Robotics: Science and Systems 2022
New York City, NY, USA, June 27-July 1, 2022

SVAM: Saliency-guided Visual Attention Modeling by Autonomous Underwater Robots

Md Jahidul Islam*, Ruobing Wang† and Junaed Sattar‡
*RoboPI Group, Dept. of ECE, University of Florida, FL, USA
†‡IRVLab, Dept. of CS, University of Minnesota, Twin Cities, MN, USA
Email: *jahid@ece.ufl.edu, {†wang8063, ‡junaed}@umn.edu

***Abstract*—This paper presents a holistic approach to saliency-guided visual attention modeling (SVAM) for use by autonomous underwater robots. Our proposed model, named SVAM-Net, integrates deep visual features at various scales and semantics for effective salient object detection (SOD) in natural underwater images. The SVAM-Net architecture is configured in a unique way to jointly accommodate bottom-up and top-down learning within two separate branches of the network while sharing the same encoding layers. We design dedicated spatial attention modules (SAMs) along these learning pathways to exploit the coarse-level and fine-level semantic features for SOD at four stages of abstractions. The bottom-up branch performs a rough yet reasonably accurate saliency estimation at a fast rate, whereas the deeper top-down branch incorporates a residual refinement module (RRM) that provides fine-grained localization of the salient objects. Extensive performance evaluation of SVAM-Net on benchmark datasets clearly demonstrates its effectiveness for underwater SOD. We also validate its generalization performance by several ocean trials' data that include test images of diverse underwater scenes and waterbodies, and also images with unseen natural objects. Moreover, we analyze its computational feasibility for robotic deployments and demonstrate its utility in several important use cases of visual attention modeling.**

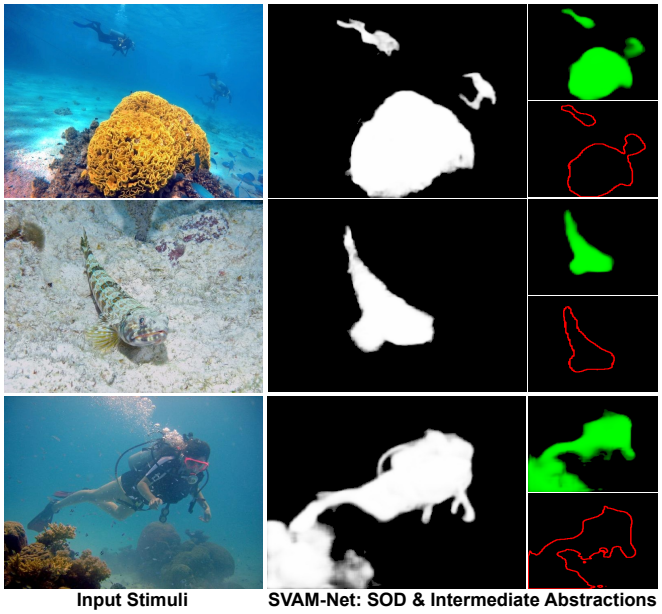


Fig. 1: The proposed SVAM-Net model identifies salient objects and interesting image regions to facilitate effective visual attention modeling by autonomous underwater robots. It also generates abstract saliency maps (shown in green intensity channel and red object contours) from an early bottom-up SAM which can be used for fast processing on single-board devices.

I. Introduction

Salient object detection (SOD) aims at modeling human visual attention behavior to highlight the most important and distinct objects in a scene. It is a well-studied problem in the domains of robotics and computer vision [8, 27, 39] for its usefulness in identifying regions of interest (RoI) in an image for fast and effective visual perception. The SOD capability is essential for visually-guided robots because they need to make critical navigational and operational decisions based on the relative *importance* of various objects in their field-of-view (FOV). The autonomous underwater vehicles (AUVs), in particular, rely heavily on visual saliency estimation for tasks such as exploration and surveying [13, 30, 26], ship-hull inspection [27], event detection [12], place recognition [43], target localization [67], and more.

In the pioneering work on SOD, Itti *et al.* [23] used local feature contrast in image regions to infer visual saliency. Numerous methods have been subsequently proposed [29, 10] that utilize local point-based features and also global contextual information as reference for saliency estimation. In recent years, the state-of-the-art (SOTA) approaches have used powerful deep visual models [62] to imitate human visual information processing through top-down or bottom-up computational pipelines. The bottom-up models learn to gradually infer high-level semantically rich features [61]; hence the shallow layers' structural knowledge drives their multi-scale saliency learning. Conversely, the the top-down approaches progressively integrate high-level semantic knowledge with low-level features for learning *coarse-to-fine* saliency estimation [39]. Moreover, the contemporary models have introduced techniques to learn boundary refinement [46, 58], pyramid feature attention [62], and contextual awareness [39], which significantly boost the SOD performance on benchmark datasets.

However, the applicability of such powerful learning-based SOD models in real-time underwater robotic vision has been rather limited. The underlying challenges and practicalities are twofold. First, the visual content of underwater imagery is uniquely diverse due to domain-specific object categories, background waterbody patterns, and a host of optical distortion artifacts [3, 22]; hence, the SOTA models trained

on terrestrial data are not transferable off-the-shelf. A lack of large-scale annotated underwater datasets aggravates the problem; the existing datasets and relevant methodologies are tied to specific applications such as coral reef classification and coverage estimation [6, 4], object detection [47], and foreground segmentation [68]. Consequently, these do not provide a comprehensive data representation for effective learning of underwater SOD. Secondly, learning a generalizable SOD function demands the extrapolation of multi-scale hierarchical features by high-capacity deep network models. This results in a heavy computational load and makes real-time inference impossible, particularly on single-board robotic platforms.

To this end, traditional approaches based on various feature contrast evaluation techniques [13, 67] are often practical choices for saliency estimation by visually-guided underwater robots. These techniques encode low-level image-based features (*e.g.*, color, texture, object shapes or contours) into super-pixel descriptors [32, 24] to subsequently infer saliency by quantifying their relative *distinctness* on a global scale. Such bottom-up approaches are computationally light and are useful as pre-processing steps for faster visual search [32, 27] and exploration tasks [13, 43]. However, they do not provide a standalone generalizable solution for SOD in underwater imagery. A few recently proposed approaches attempt to address this issue by learning more generalizable SOD solutions from large collection of annotated underwater data [19, 50]. These approaches and other SOTA deep visual models have reported inspiring results for underwater SOD and relevant problems [24, 21]. Nevertheless, their utility and performance margins for real-time underwater robotic applications have not been explored in-depth in the literature.

