of our underwater SOD dataset: The first row are original images, and the second one are annotated images.

- Extended Complex Scene Saliency Dataset (ECSSD) [67] that has 1000 images with semantically meaningful but structurally complex scenes;
- DUT-OMRON [68] that contains 5172 challenging images;
- PASCAL-S [69] that has 850 validation images used in Pascal visual object classes challenge;
- HKU-IS [21] consisting of 4447 images with low contrast or having image boundary overlapped;
- DUTS [70] containing 15572 images and served as a generic training set of SOD tasks;
- U-SOD consists of 1,111 underwater images with various contents, backgrounds, and watercolors.

We further analyzed our U-SOD dataset and classified the images according to their contents, such as coral, fish, diving and vehicles, aquatic animals, wreckage, and sculpture, and rock, shown in Fig. 3.2(a), where the fish-related content accounts for more than 45% of the entire dataset, containing various fish species, including sharks, stingrays, eels, anglerfish, and so on. Due to unstable lighting conditions and varying attenuation rates in the water, underwater images suffer from blurring, low contrast, and color distortions. To

DOI:10.6814/NCCU202101682

further analyze our dataset, we quantitatively compared the contrast and color distribution (Hue) of our U-SOD and the abovementioned terrestrial datasets.

Local contrast: We measured the local contrast of images based on the minimal and maximal luminance in a 3×3 image patch $\Omega(x)$ centered at the pixel x in the CIELAB color space. We then averaged contrast over all the local patches sliding across the whole image as $contrast = \frac{1}{N}\sum_x \frac{max_{y\in\Omega(x)}(L_y) - min_{y\in\Omega(x)}(L_y)}{max_{y\in\Omega(x)}(L_y) + min_{y\in\Omega(x)}(L_y)}$, where N is the image size. As shown in Fig. 3.2(c), our U-SOD dataset has the smallest value of contrast, indicating that it is more challenging for SOD since better contrast makes salient objects more prominent and easier to be detected.

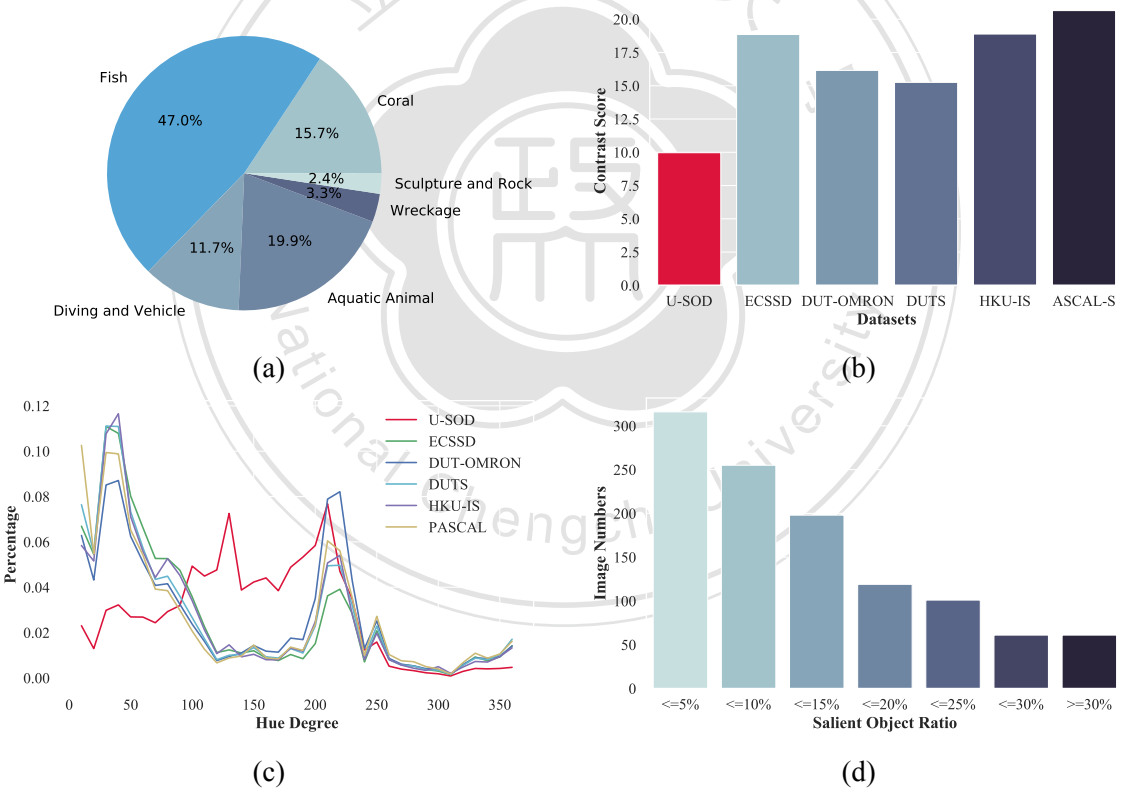


Figure 3.2: Statistics of the U-SOD dataset. (a) Percentage of content categories for U-SOD. Comparison of our U-SOD and the abovementioned terrestrial datasets in (b) Contrast and (c) Hue. (d) U-SOD Salient Object Ratio.

Color distribution: We compared the color distribution of different datasets in Fig. 3.2(a) to show that underwater images are distorted with a non-uniform color cast. We converted images into HSV color space and computed the Hue distribution with 36 bins. Hue ranges

DOI:10.6814/NCCU202101682

from 0° to 360°, and the hue for red is 0°. According to Fig. 3.2(a), our underwater SOD dataset has significantly fewer red components as the red light attenuates faster in the water.