In this paper, we formulate a robust and efficient solution for saliency-guided visual attention modeling (SVAM) by harnessing the power of both bottom-up and top-down learning in a novel encoder-decoder model named **SVAM-Net**. We design two spatial attention modules (SAMs) named **SAM$^{\text{bu}}$** and **SAM$^{\text{td}}$** to effectively exploit the coarse-level and fine-level semantic features along the bottom-up and top-down learning pathways, respectively. SAM$^{\text{bu}}$ utilizes the semantically rich low-dimensional features extracted by the encoder to perform an abstract yet reasonably accurate saliency estimation. Concurrently, SAM$^{\text{td}}$ combines the multi-scale hierarchical features of the encoder to progressively decode the information for robust SOD. A residual refinement module (**RRM**) further sharpens the initial SAM$^{\text{td}}$ predictions to provide fine-grained localization of the salient objects. To balance the high degree of *refined* gradient flows from the later SVAM-Net layers, we deploy an auxiliary SAM named **SAM$^{\text{aux}}$** that guides the spatial activations of early encoding layers and ensures smooth end-to-end learning.

In addition to sketching the conceptual design, we present a holistic training pipeline of SVAM-Net and its variants. The end-to-end learning is supervised by six loss functions which are selectively applied at the final stages of SAM$^{\text{aux}}$, SAM$^{\text{bu}}$, SAM$^{\text{td}}$, and RRM. These functions evaluate information loss and boundary localization errors in the respective SVAM-Net predictions and collectively ensure effective SOD learning. In our evaluation, we analyze SVAM-Net's performance in standard quantitative and qualitative terms on three benchmark datasets named UFO-120 [21], MUED [25], and SUIM [19]. We also conduct performance evaluation on **USOD**, which we prepare as a new challenging test set for underwater SOD. Without data-specific tuning or task-specific model adaptation, SVAM-Net outperforms other existing solutions on these benchmark datasets; more importantly, it exhibits considerably better generalization performance on random unseen test cases of natural underwater scenes.

Lastly, we present several design choices of SVAM-Net, analyze their computational aspects, and discuss the corresponding use cases. The end-to-end SVAM-Net model offers over 20 frames per second (FPS) inference rate on a single GPU. Moreover, the decoupled SAM$^{\text{bu}}$ branch offers significantly faster rates, *e.g.*, over 86 FPS on a GPU and over 21 FPS on single-board computers. As illustrated in Fig. 1, robust saliency estimates of SVAM-Net at such speeds are ideal for fast visual attention modeling in robotic deployments. We further demonstrate its usability benefits for important applications such as object detection, image enhancement, and image super-resolution by visually-guided underwater robots. The SVAM-Net model, USOD dataset, and relevant resources are released at `http://irvlab.cs.umn.edu/visual-attention-modeling/svam`.

II. Background & Related Work

A. Salient Object Detection (SOD)

SOD is a successor to the human fixation prediction (FP) problem [23] that aims to identify *fixation points* that human viewers would focus on at first glance. While FP originates from research in cognition and psychology [31, 59], SOD is more of a visual perception problem explored by the computer vision and robotics community [8, 27, 39]. The history of SOD dates back to the work of Liu *et al.* [40] and Achanta *et al.* [2], which make use of multi-scale contrast, center-surround histogram, and frequency-domain cues to (learn to) infer saliency in image space. Other traditional SOD models rely on various low-level saliency cues such as point-based features [12], local and global contrast [10, 29], background prior [64], etc. Please refer to [7] for a more comprehensive overview of non-deep learning-based SOD models.

Recently, deep convolutional neural network (CNN)-based models have set new SOTA for SOD [8, 60]. Li *et al.* [35] and Zhao *et al.* [65] use sequential CNNs to extract multi-scale hierarchical features to infer saliency on global and local contexts. Recurrent fully convolutional networks (FCNs) [57, 5] are also used to progressively refine saliency estimates. In particular, Wang *et al.* [59] use multi-stage convolutional LSTMs for saliency estimation guided by fixation maps. Later in [61], they explore the benefits of integrating bottom-up and top-down recurrent modules for co-operative SOD learning. Since the feed-forward computational pipelines lack a feedback strategy [36, 33], recurrent modules offer more learning capacity via self-correction. However, they are prone to the vanishing

gradient problem and also require meticulous design choices in their feedback loops [61]. To this end, top-down models with UNet-like architectures [39, 38] provide more consistent learning behavior. These models typically use a powerful backbone network (*e.g.*, VGG [53], ResNet [15]) to extract a hierarchical pyramid of features, then perform a coarse-to-fine feature distillation via mirrored skip-connections. Subsequent research introduces the notions of short connections [16] and guided super-pixel filtering [17] to learn to infer compact and uniform saliency maps.

B. SOD and SVAM by Underwater Robots

The most essential capability of visually-guided AUVs is to identify interesting and relevant image regions to make effective operational decisions. The existing systems and solutions for visual saliency estimation can be categorically discussed from the perspectives of model adaptation [30, 19], high-level robot tasks [13, 48], and feature evaluation pipeline [24, 43]. Since we already discussed the particulars of bottom-up and top-down computational pipelines, our following discussion is schematized based on the *model* and *task* perspectives.

Visual saliency estimation approaches can be termed as either *model-based* or *model-free*, depending on whether the robot models any prior knowledge of the target salient objects and features. The model-based techniques are particularly beneficial for fast visual search [30, 26], enhanced object detection [67, 50], and monitoring applications [44]. For instance, Maldonado-Ramírez *et al.* [43] use ad hoc visual descriptors learned by a convolutional autoencoder to identify salient landmarks for fast place recognition. Moreover, Koreitem *et al.* [30] use a bank of pre-specified image patches (containing interesting objects or relevant scenes) to learn a similarity operator that guides the robot's visual search in an unconstrained setting. Such similarity operators are essentially spatial saliency predictors which assign a degree of *relevance* to the visual scene based on the prior model-driven knowledge of what may constitute as salient, *e.g.*, coral reefs [4], companion divers [67], wrecks [19], fish [47], etc.

On the other hand, model-free approaches are more feasible for autonomous exploratory applications [14, 49]. The early approaches date back to the work of Edgington *et al.* [12] that uses binary morphology filters to extract salient features for automated event detection. Subsequent approaches adopt various feature contrast evaluation techniques that encode low-level image-based features (*e.g.*, color, luminance, texture, object shapes) into super-pixel descriptors [42, 32, 55]. These low-dimensional representations are then exploited by heuristics or learning-based models to infer global saliency. For instance, Girdhar *et al.* [13] formulate an online topic-modeling scheme that encodes visible features into a low-dimensional semantic descriptor, then adopt a probabilistic approach to compute a *surprise score* for the current observation based on the presence of high-level patterns in the scene. Moreover, Kim *et al.* [27] introduce an online bag-of-words scheme to measure intra- and inter-image saliency estimation for robust key-frame selection in SLAM-based navigation. Wang *et al.* [55] encode multi-scale image features into a topographical descriptor, then apply Bhattacharyya measure to extract salient RoIs by segmenting out the background. These bottom-up approaches are effective in pre-processing raw visual data to identify point-based or region-based salient features; however, they do not provide a generalizable object-level solution for underwater SOD.

Nevertheless, several contemporary research [9, 24, 68, 37] report inspiring results for object-level saliency estimation and foreground segmentation in underwater imagery. Chen *et al.* [9] use a level set-based formulation that exploits various low-level features for underwater SOD. Moreover, Jian *et al.* [24] perform principal components analysis (PCA) in quaternionic space to compute pattern distinctness and local contrast to infer directional saliency. These methods are also model-free and adopt a bottom-up feature evaluation pipeline. In contrast, Islam *et al.* [21] incorporates multi-scale hierarchical features extracted by a top-down deep residual model to identify salient foreground pixels for global contrast enhancement. In this paper, we formulate a generalized solution for underwater SOD and demonstrate its utility for SVAM by visually-guided underwater robots. It combines the benefits of bottom-up and top-town feature evaluation in a compact end-to-end pipeline, provides SOTA performance, and ensures computational efficiency for robotic deployments in both search-based and exploration-based applications.

III. Model & Training Pipeline

A. SVAM-Net Architecture

As illustrated in Fig. 2, the major components of our SVAM-Net model are: the backbone encoder network, the top-down SAM (SAM^{td}), the residual refinement module (RRM), the bottom-up SAM (SAM^{bu}), and the auxiliary SAM (SAM^{aux}). These components are tied to an end-to-end architecture for a supervised SOD learning.

1) Backbone Encoder Network: We use the first five sequential blocks of a standard VGG-16 network [53] as the backbone encoder in our model. Each of these blocks consist of two or three convolutional (`Conv`) layers for feature extraction, which are then followed by a pooling (`Pool`) layer for spatial down-sampling. For an input dimension of $256 \times 256 \times 3$, the composite encoder blocks $\mathbf{e_1} \rightarrow \mathbf{e_5}$ learn $128 \times 128 \times 64$, $64 \times 64 \times 128$, $32 \times 32 \times 256$, $16 \times 16 \times 512$, and $8 \times 8 \times 512$ feature-maps, respectively. These multi-scale deep visual features are jointly exploited by the attention modules of SVAM-Net for effective learning.

2) Top-Down SAM (SAM^{td}): Unlike the existing U-Net-based architectures, we adopt a partial top-down decoder $\mathbf{d_5} \rightarrow \mathbf{d_2}$ that allows skip-connections from mirrored encoding layers. We consider the mirrored conjugate pairs as $\mathbf{e_4} \sim \mathbf{d_5}$, $\mathbf{e_3} \sim \mathbf{d_4}$, $\mathbf{e_2} \sim \mathbf{d_3}$, and $\mathbf{e_1} \sim \mathbf{d_2}$. Such asymmetric pairing facilitates the use of a standalone de-convolutional (`DeConv`) layer following $\mathbf{d_2}$ rather than using another composite decoder block, which we have found to be redundant (during ablation experiments). The composite blocks $\mathbf{d_5} \rightarrow \mathbf{d_2}$

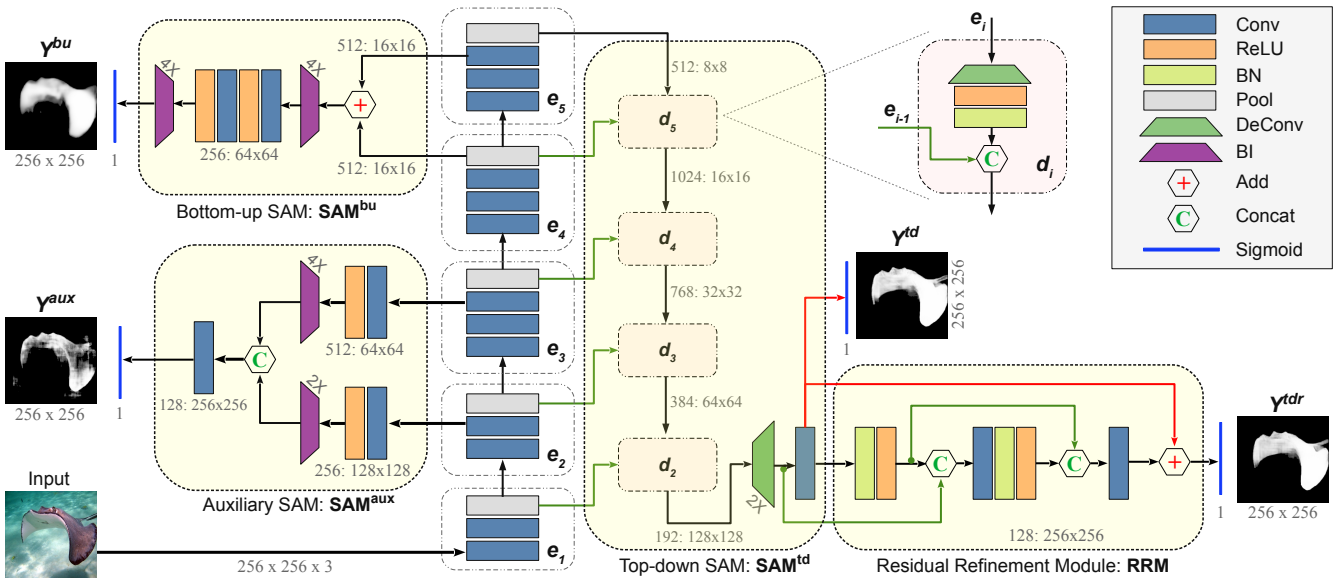


Fig. 2: The detailed architecture of SVAM-Net is shown. The input image is passed over to the sequential encoding blocks $\{\mathbf{e_1} \rightarrow \mathbf{e_5}\}$ for multi-scale convolutional feature extraction. Then, SAM$^{\text{td}}$ gradually up-samples these hierarchical features and fuses them with mirrored skip-connections along the top-down pathway $\{\mathbf{d_5} \rightarrow \mathbf{d_2}\}$ to subsequently generate an intermediate output Y^{td}; the RRM refines this intermediate representation and produces the final SOD output Y^{tdr}. Moreover, SAM$^{\text{bu}}$ exploits the features of $\mathbf{e_4}$ and $\mathbf{e_5}$ to generate an abstract SOD prediction Y^{bu} along the bottom-up pathway; additionally, SAM$^{\text{aux}}$ performs an auxiliary refinement on the $\mathbf{e_2}$ and $\mathbf{e_3}$ features that facilitates a smooth end-to-end SOD learning. Implementation of this training pipeline is here: `https://github.com/xahidbuffon/SVAM-Net`.

decode $16 \times 16 \times 1024$, $32 \times 32 \times 768$, $64 \times 64 \times 384$, and $128 \times 128 \times 192$ feature-maps, respectively. Following $\mathbf{d_2}$ and the standalone `DeConv` layer, an additional `Conv` layer learns $256 \times 256 \times 128$ feature-maps to be the final output of SAM$^{\text{td}}$ as $\mathbf{S}^{td}_{coarse} = \mathbf{SAM}^{\text{td}}(\mathbf{e_1} : \mathbf{e_5})$. These feature-maps are passed along two branches (see Fig. 2); on the shallow branch, a `Sigmoid` layer is applied to generate an intermediate SOD prediction Y^{td}, while the other deeper branch incorporates residual layers for subsequent refinement.