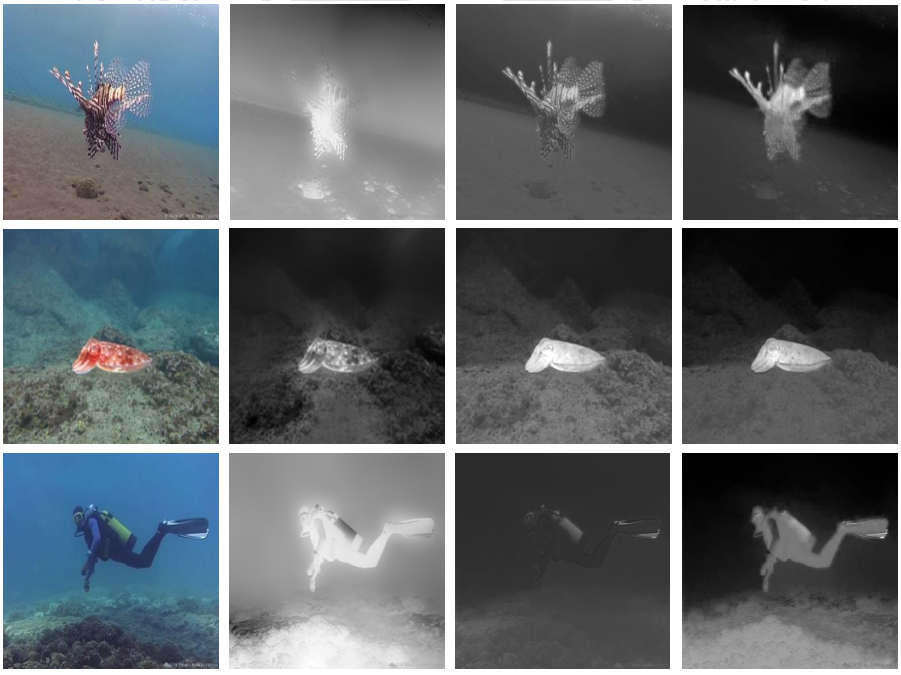
Object Ratio: We analyzed how large the salient objects take in each image, shown in Fig. 3.2(d). In the underwater scene, objects tend to have a similar color to the background. Thus, a large object should be easier to be identified than small objects, which could be considered as noise. Fig. 3.2(d) shows the U-SOD dataset has salient objects with various scales in sizes.

DOI:10.6814/NCCU202101682

4 PROPOSED METHOD

4.1 Blurriness-guided SOD

4.1.1 Blurriness map generation



(a) (b) (c) (d)

Figure 4.1: Comparison of different depth estimation method. (a) Input images; (b) Semi-inverse [71]; (c) Red Channel [72]; (d) Blurriness-based [30].

The images taken from underwater scenes usually suffer from blur effects. Naturally, salient objects in images should be in the foreground, thus less blurry for underwater images due to the fact that when light travels longer, the light scatters more, causing more

DOI:10.6814/NCCU202101682

blurry. Thus, we can take advantage of the blurriness of underwater images as a cue to boost the accuracy of SOD. In our work, we chose the work [30] to estimate image blurriness and generate a blurriness map for an underwater image.

There are several works that can do a similar thing. Galdran *et al*. [72] extended the DCP method to the Red Channel prior to underwater scenes based on the assumption that the red channel almost always has low intensities. One could use the red channel to estimate underwater scene depth, which can roughly help differentiate the foreground and background. Nevertheless, the assumption often fails in a haze-like lighting condition where farther scenes have more red. In [71], Xiao *et al*. estimated scene depth based on the semi-inverse of an image and integrated them into their SOD model. They adaptively inverse images according to different light transmission scenarios, either by medium transmission model or reversing the strength of images. However, images captured through the water medium can not fit the model because the disturbed light transmission fails.

We compare the performance of different approaches based on our U-Net baseline SOD model where a guided map is fused. The more details about our baseline SOD model is in Sec. 4.1.2. As can be seen in Tab. 4.1, the guided map can be the red Channel [72], Semi-inverse [71], and the blurriness map. We can see in in Tab. 4.1 that blurriness estimation is more advantageous than the other two approaches.

Table 4.1: Ablation study on different cues for our baseline model with CCAF described in Sec. 4.1.2.

Method	MaxF↑	MeanF↑	WFm↑	Emeasure↑	Smeasure↑	MAE↓
No cues	0.875	0.828	0.828	0.928	0.843	.0335
Red Channel [72]	0.886	0.843	0.844	0.935	0.850	.0300
Semi-inverse [71]	**0.890**	0.838	0.829	0.919	0.831	.0315
Blurriness [30]	0.888	**0.848**	**0.848**	**0.941**	**0.852**	**.0295**

DOI:10.6814/NCCU202101682

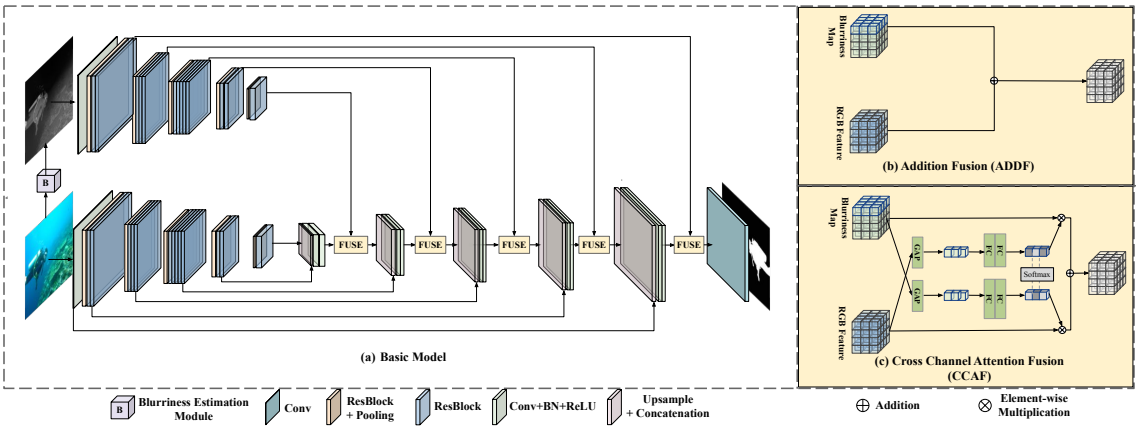


Figure 4.2: An illustration of the proposed SOD model. (a) Proposed SOD model integrate with blurriness. The FUSE in (a) represents either: (b) ADDition Fusion (ADDF) or (c) Cross Channel Attention Fusion (CCAF).