3) Residual Refinement Module (RRM): We further design a residual module to effectively refine the top-down coarse saliency predictions by learning the desired residuals as

$$\mathbf{S}^{tdr}_{refined} = \mathbf{S}^{td}_{coarse} + \mathbf{S}^{rrm}_{residual}.$$

Such refinement modules [46, 58] are designed to address the loss of regional probabilities and boundary localization in intermediate SOD predictions. While the existing methodologies use iterative recurrent modules or additional residual encoder-decoder networks [46], we deploy only two sequential residual blocks and a `Conv` layer for the refinement. Each residual block consists of a `Conv` layer followed by batch normalization (`BN`) [18] and a rectified linear unit (`ReLU`) activation [45]. The entire RRM operates on a feature dimension of $256 \times 256 \times 128$; following refinement, a `Sigmoid` layer squashes the feature-maps to generate a single-channel output Y^{tdr}, which is the final SOD prediction of SVAM-Net.

4) Bottom-Up SAM (SAM^{bu}): A high degree of supervision at the final layers of RRM and SAM$^{\text{td}}$ forces the backbone encoding layers to learn effective multi-scale features. In SAM$^{\text{bu}}$, we exploit these low-resolution yet semantically rich features for efficient bottom-up SOD learning. Specifically, we combine the feature-maps of dimension $16 \times 16 \times 512$ from $\mathbf{e_4}$ (`Pool4`) and $\mathbf{e_5}$ (`Conv53`), and subsequently learn the bottom-up spatial attention as $\mathbf{S}^{bu} = \mathbf{SAM}^{\text{bu}}(\mathbf{e_4}\texttt{.Pool4}, \mathbf{e_5}\texttt{.Conv53})$. On the combined input feature-maps, SAM$^{\text{bu}}$ incorporates $4\times$ bilinear interpolation (`BI`) followed by two `Conv` layers with `ReLU` activation to learn $64 \times 64 \times 256$ feature-maps. Subsequently, another `BI` layer performs $4\times$ spatial up-sampling to generate $\mathbf{S}^{bu}$; lastly, a `Sigmoid` layer is applied to generate the single-channel output Y^{bu}.

5) Auxiliary SAM (SAM^{aux}): We excluded the features of early encoding layers for bottom-up SOD learning in SAM$^{\text{bu}}$ for two reasons: *i*) they lack important semantic details despite their higher resolutions [66, 63], and *ii*) it is counter-intuitive to our goal of achieving fast bottom-up inference. Nevertheless, we adopt a separate attention module that refines the features of $\mathbf{e_2}$ (`Conv22`) and $\mathbf{e_3}$ (`Conv33`) as $\mathbf{S}^{aux} = \mathbf{SAM}^{\text{aux}}(\mathbf{e_2}\texttt{.Conv22}, \mathbf{e_3}\texttt{.Conv33})$. Here, a `Conv` layer with `ReLU` activation is applied separately on these inputs, followed by a $2\times$ or $4\times$ `BI` layer (see Fig. 2). Their combined output features are passed to a `Conv` layer to subsequently generate $\mathbf{S}^{aux}$ of dimension $256 \times 256 \times 128$. The sole purpose of this module is to backpropagate additional *refined* gradients via supervised loss applied to the `Sigmoid` output Y^{aux}. This auxiliary refinement facilitates smooth feature learning while adding no computational overhead in the bottom-up inference through SAM$^{\text{bu}}$ (as we discard SAM$^{\text{aux}}$ after training).

B. Learning Objectives and Training

SOD is a pixel-wise binary classification problem that refers to the task of identifying all *salient* pixels in a given image.

We formulate the problem as learning a function $f : X \rightarrow Y$, where X is the input image domain and Y is the target saliency map, *i.e.*, saliency probability for each pixel. As illustrated in Fig. 2, SVAM-Net generates saliency maps from four output layers, namely $Y^{aux} = \sigma(\mathbf{S}^{aux})$, $Y^{bu} = \sigma(\mathbf{S}^{bu})$, $Y^{td} = \sigma(\mathbf{S}^{td}_{coarse})$, and $Y^{tdr} = \sigma(\mathbf{S}^{tdr}_{refined})$ where σ is the `Sigmoid` function. Hence, the learning pipeline of SVAM-Net is expressed as $f : X \rightarrow Y^{aux},\ Y^{bu},\ Y^{td},\ Y^{tdr}$.

We adopt six loss components to collectively evaluate the information loss and boundary localization error for the supervised training of SVAM-Net. To quantify the information loss, we use the standard binary cross-entropy (BCE) function [11] that measures the disparity between predicted saliency map $\hat{Y}$ and ground truth Y as

$$\mathcal{L}_{BCE}(\hat{Y}, Y) = \mathbb{E}\big[-Y_p \log \hat{Y}_p - (1 - Y_p) \log(1 - \hat{Y}_p)\big]. \quad (1)$$

We also use the analogous weighted cross-entropy loss function $\mathcal{L}_{WCE}(\hat{Y}, Y)$, which is widely adopted in SOD literature to handle the imbalance problem of the number of salient pixels [59, 66]. While $\mathcal{L}_{WCE}$ provides general guidance for accurate saliency estimation, we use the 2D Laplace operator to further ensure robust boundary localization of salient objects. Specifically, we utilize the 2D Laplacian kernel $K_{Laplace}$ to evaluate the divergence of image gradients [66] in the predicted saliency map and respective ground truth as

$$\Delta\hat{Y} = \big|\texttt{tanh}\big(\texttt{conv}(\hat{Y},\ K_{Laplace})\big)\big|, \text{ and} \quad (2)$$
$$\Delta Y = \big|\texttt{tanh}\big(\texttt{conv}(Y,\ K_{Laplace})\big)\big|. \quad (3)$$

Then, we measure the boundary localization error as

$$\mathcal{L}_{BLE}(\hat{Y}, Y) = \mathcal{L}_{WCE}(\Delta\hat{Y}, \Delta Y). \quad (4)$$

We deploy a two-step training process for SVAM-Net to ensure robust and effective SOD learning. First, the backbone encoder and SAM^{td} are pre-trained holistically with combined terrestrial (DUTS [56]) and underwater data (SUIM [19], UFO-120 [21]). The DUTS training set (DUTS-TR) has 10553 terrestrial images, whereas the SUIM and UFO-120 datasets contain a total of 3025 underwater images for training and validation. This large collection of diverse training instances facilitates a comprehensive learning of a generic SOD function. We supervise the training by applying $\mathcal{L}_{PT} \equiv \mathcal{L}_{BCE}(Y^{td}, Y)$ loss at the sole output layer of SAM^{td}. The SGD optimizer [28] is used for the iterative learning with an initial rate of $1e^{-2}$ and 0.9 momentum, which is decayed exponentially by a drop rate of 0.5 after every 8 epochs; other hyper-parameters are listed in Table I.

Subsequently, the pre-trained weights are exported into the SVAM-Net model for its end-to-end training on underwater imagery. The loss components applied at the output layers of SAM^{aux}, SAM^{bu}, SAM^{td}, and SAM^{tdr} are

$$\mathcal{L}^{aux}_{E2E} \equiv \mathcal{L}_{BCE}(Y^{aux}, Y), \quad (5)$$
$$\mathcal{L}^{bu}_{E2E} \equiv \lambda_w \mathcal{L}_{WCE}(Y^{bu}, Y) + \lambda_b \mathcal{L}_{BLE}(Y^{bu}, Y), \quad (6)$$
$$\mathcal{L}^{td}_{E2E} \equiv \lambda_w \mathcal{L}_{WCE}(Y^{td}, Y) + \lambda_b \mathcal{L}_{BLE}(Y^{td}, Y), \text{ and} \quad (7)$$
$$\mathcal{L}^{tdr}_{E2E} \equiv \mathcal{L}_{BCE}(Y^{tdr}, Y). \quad (8)$$

TABLE I: The two-step training process of SVAM-Net and corresponding learning parameters [b: batch size; e: number of epochs; N_{train}: size of the training data; f_{opt}: global optimizer; η_o: initial learning rate; m: momentum; τ: decay drop rate].

	Backbone Pre-training	End-to-end Training
Pipeline	$\{\mathbf{e_{1:5}} \rightarrow \text{SAM}^{td}\}$	Entire SVAM-Net
Objective	$\mathcal{L}_{PT} \equiv \mathcal{L}_{BCE}(Y^{td}, Y)$	$\mathcal{L}_{E2E}$ (see Eq. 9)
Data	DUTS + SUIM + UFO-120	SUIM + UFO-120
$b \odot e\ /\ N_{train}$	$4 \odot 90\ /\ 13578$	$4 \odot 50\ /\ 3025$
$f_{opt}(\eta_o, m, \tau)$	SGD$(1e^{-2}, 0.9, 0.5)$	Adam$(3e^{-4}, 0.5, \times)$

We formulate the combined objective function as:

$$\mathcal{L}_{E2E} = \lambda_{aux}\mathcal{L}^{aux}_{E2E} + \lambda_{bu}\mathcal{L}^{bu}_{E2E} + \lambda_{td}\mathcal{L}^{td}_{E2E} + \lambda_{tdr}\mathcal{L}^{tdr}_{E2E}. \quad (9)$$

Here, $\lambda_\square$ symbols are scaling factors that represent the contributions of respective loss components; their values are empirically tuned as hyper-parameters. In our evaluation, the selected values of λ_w, λ_b, λ_{aux}, λ_{bu}, λ_{td}, and λ_{tdr} are 0.7, 0.3, 0.5, 1.0, 2.0, and 4.0, respectively. As shown in Table I, we use the Adam optimizer [28] for the global optimization of $\mathcal{L}_{E2E}$ with a learning rate of $3e^{-4}$ and a momentum of 0.5.

C. SVAM-Net Inference

Once the end-to-end training is completed, we decouple a bottom-up and a top-down branch of SVAM-Net for fast inference. As illustrated in Fig. 3, the $\{\mathbf{e_{1:5}} \rightarrow \text{SAM}^{td} \rightarrow \text{RRM}\}$ branch is the default SVAM-Net top-down pipeline that generates fine-grained saliency maps; here, we discard the SAM^{aux} and SAM^{bu} modules to avoid unnecessary computation. On the other hand, we exploit the shallow bottom-up branch, *i.e.*, the $\{\mathbf{e_{1:5}} \rightarrow \text{SAM}^{bu}\}$ pipeline to generate rough yet reasonably accurate saliency maps at a significantly faster rate. Here, we discard SAM^{aux} and both the top-down modules (SAM^{td} and RRM); we denote this computationally light pipeline as SVAM-Net^{Light}.

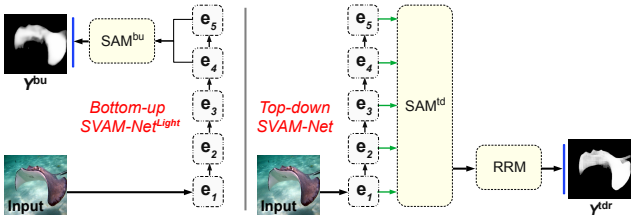
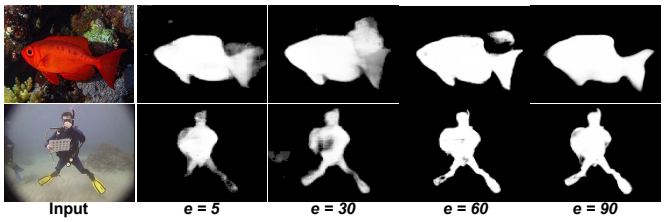


Fig. 3: The decoupled pipelines for bottom-up and top-down inference: SVAM-Net^{Light} and SVAM-Net (default), respectively.

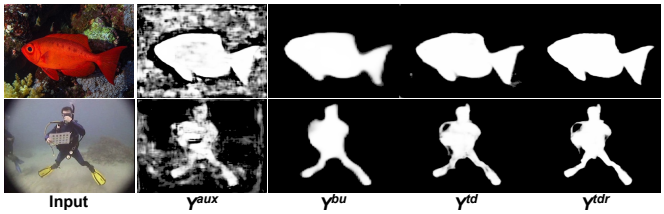
IV. Experimental Evaluation

A. Implementation Details and Ablation Studies

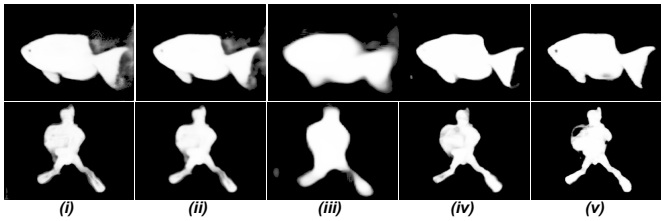
As mentioned in §III-B, SVAM-Net training is supervised by paired data $(\{X\}, \{Y\})$ to learn a pixel-wise predictive function $f : X \rightarrow Y^{aux},\ Y^{bu},\ Y^{td},\ Y^{tdr}$. TensorFlow and Keras libraries [1] are used to implement its network architecture and optimization pipelines (Eq. 1-9). A Linux machine with two NVIDIA™ GTX 1080 graphics cards is used for its backbone pre-training and end-to-end training with the learning parameters provided in Table I.