4.1.2 Blurriness-guided SOD model

As can be seen in Fig. 4.2(a), our baseline model has two data streams, RGB and blurriness, which stand for the underwater image input and its self-derived blurriness map input. We use a U-Net model with a ResNet-50 backbone to process the RGB stream. Likewise, we use the same backbone structure to extract features from the blurriness stream. To address cross-modality between RGB and blurriness, we use two ways to fuse the two streams: one with addition and the other with the proposed Cross Channel Attention Fusion (CCAF). For the addition fusion, we add blurriness features extracted from different layers to those corresponding layers (with the same resolutions) in the decoder of the RGB stream shown in Fig. 4.2(b). To better address cross-modality between RGB and blurriness, we employ a modified SK block from [40] to integrate them in this thesis, called CCAF, where RGB features(x^{rgb}) and blurriness features (x^b) become $1 \times 1 \times C$ after global average pooling, denoted as $s_c^{rgb}, s_c^b = f_{gap}(x^{rgb}, x^b)$, and they are fed into two consecutive fully connected layers as $a_C^{rgb}, a_C^b = f_{fcs}(s_c^{rgb}, s_c^b)$ respectively. At last, a softmax function is applied to the output features of those two streams as:

DOI:10.6814/NCCU202101682

$$A_c^{rgb} = \frac{e^{a_C^{rgb}}}{e^{a_C^{rgb}} + e^{a_C^b}}, A_c^b = \frac{e^{a_C^b}}{e^{a_C^{rgb}} + e^{a_C^b}},$$

where $A_c^{rgb} + A_c^b = [1, 1, ..., 1] \in R^C$ denote the attention weights for two streams. To this end, final output O is obtained by multiplying the attention weights from one stream by the features from the other stream as:

$$O = x^{rgb} \cdot A_c^b + x^b \cdot A_c^{rgb},$$

where $O = [O_1, O_2, ..., O_C] \in R^{H \times W \times C}$. The CCAF is shown in Fig. 4.2(c).

We adopt ResNet-50 pre-trained on the ImageNet dataset for the RGB stream but no pre-trained weights for the blurriness stream for our baseline U-Net structure. The loss function used is the multi-scale hybrid loss $\mathcal{L}$ originally proposed in [36], described, to demonstrate the effectiveness and accuracy gain of the fused blurriness cue. In Chapter 5, in addition to our baseline architecture, we integrate the proposed approach into the state-of-the-art U-net-like SOD model *i.e.*, MINet [2] and BASNet [36] to demonstrate our method's effectiveness.

4.2 Blurriness-guided augmentation for underwater SOD

In the previous section, we describe a blurriness-guided underwater SOD approach. However, currently, it can only work for U-Net-like architecture. In this section, we introduce a data augmentation approach designed for underwater SOD. Data augmentation (DA) is a useful technique to enlarge the training dataset and improve the performance of deep learning models. Our DA approach exploits image blurriness [30], called FocusAugment, specifically for underwater SOD. FocusAugment adjusts the image intensity based on the blurriness map since salient objects in an image are usually less blurred. The gen-

DOI:10.6814/NCCU202101682

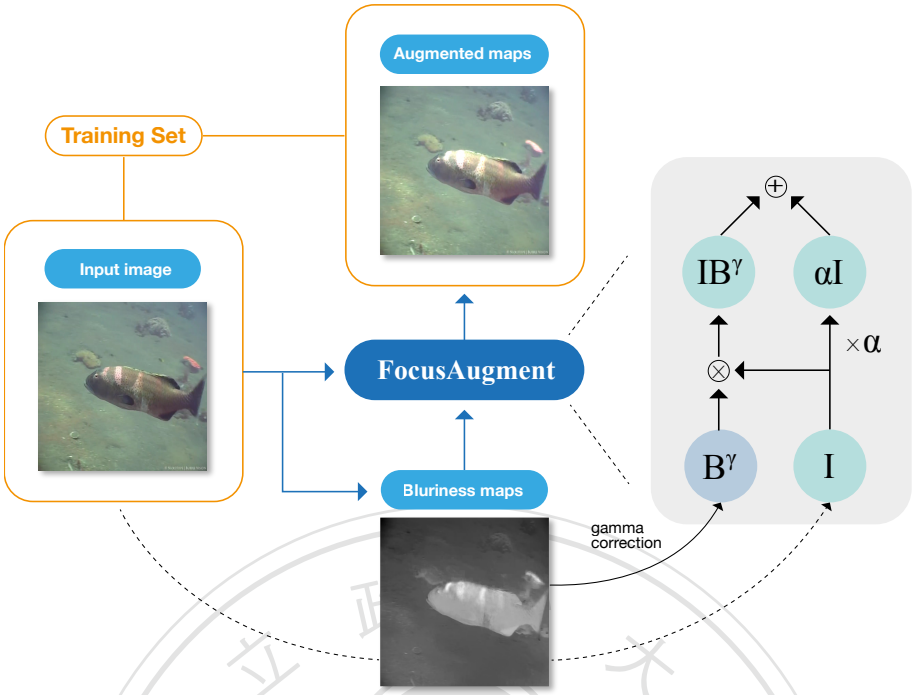


Figure 4.3: The flowchart of FocusAugment.

erated augmented images are modified by highlighting less blurred regions or enlarging the difference of pixels based on the blurriness maps. The intuition is that salient objects in a scene are generally in focus; thus, we would like to change the prominence of objects based on the focus to diversify data distribution. We first generate the blurriness map based on [30], where pixels in the foreground often have larger values while those in the background have small values. Next, we apply gamma correction to adjust the blurriness map and multiply it by the original image. Here, we use α to make the augmented image brighter or darker than the original. The augmented image I_{da} can be derived as

$$I_{da} = I \times (\alpha + B^{\gamma}), \tag{4.1}$$

where I is the original image, and B is the blurriness map. Here, the power for B, denoted γ, controls the intensity gain or decrease. Based on our experiment, we set $\gamma = 2$ and $\alpha = 1$, meaning augmented images are brighter than the original versions. Sec. 5.4 will

DOI:10.6814/NCCU202101682

give more detailed results. Hence, the augmented dataset S_{aug} for training is described as

$$S_{aug} = \{S \cup S_{da}\}, S_{da} = \{I \times (\alpha + B^{\gamma}),\ \forall I \in S\}, \tag{4.2}$$

where S represents the original dataset.