(a) Spatial saliency learning over $e = 100$ epochs of backbone pre-training; outputs of Y^{td} are shown after 5, 30, 60, and 90 epochs.



(b) Snapshots of SVAM-Net output after 40 epochs of subsequent end-to-end training; notice the spatial attention of early encoding layers (in Y^{aux}) and the gradual progression and refinement by the deeper layers (through $Y^{bu} \rightarrow Y^{td} \rightarrow Y^{tdr}$).



(c) Results of ablation experiments (for the same input images) showing contributions of various attention modules and loss functions in the SOD learning: (i) without $\mathcal{L}_{BLE}$ ($\lambda_b = 0$, $\lambda_w = 1$), (ii) without SAM$^{\text{aux}}$ and SAM$^{\text{bu}}$ ($\lambda_{aux} = \lambda_{bu} = 0$), (iii) without SAM$^{\text{td}}$ and RRM ($\lambda_{td} = \lambda_{tdr} = 0$), (iv) without RRM ($\lambda_{tdr} = 0$), and (v) without backbone pre-training.

Fig. 4: Demonstrations of progressive learning behavior of SVAM-Net and effectiveness of its learning components.

We demonstrate the progression of SOD learning by SVAM-Net and visualize the contributions of its learning components in Fig. 4. The first stage of learning is guided by supervised pre-training with over 13.5K instances including both terrestrial and underwater images. This large-scale training facilitates effective feature learning in the backbone encoding layers and by SAM$^{\text{td}}$. As Fig. 4a shows, the $\{\mathbf{e_{1:5}} \rightarrow \text{SAM}^{\text{td}}\}$ pipeline learns spatial attention with a reasonable precision within 90 epochs. We found that it is crucial to not over-train the backbone for ensuring a smooth and effective end-to-end learning with the integration of SAM$^{\text{aux}}$, SAM$^{\text{bu}}$, and RRM. As illustrated in Fig. 4b, the subsequent end-to-end training on underwater imagery enables more accurate and fine-grained saliency estimation by SVAM-Net.

Moreover, we conduct a series of ablation experiments to visually inspect the effects of various loss functions and attention modules in the learning. As Fig. 4c demonstrates, the boundary awareness (enforced by $\mathcal{L}_{BLE}$) and bottom-up attention modules (SAM$^{\text{aux}}$ and SAM$^{\text{bu}}$) are essential to achieve precise localization and sharp contours of the salient objects. It also shows that important details are missed when we incorporate only bottom-up learning, *i.e.*, without SAM$^{\text{td}}$ and subsequent delicate refinements by RRM. Besides, the backbone pre-training step is important to ensure generalizability in the SOD learning and as an effective way to combat the lack of large-scale annotated underwater datasets.

B. Evaluation Data Preparation and Metrics

For performance evaluation of SVAM-Net and other existing SOD methods, we use four widely-used metrics [8, 46, 58]: F-measure ($\mathbf{F}_\beta$), S-measure ($\mathbf{S}_m$), Mean absolute error (MAE), and Precision-recall (PR) curves. We conduct the evaluation on the test sets of three publicly available datasets: SUIM [19], UFO-120 [21], and MUED [25]. In addition, we prepare a challenging test set named **USOD** to evaluate underwater SOD methods. It contains 300 natural underwater images which we exhaustively compiled to ensure diversity in the objects categories, water-body, optical distortions, and aspect ratio of the salient objects. More detailed explanations of the data preparation processes and evaluation metrics are provided in the supplementary materials.

C. SOD Performance Evaluation

1) Quantitative and Qualitative Analysis: For performance comparison, we consider the following six methods that are widely used for underwater SOD and/or saliency estimation: (i) SOD by Quaternionic Distance-based Weber Descriptor (QDWD) [24], (ii) saliency estimation by the Segmentation of Underwater IMagery Network (SUIM-Net) [19], (iii) saliency prediction by the Deep Simultaneous Enhancement and Super-Resolution (Deep SESR) model [21], (iv) SOD by a Level Set-guided Method (LSM) [9], (v) Saliency Segmentation by evaluating Region Contrast (SSRC) [37], and (vi) SOD by Saliency-based Adaptive Object Extraction (SAOE) [55]. We also include the performance margins of four SOTA SOD models: (i) Boundary-Aware Saliency Network (BASNet) [46], (ii) Pyramid Attentive and salient edGE-aware Network (PAGE-Net) [62], (iii) Attentive Saliency Network (ASNet) [59], and (iv) Cascaded Partial Decoder (CPD) [63]. We use their publicly released weights (pre-trained on terrestrial imagery) and further train them on combined SUIM and UFO-120 data by following the same setup as SVAM-Net (see Table I).

As the results in the first part of Table II suggest, SVAM-Net outperforms all the underwater SOD models in comparison with significant margins. Although QDWD and SUIM-Net perform reasonably well on particular datasets (*e.g.*, SUIM, and MUED, respectively), their $\mathbf{F}_\beta^{max}$, $\mathbf{S}_m$, and $\mathbf{MAE}$ scores are much lower; in fact, their scores are comparable to and often lower than those of SVAM-Net$^{\text{Light}}$. The LSM, Deep SESR, SSRC, and SAOE models offer even lower scores than SVAM-Net$^{\text{Light}}$. The respective comparisons of PR curves shown in Fig. 5 and Fig. 6 further validate the superior performance of SVAM-Net and SVAM-Net$^{\text{Light}}$ by an area-under-the-curve (AUC)-based analysis. Moreover, Fig. 7 demonstrates that SVAM-Net-generated saliency maps are accurate with precisely segmented boundary pixels in general. Although not as fine-grained, SVAM-Net$^{\text{Light}}$ also generates reasonably

TABLE II: Quantitative performance comparison of SVAM-Net and SVAM-Net$^{\text{Light}}$ with existing SOD solutions and SOTA methods for both underwater (first six) and terrestrial (last four) domains are shown. All scores of maximum F-measure ($\mathbf{F}_\beta^{max}$), S-measure ($\mathbf{S}_m$), and mean absolute error ($\mathbf{MAE}$) are evaluated in $[0, 1]$; top two scores (column-wise) are indicated by red (best) and blue (second best) colors.

Method	SUIM [19]			UFO-120 [21]			MUED [25]			**USOD**		
	$\mathbf{F}_\beta^{max}$ (↑)	$\mathbf{S}_m$ (↑)	$\mathbf{MAE}$ (↓)	$\mathbf{F}_\beta^{max}$ (↑)	$\mathbf{S}_m$ (↑)	$\mathbf{MAE}$ (↓)	$\mathbf{F}_\beta^{max}$ (↑)	$\mathbf{S}_m$ (↑)	$\mathbf{MAE}$ (↓)	$\mathbf{F}_\beta^{max}$ (↑)	$\mathbf{S}_m$ (↑)	$\mathbf{MAE}$ (↓)
SAOE [55]	0.2698	0.3965	0.4015	0.4011	0.4420	0.3752	0.2978	0.3045	0.3849	0.2520	0.2418	0.4678
SSRC [37]	0.3015	0.4226	0.3028	0.3836	0.4534	0.4125	0.4040	0.3946	0.2295	0.2143	0.2846	0.3872
Deep SESR [21]	0.3838	0.4769	0.2619	0.4631	0.5146	0.3437	0.3895	0.3565	0.2118	0.3914	0.4868	0.3030
LSM [9]	0.5443	0.5873	0.1504	0.6908	0.6770	0.1396	0.4174	0.4025	0.1934	0.6775	0.6768	0.1186
SUIM-Net [19]	0.8413	0.8296	0.0787	0.6628	0.6790	0.1427	0.5686	0.5070	0.1227	0.6818	0.6754	0.1386
QDWD [24]	0.7328	0.6978	0.1129	0.7074	0.7044	0.1368	0.6248	0.5975	0.0771	0.7750	0.7245	0.0989
SVAM-Net$^{\text{Light}}$	0.8254	0.8356	0.0805	0.8428	0.8613	0.0663	0.8492	0.8588	0.0184	0.8703	0.8723	0.0619
SVAM-Net	0.8830	0.8607	0.0593	0.8919	0.8808	0.0475	0.9013	0.8692	0.0137	0.9162	0.8832	0.0450
BASNet [46]	0.7212	0.6873	0.1142	0.7609	0.7302	0.1108	0.8556	0.8820	0.0145	0.8425	0.7919	0.0745
PAGE-Net [62]	0.7481	0.7207	0.1028	0.7518	0.7522	0.1062	0.6849	0.7136	0.0442	0.8430	0.8017	0.0713
ASNet [59]	0.7344	0.6740	0.1168	0.7540	0.7272	0.1153	0.6413	0.7476	0.0370	0.8310	0.7732	0.0798
CPD [63]	0.6679	0.6254	0.1387	0.6947	0.6880	0.3752	0.7624	0.7311	0.0330	0.7877	0.7436	0.0917

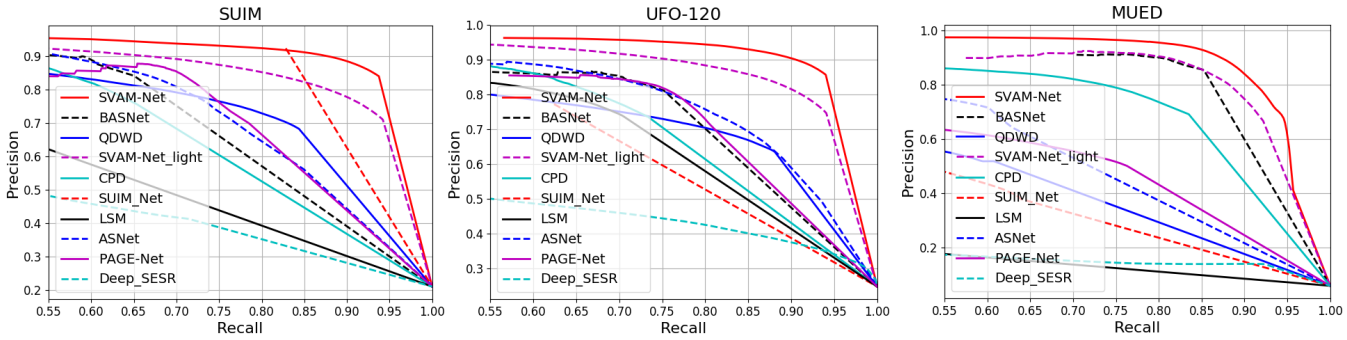


Fig. 5: Comparisons of PR curves on three benchmark datasets are shown; to maintain clarity, we consider the top ten SOD models based on the results shown in Table II.

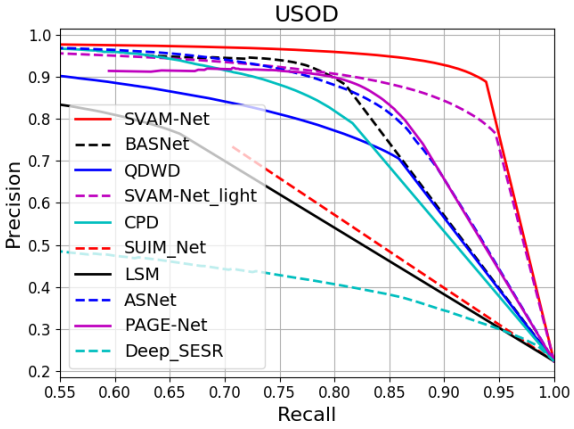


Fig. 6: Comparison of PR curves on USOD dataset is shown for the top ten SOD models based on the results shown in Table II.

well-localized saliency maps that are still more accurate and consistent compared to the existing models. These results corroborate our earlier discussion on the challenges and lack of advancements of underwater SOD literature .

For a comprehensive validation of SVAM-Net, we compare the performance margins of SOTA SOD models trained through the same learning pipeline. As shown in Fig. 7, the saliency maps of BASNet, PAGE-Net, ASNet, and CPD are mostly accurate and often comparable to SVAM-Net-generated maps. The quantitative results of Table II and Fig. 5-6 also confirm their competitive performance over all datasets. Given the substantial learning capacities of these models, one may exhaustively find a better choice of hyper-parameters that further improves their baseline performances. Nevertheless, unlike these standard models, SVAM-Net incorporates a considerably shallow computational pipeline and offers an even lighter bottom-up sub-network (SVAM-Net$^{\text{Light}}$) that ensures fast inference on single-board devices. Next, we demonstrate SVAM-Net's generalization performance and discuss its utility for underwater robotic deployments.