DOI:10.6814/NCCU202101682

5 EXPERIMENTS

5.1 Implementation and Experimental Setup

We evaluate each compared method on the U-SOD dataset, where all images are resized to 320×320. We implemented our approaches in Pytorch and ran all the experiments on a desktop with Intel Core i7-9700 CPU (3.00GHz), 32GB RAM, and an NVIDIA RTX 2080 Ti GPU. All experiments were conducted based on the testing set, except for fusion architecture and parameter selection on FocusAugment.

5.2 Evaluation Metrics

We have adopted seven metrics to evaluate and compare model performance, which are listed below.

5.2.1 Precision-Recall (PR) curve

With a given threshold, we can binarize predicted salient maps and compute Precision and Recall as:

$$Precision = \frac{TP}{(TP + FP)}$$

,

$$Recall = \frac{TP}{(TP + FN)}$$

DOI:10.6814/NCCU202101682

, where TP, FP, and FN stand for true-positive, false-positive, and false-negative, respectively. We can plot a PR curve by a sequence of Precision and Recall pairs from different thresholds.

5.2.2 F-measure family

F-measure [73], F_β, considering both Precision and Recall, is formulated as:

$$F_\beta = \frac{(1+\beta^2)Precision \times Recall}{\beta^2 Precision + Recall}$$

, where β^2 is set to 0.3 to emphasize precision. $MaximalF_\beta$, denoted as F_{max} reported according to the PR curve. And the F_{avg} calculated by the threshold that twice the mean value of the predicted salient map can reflect whether objects are uniformly highlighted in salient maps. We also compute the weighted F-measure [74], F_w formulating weighted Precision, which is a measure of exactness, and weighted Recall, a completeness measure.

5.2.3 Mean absolute error (MAE)

MAE [75] calculates the mean of the absolute difference between the predicted saliency map and its ground truth.

5.2.4 S-measure

S-measure [76], S_m is the linear combination of object-aware and region-aware structure similarities, denoted as S_o and S_r, between the prediction and the ground truth. The previous term exploits the relationship between the foreground and background, and the latter is a summation of patch-wise structural similarity. S_m can be described as: $S_m = \alpha \cdot S_o + (1 - \alpha) \cdot S_r$, where α usually is set to 0.5.

DOI:10.6814/NCCU202101682

5.2.5 E-measure

Enhanced-alignment Measure [77], E-measure simultaneously considers local pixel-level and image-level error to evaluate the similarity between the prediction and the ground truth.

5.3 Experimental results

5.3.1 Blurriness-guided underwater SOD

First, we evaluate the performance of our blurriness-guided SOD approach on the testing set. As mentioned before, we can only apply this approach to U-Net-like SOD architectures. Thus, we choose a U-Net baseline model, BASNet [36], and MINet [2] with two fusion settings, ADDF and CCAF, to compare their performance in our experiment. We use ResNet-50 as the backbone for the RGB and blurriness streams and adopt the hybrid loss in BASNet [36] for our baseline model. As seen in the Tab. 5.1, blurriness-guided approach with AADF merely improves the SOD accuracy across three models. MINet [2] which has a more complex structure even gets worse scores. It indicates that a meticulous design *i.e.*, CCAF is needed to help the model to learn. As for CCAF on Baseline, BASNet [36] and MINet [2], we could find that using the proposed blurriness-guided SOD with CCAF for fusing a blurriness cue achieves better accuracy, shown in Tab. 5.1 and Figs. 5.1(b, d). Furthermore, in addition to blurriness-guided models with CCAF, adopting the proposed FocusAugment can have a significant improvement.

5.3.2 FocusAugment

To examine the effectiveness of our FocusAugment, we test it on an U-Net baseline SOD network (the same one mentioned in Tab. 5.1) and five state-of-the-art SOD models,

DOI:10.6814/NCCU202101682

Table 5.1: Quantitative comparison of different fusion methods on state-of-the-art SOD models.

Model	MaxF↑	MeanF↑	WFm↑	Emeasure↑	Smeasure↑	MAE↓
Baseline	0.875	0.828	0.828	0.928	0.843	.0335
Baseline + ADDF	0.874	0.824	0.826	0.930	0.841	.0327
Baseline + CCAF	0.882	0.838	0.839	0.931	0.842	.0314
Baseline + CCAF + FocusAugment	**0.895**	**0.851**	**0.858**	**0.949**	**0.858**	**.0262**
BASNet	0.885	0.835	0.836	0.927	0.844	.0327
BASNet + ADDF	0.889	0.844	0.838	0.932	0.843	.0308
BASNet + CCAF	0.890	0.839	0.839	0.932	0.845	.0288
BASNet + CCAF + FocusAugment	**0.898**	**0.851**	**0.856**	**0.940**	**0.857**	**.0275**
MINet	0.894	0.875	0.872	0.963	0.878	.0249
MINet + ADDF	0.890	0.872	0.867	0.961	0.873	.0256
MINet + CCAF	0.896	0.877	0.875	0.966	0.881	.0240
MINet + CCAF + FocusAugment	**0.897**	**0.883**	**0.879**	**0.966**	**0.881**	**.0235**

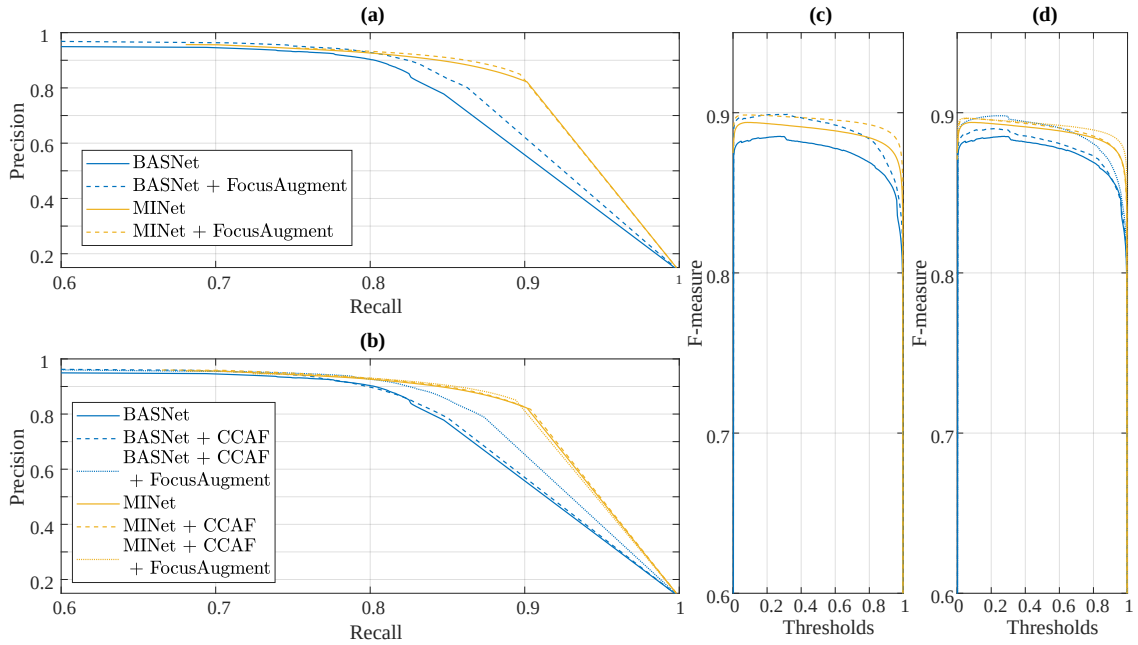


Figure 5.1: Illustration of Precision-Recall curves and F-measure curves on BASNet [36], and MINet [2], with and without proposed methods. (a, c) PR/ F-measure curves of SOD methods with and without FocusAugment; (b, d) PR/ F-measure curves of SOD methods with and without CCAF. Note that (a, c) and (b, d) share the same legends.

including BASNet [36], GateNet [37], MINet [2], F3Net [35], and ITSD [38].

We also use IDA [1] to train MINet [2], the best performer for the underwater SOD, but it is not helpful at all since the inpainting process in IDA [1] does not work for low-contrast, color-distorted, noise-prone underwater scenes. As shown in Fig. 5.2, the inpainted fish and rock seem distorted and not nature-looking.

After discussing the performance of CCAF and FocusAugment separately, we com-

DOI:10.6814/NCCU202101682

(a) (b)

Figure 5.2: Illustration of the inpainting process in IDA [1]. (a) Original input image. (b) Image after removing the salient object and the impainting process.

Table 5.2: Quantitative comparison of SOD models with or without FocusAugment. Note that these models have their own augmentation enabled with or without our FocusAugment.

Model	MaxF↑	MeanF↑	WFm↑	E measure↑	S measure↑	MAE↓
Baseline	0.875	0.828	0.828	0.928	0.843	.0335
Baseline w/ FocusAugment	**0.892**	**0.853**	**0.860**	**0.952**	**0.861**	**.0267**
BASNet [36]	0.885	0.835	0.836	0.927	0.844	.0327
BASNet w/ FocusAugment	**0.899**	**0.852**	**0.859**	**0.942**	**0.854**	**.0282**
GateNet [37]	0.908	0.892	0.879	0.954	0.873	.0241
GateNet w/ FocusAugment	**0.912**	**0.897**	**0.886**	**0.960**	**0.876**	**.0226**
ITSD [38]	0.904	0.824	0.843	0.941	0.894	.0281
ITSD w/ FocusAugment	**0.907**	**0.840**	**0.852**	**0.948**	**0.896**	**.0262**
F3Net [35]	0.888	0.852	0.856	0.949	**0.858**	.0282
F3Net w/ FocusAugment	**0.889**	**0.861**	**0.860**	**0.950**	0.856	**.0265**
MINet [2]	0.894	0.875	0.872	0.963	0.878	.0249
MINet w/ IDA [1]	0.895	0.879	0.874	0.962	0.879	.0244
MINet w/ FocusAugment	**0.899**	**0.886**	**0.882**	**0.967**	**0.883**	**.0232**

pare the combination of two methods in Tab. 5.1 and only FocusAugment in Tab. 5.2, which shows that combining blurriness-guided SOD and FocusAugment can make detection more accurate. However, since fusing blurriness cues into the RGB stream requires the model to be U-Net-like, it is not that flexible. FocusAugment, by contrast, is more adaptable to working with all the SOD models and achieving better performance.

DOI:10.6814/NCCU202101682

5.4 Ablation sutdy

5.4.1 Blurriness-guided underwater SOD

As mentioned in Sec. 2.1, we aim to integrate blurriness features into deep networks with the two-streams manner. Hence, we validate the baseline network with CCAF fusing RGB and blurriness features in either encoder or decoder side of the RGB stream on the validation set. As a result, shown in the Tab. 5.3, fusion conducted on the decoder side (B-Dec) has a better performance over on the encoder side (B-Enc) (row 3 and 2). We observe that fusing blurriness features closer to the final prediction could be more helpful since the blurriness is considered homogeneous to saliency and more similar to saliency maps. To prove that the performance gain is not merely from the increase of the model size, we conduct an ablation on replacing the blurriness map with the original RGB image (RGB-Enc and RGB-Dec). As can be seen in Row 4 and 5 of Tab. 5.3, fusing the blurriness cue works better.

Table 5.3: Validation accuracy for different configurations on CCAF. B, Enc, Dec denote blurriness map, fusion at encoder, and fusion at the decoder side.

Model	MaxF↑	MeanF↑	WFm↑	Emeasure↑	Smeasure↑	MAE↓
Baseline	0.850	0.805	0.812	0.938	0.828	.0309
Baseline + CCAF B-Enc	0.869	0.831	0.834	0.947	0.838	.0280
Baseline + CCAF B-Dec	**0.877**	**0.838**	**0.847**	**0.948**	**0.846**	**.0248**
Baseline + CCAF RGB-Enc	0.865	0.819	0.823	0.935	0.835	.0275
Baseline + CCAF RGB-Dec	0.863	0.822	0.827	0.939	0.836	.0275

5.4.2 FocusAugment

Here, we analyze FocusAugment with other augmentation methods to validate its effectiveness. Moreover, we also show how its hyperparameters listed in Equation 4.1 are determined. All these experiments are conducted using the state-of-the-art SOD model, MINet [2] and on the validation set.

DOI:10.6814/NCCU202101682

Effectiveness of the FocusAugment: To demonstrate the effectiveness of our FocusAugment, we compare it with six commonly used augmentation methods, including color jittering, horizontal flipping, vertical flipping, cropping, rotation, and random scaling. We employed each of the above techniques individually on MINet. As we can see from Tab. 5.4, all the methods can increase the model accuracy except for IDA [1]. FocusAugment reaches the top three scores in most metrics, indicating that our FocusAugment is beneficial to SOD tasks.

Configuration of FocusAugment: To determine a better set of hyperparameters for Equation 4.1, we use a grid search approach on the validation set based on MINet ($0 \leq \alpha \leq 1$ and $0 \leq \gamma leq 2.5$). In Tab. 5.5, the first row (α and γ equal to zero) shows the result of the original MINet without FocusAugment. The third row ($\alpha = 0$ and $\gamma = 1$) shows the result of MINet trained with the product of the original image and its blurriness map, which works slightly better. If we keep α the same and tune γ, there is no further improvement. To this end, we moderately select $\alpha = 1$ and $\gamma = 2$ to balance fidelity ($\alpha = 1$) and diversity ($\gamma = 2$). The last two rows of Tab. 5.5, we test if further potential synergy exists, we add the best two options, $\alpha = 0$ & $\gamma = 1$ and $\alpha = 1$ & $\gamma = 2$, to testing. As can be seen, it does increase MaxF a little, but overall there is no significant gain for other metrics. Thus, we only choose one set of hyperparameters for our data augmentation.

The impact of FocusAugment: As mentioned above, we have compared six commonly used augmentation methods to validate the effectiveness of FocusAugment. We further compare our FocusAugment with other Photometric transformations, such as contrast and brightness adjustment. Fig. 5.3 shows the difference between the original image and the data-augmented version of it using heatmaps, where larger values are present in red while small values are bluish. As can be seen, our FocusAugment emphasizes the salient region, the fish, more than global brightness or contrast adjustment do. Tab. 5.6

DOI:10.6814/NCCU202101682

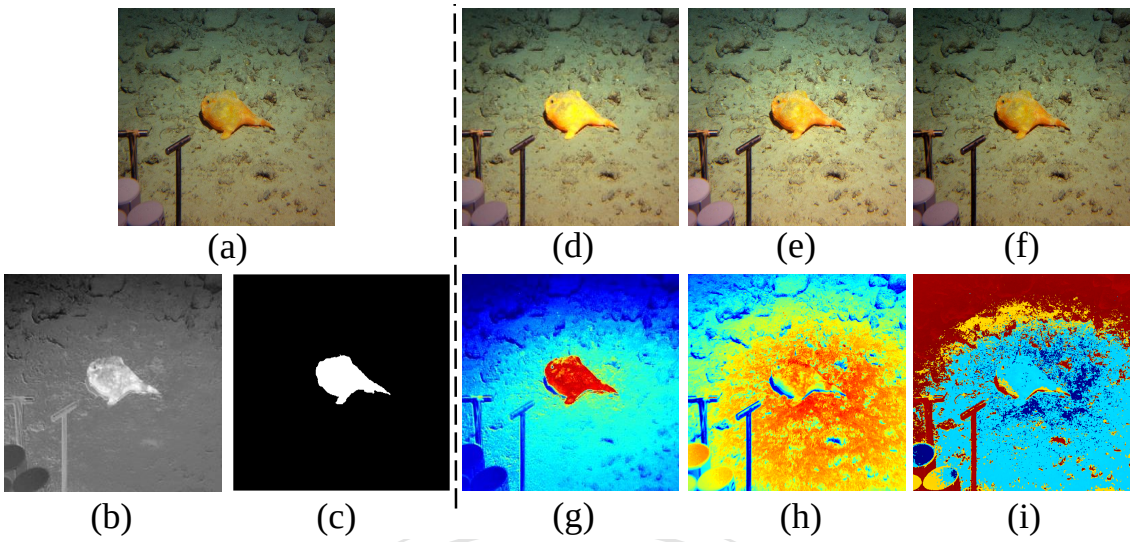


Figure 5.3: Visualization of different augmentation methods. (a) input image; (b) Blurriness map; (c) Saliency map; (d) Augmented by FocusAugment; (e) Augmented by brightness; (f) Augmented by contrast; (g-i) differences of d-f and a.

Table 5.4: Quantitative comparison of different augmentation method based on MINet [2]

MINet	MaxF↑	MeanF↑	WFm↑	Emeasure↑	Smeasure↑	MAE↓
None	0.889	0.877	0.867	0.959	0.881	.0274
+ H flipping	**0.893**	0.878	0.872	0.961	**0.883**	.0257
+ V flipping	0.892	**0.880**	**0.872**	**0.961**	0.882	.0255
+ Cropping	0.892	0.864	0.860	0.956	0.862	**.0250**
+ Rotation	0.890	0.870	0.866	0.961	0.876	.0258
+ Random scaling	0.888	0.874	0.862	0.955	0.875	.0287
+ Color Jittering	0.886	0.875	0.865	0.958	0.877	.0274
+ IDA [1]	0.887	0.876	0.862	0.951	0.873	.0290
+ FocusAugment	0.889	0.878	0.869	0.960	0.881	.0270

shows the quantitative results, which indicates that the proposed FocusAugment alone works slightly better than indiscriminate transformations.

DOI:10.6814/NCCU202101682

Table 5.5: Validation accuracy for different configuration on FocusAugment on MINet [2]. The top one performance is highlighted in red and the second one is in blue.

Method	α	γ	MaxF↑	MeanF↑	WFm↑	Emeasure↑	Smeasure↑	MAE↓
MINet	0	0	0.886	0.871	0.870	0.968	0.871	.0211
	0	0.5	0.885	0.871	0.870	0.970	0.875	.0212
	0	1	0.888	0.876	0.875	0.971	0.880	.0196
	0	1.5	0.887	0.874	0.870	0.967	0.874	.0207
	0	2	0.882	0.870	0.867	0.966	0.876	.0225
	0.5	0.5	0.885	0.871	0.871	0.970	0.876	.0205
	0.5	1	0.888	0.875	0.871	0.965	0.881	.0202
	0.5	1.5	0.889	0.877	0.875	0.970	0.874	.0211
	0.5	2	0.891	0.875	0.875	0.969	0.873	.0210
	1	0.5	0.884	0.871	0.871	0.969	0.871	.0204
	1	1	0.882	0.868	0.866	0.967	0.871	.0225
	1	1.5	0.886	0.872	0.872	0.969	0.875	.0207
	1	2	0.887	0.876	0.877	0.975	0.879	.0196
	1	2.5	0.887	0.874	0.876	0.973	0.879	.0207
	0	1	0.896	0.882	0.879	0.964	0.884	.0230
	1	2						

Table 5.6: Quantitative comparison of Photometric-based augmentation methods on MINet [2].

MINet	MaxF↑	MeanF↑	WFm↑	Emeasure↑	Smeasure↑	MAE↓
+ Brightness	0.888	0.876	0.867	0.959	**0.882**	0.0270
+ Contrast	0.888	0.876	0.865	0.957	0.880	0.0286
+ FocusAugment	**0.889**	**0.878**	**0.869**	**0.960**	0.881	**0.0270**

DOI:10.6814/NCCU202101682

6 CONCLUSION

In this work, we have proposed a blurriness-guided underwater SOD model and a data augmentation method, FocusAugment, for underwater SOD. We constructed an underwater SOD dataset including a wide variety of underwater scenes to verify the performance. Experimental results show that fusing a blurriness cue into salient object detection can increase detection accuracy. Besides, applying the proposed FocusAugment to training SOD models can further boost performance. Combining blurriness cues and FocusAugment achieves the best results. Comparing the two proposed approaches, unlike fusing blurriness cues into SOD requires the model to be U-Net-like, FocusAugment, by contrast, can work with all the SOD models and achieve better performance.

DOI:10.6814/NCCU202101682

Reference

[1] D. V. Ruiz, B. A. Krinski, and E. Todt, "Ida: Improved data augmentation applied to salient object detection," in *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. IEEE, 2020.

[2] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[3] D. Zhang, D. Meng, L. Zhao, and J. Han, "Bridging saliency detection to weakly supervised object detection based on self-paced curriculum learning," *arXiv preprint arXiv:1703.01290*, 2017.

[4] Z. Ren, S. Gao, L.-T. Chia, and I. W.-H. Tsang, "Region-based saliency detection and its application in object recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, 2013.

[5] S. P. Bharati, S. Nandi, Y. Wu, Y. Sui, and G. Wang, "Fast and robust object tracking with adaptive detection," in *2016 IEEE 28th international conference on tools with artificial intelligence (ICTAI)*. IEEE, 2016.

[6] H. Lee and D. Kim, "Salient region-based online object tracking," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018.

DOI:10.6814/NCCU202101682

[7] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Paying more attention to saliency: Image captioning with saliency and context attention," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2018.

[8] H. Hadizadeh and I. V. Bajić, "Saliency-aware video compression," *IEEE Transactions on Image Processing*, vol. 23, no. 1, pp. 19–33, 2013.

[9] Q.-G. Ji, Z.-D. Fang, Z.-H. Xie, and Z.-M. Lu, "Video abstraction based on the visual attention model and online clustering," *Signal Processing: Image Communication*, vol. 28, no. 3, pp. 241–253, 2013.

[10] Y. Ban and K. Lee, "Re-enrichment learning: Metadata saliency for the evolutive personalization of a recommender system," *Applied Sciences*, vol. 11, no. 4, p. 1733, 2021.

[11] J. Zhang, X. Yu, A. Li, P. Song, B. Liu, and Y. Dai, "Weakly-supervised salient object detection via scribble annotations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 546–12 555.

[12] P. Siva, C. Russell, T. Xiang, and L. Agapito, "Looking beyond the image: Unsupervised learning for object saliency and detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3238–3245.

[13] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 136–145.

[14] D.-P. Fan, T. Li, Z. Lin, G.-P. Ji, D. Zhang, M.-M. Cheng, H. Fu, and J. Shen, "Re-thinking co-salient object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

DOI:10.6814/NCCU202101682

[15] M. J. Islam, R. Wang, K. de Langis, and J. Sattar, "Svam: Saliency-guided visual attention modeling by autonomous underwater robots," *arXiv preprint arXiv:2011.06252*, 2020.

[16] M. J. Islam, P. Luo, and J. Sattar, "Simultaneous enhancement and super-resolution of underwater imagery for improved visual perception," *arXiv preprint arXiv:2002.01155*, 2020.

[17] L. Zhang, B. He, Y. Song, and T. Yan, "Underwater image feature extraction and matching based on visual saliency detection," in *OCEANS 2016-Shanghai*. IEEE, 2016.

[18] A. Maldonado-Ramírez and L. A. Torres-Méndez, "Robotic visual tracking of relevant cues in underwater environments with poor visibility conditions," *Journal of Sensors*, 2016.

[19] Microsoft azure cognitive services computer vision. Accessed on Nov. 2020. [Online]. Available: https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/

[20] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1265–1274.

[21] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015.

[22] X. Li, L. Zhao, L. Wei, M.-H. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang, "Deepsaliency: Multi-task deep neural network model for salient object detection," *IEEE transactions on image processing*, vol. 25, no. 8, pp. 3919–3930, 2016.

DOI:10.6814/NCCU202101682

[23] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency detection with recurrent fully convolutional networks," in *European conference on computer vision*. Springer, 2016, pp. 825–841.

[24] D. A. Klein and S. Frintrop, "Center-surround divergence of feature statistics for salient object detection," in *2011 International Conference on Computer Vision*. IEEE, 2011.

[25] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE transactions on pattern analysis and machine intelligence*, 2014.

[26] Z. Wang, D. Xiang, S. Hou, and F. Wu, "Background-driven salient object detection," *IEEE transactions on multimedia*, 2016.

[27] Ç. Aytekin, H. Possegger, T. Mauthner, S. Kiranyaz, H. Bischof, and M. Gabbouj, "Spatiotemporal saliency estimation by spectral foreground detection," *IEEE Transactions on Multimedia*, 2017.

[28] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "Rgbd salient object detection: a benchmark and algorithms," in *European conference on computer vision*. Springer, 2014, pp. 92–109.

[29] J. Han, H. Chen, N. Liu, C. Yan, and X. Li, "Cnns-based rgb-d saliency detection via cross-view transfer and multiview fusion," *IEEE transactions on cybernetics*, vol. 48, no. 11, pp. 3171–3183, 2017.

[30] Y.-T. Peng, X. Zhao, and P. C. Cosman, "Single underwater image enhancement using depth estimation based on blurriness," in *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 4952–4956.

DOI:10.6814/NCCU202101682

[31] D. V. Ruiz, B. A. Krinski, and E. Todt, "Anda: A novel data augmentation technique applied to salient object detection," in *2019 19th International Conference on Advanced Robotics (ICAR)*. IEEE, 2019.

[32] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. Torr, "Deeply supervised salient object detection with short connections," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3203–3212.

[33] G. Lee, Y.-W. Tai, and J. Kim, "Deep saliency with encoded low level distance map and high level features," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 660–668.

[34] N. Liu and J. Han, "Dhsnet: Deep hierarchical saliency network for salient object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 678–686.

[35] J. Wei, S. Wang, and Q. Huang, "F^3net: Fusion, feedback and focus for salient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[36] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: Boundary-aware salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[37] X. Zhao, Y. Pang, L. Zhang, H. Lu, and L. Zhang, "Suppress and balance: A simple gated network for salient object detection," in *European Conference on Computer Vision*. Springer, 2020, pp. 35–51.

[38] H. Zhou, X. Xie, J.-H. Lai, Z. Chen, and L. Yang, "Interactive two-stream decoder for accurate and fast saliency detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

DOI:10.6814/NCCU202101682

[39] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3907–3916.

[40] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 510–519.

[41] N. Liu, N. Zhang, and J. Han, "Learning selective self-mutual attention for rgb-d saliency detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 756–13 765.

[42] B. Jiang, Z. Zhou, X. Wang, J. Tang, and B. Luo, "Cmsalgan: Rgb-d salient object detection with cross-view generative adversarial networks," *IEEE Transactions on Multimedia*, vol. 23, pp. 1343–1353, 2020.

[43] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "Egnet: Edge guidance network for salient object detection," in *Proc. Int'l Conf. Computer Vision*, 2019.

[44] W. Wang, S. Zhao, J. Shen, S. C. Hoi, and A. Borji, "Salient object detection with pyramid attention and salient edges," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1448–1457.

[45] H. Feng, X. Yin, L. Xu, G. Lv, Q. Li, and L. Wang, "Underwater salient object detection jointly using improved spectral residual and fuzzy c-means," *Journal of Intelligent & Fuzzy Systems*, vol. 37, no. 1, pp. 329–339, 2019.

[46] Z. Chen, H. Gao, Z. Zhang, H. Zhou, X. Wang, and Y. Tian, "Underwater salient object detection by combining 2d and 3d visual features," *Neurocomputing*, vol. 391, pp. 249–259, 2020.

DOI:10.6814/NCCU202101682

[47] B. Xu, H. Liang, R. Liang, and P. Chen, "Locate globally, segment locally: A progressive architecture with knowledge review network for salient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[48] M. Ma, C. Xia, and J. Li, "Pyramidal feature shrinking for salient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[49] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-net: Going deeper with nested u-structure for salient object detection," *Pattern Recognition*, 2020.

[50] Z. Chen, Q. Xu, R. Cong, and Q. Huang, "Global context-aware progressive aggregation network for salient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[51] J. Wei, S. Wang, Z. Wu, C. Su, Q. Huang, and Q. Tian, "Label decoupling framework for salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[52] S.-H. Gao, Y.-Q. Tan, M.-M. Cheng, C. Lu, Y. Chen, and S. Yan, "Highly efficient salient object detection with 100k parameters," in *European Conference on Computer Vision*. Springer, 2020.

[53] S. Mohammadi, M. Noori, A. Bahri, S. G. Majelan, and M. Havaei, "Cagnet: Content-aware guidance for salient object detection," *Pattern Recognition*, 2020.

[54] M. Feng, H. Lu, and E. Ding, "Attentive feedback network for boundary-aware salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

DOI:10.6814/NCCU202101682

[55] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[56] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation policies from data," *arXiv preprint arXiv:1805.09501*, 2018.

[57] S. Lim, I. Kim, T. Kim, C. Kim, and S. Kim, "Fast autoaugment," *arXiv preprint arXiv:1905.00397*, 2019.

[58] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

[59] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," *arXiv preprint arXiv:1712.04621*, 2017.

[60] J. Lemley, S. Bazrafkan, and P. Corcoran, "Smart augmentation learning an optimal data augmentation strategy," *Ieee Access*, 2017.

[61] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification," *Neurocomputing*, 2018.

[62] G. Mariani, F. Scheidegger, R. Istrate, C. Bekas, and C. Malossi, "Bagan: Data augmentation with balancing gan," *arXiv preprint arXiv:1803.09655*, 2018.

[63] Hawaii udersea research laboratory. Accessed on Jun. 2019. [Online]. Available: http://www.soest.hawaii.edu/HURL/galleries.php

[64] Bubble vision. Accessed on Jun. 2019. [Online]. Available: https://www.bubblevision.com/

DOI:10.6814/NCCU202101682

[65] National geographic. Accessed on Jun. 2019. [Online]. Available: https://nationalgeographic.com/

[66] A. Bréhéret, "Pixel annotation tool," https://github.com/abreheret/PixelAnnotationTool, 2017.

[67] J. Shi, Q. Yan, L. Xu, and J. Jia, "Hierarchical image saliency detection on extended cssd," *IEEE transactions on pattern analysis and machine intelligence*, 2015.

[68] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013.

[69] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014.

[70] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[71] X. Xiao, Y. Zhou, and Y.-J. Gong, "Rgb-‘d’ saliency detection with pseudo depth," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2126–2139, 2018.

[72] A. Galdran, D. Pardo, A. Picón, and A. Alvarez-Gila, "Automatic red-channel underwater image restoration," *Journal of Visual Communication and Image Representation*, vol. 26, pp. 132–145, 2015.

[73] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2009.

DOI:10.6814/NCCU202101682

[74] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?" in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 248–255.

[75] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2012.

[76] D. Fan, M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proc. Int'l Conf. Computer Vision*, 2017.

[77] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," *arXiv preprint arXiv:1805.10421*, 2018.

DOI:10.6814/NCCU202101682