V. Operational Feasibility & Use Cases

A. Single-board Deployments

As Table III shows, SVAM-Net offers an end-to-end runtime of 49.82 milliseconds (ms) per-frame, *i.e.*, 20.07 frames-per-second (FPS) on a single NVIDIA™ GTX 1080 GPU. Moreover, SVAM-Net$^{\text{Light}}$ operates at a much faster rate of 11.60 ms per-frame (86.15 FPS). These inference rates surpass the reported speeds of SOTA SOD models [60, 8] and are adequate for GPU-based use in real-time applications. More importantly, SVAM-Net$^{\text{Light}}$ runs at 21.77 FPS rate on a single-board computer named NVIDIA™ Jetson AGX Xavier with an on-board memory requirement of only 65 MB. These compu-

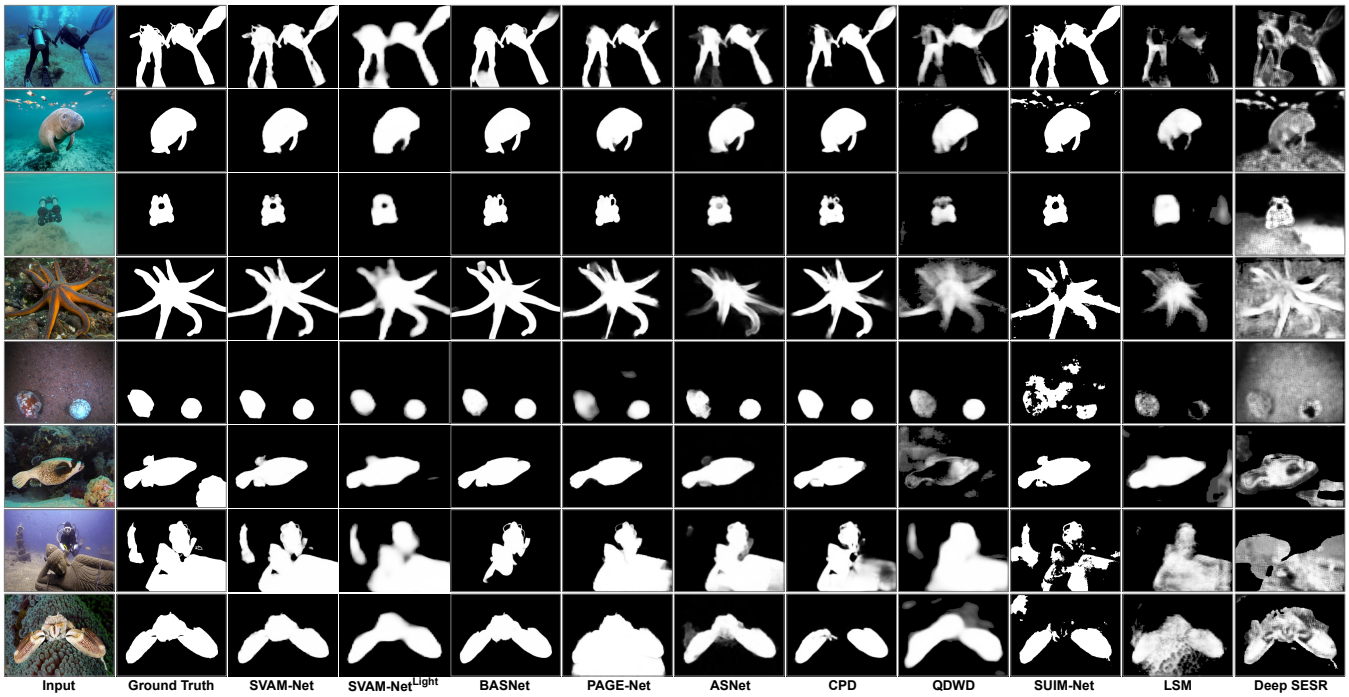


Fig. 7: A few qualitative comparisons of saliency maps generated by the top ten SOD models (based on the results of Table II). From the top: first four images belong to the test sets of SUIM [19] and UFO-120 [21], the next one to MUED [25], whereas the last three images belong to the proposed USOD dataset.

tational aspects make SVAM-Net$^{\text{Light}}$ ideally suited for single-board robotic deployments, and justify our design intuition of decoupling the bottom-up pipeline $\{\mathbf{e_{1:5}} \rightarrow \text{SAM}^{\text{bu}}\}$ from the SVAM-Net architecture (see §III-C).

B. Practical Use Cases

In the last two sections, we discussed the practicalities involved in designing a generalized underwater SOD model and identified several drawbacks of existing solutions such as QDWD, SUIM-Net, LSM, and Deep SESR. Specifically, we showed that their predicted saliency maps lack important details, exhibit improperly segmented object boundaries, and incur plenty of false-positive pixels (see §IV-C and Fig. 7). Although such sparse detection of salient pixels can be useful in specific higher-level tasks (*e.g.*, contrast enhancement [21], rough foreground extraction [37]), these models are not as effective for general-purpose SOD. It is evident from our experimental results that the proposed SVAM-Net model overcomes these limitations and offers a robust SOD solution for underwater imagery. For underwater robot vision, in particular, SVAM-Net$^{\text{Light}}$ can facilitate faster processing in a host of visual perception tasks. As seen in Fig. 8, we demonstrate its effectiveness for two such important use cases.

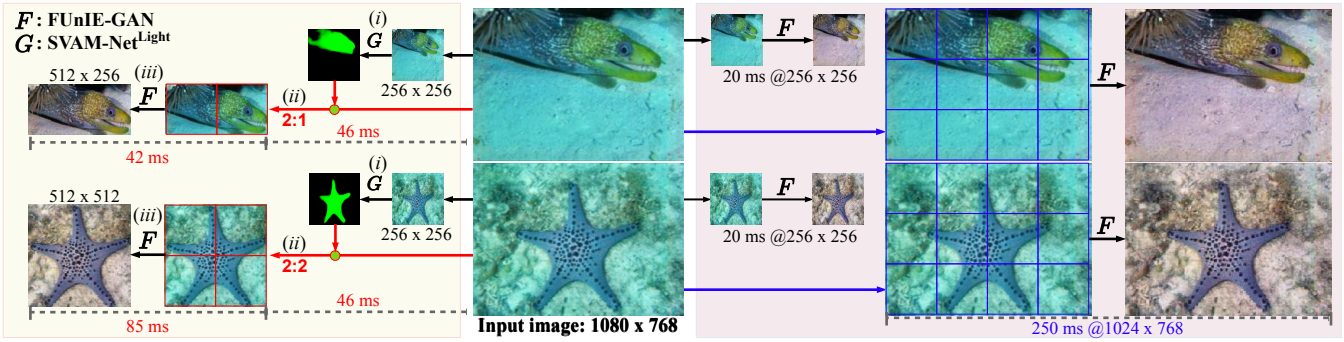
TABLE III: Run-time comparison for SVAM-Net and SVAM-Net$^{\text{Light}}$ on a GTX 1080 GPU and on a single-board AGX Xavier device.

	SVAM-Net	**SVAM-Net$^{\text{Light}}$**
GTX 1080	49.82 ms (20.07 FPS)	11.60 ms (86.15 FPS)
AGX Xavier	222.2 ms (4.50 FPS)	**45.93 ms (21.77 FPS)**

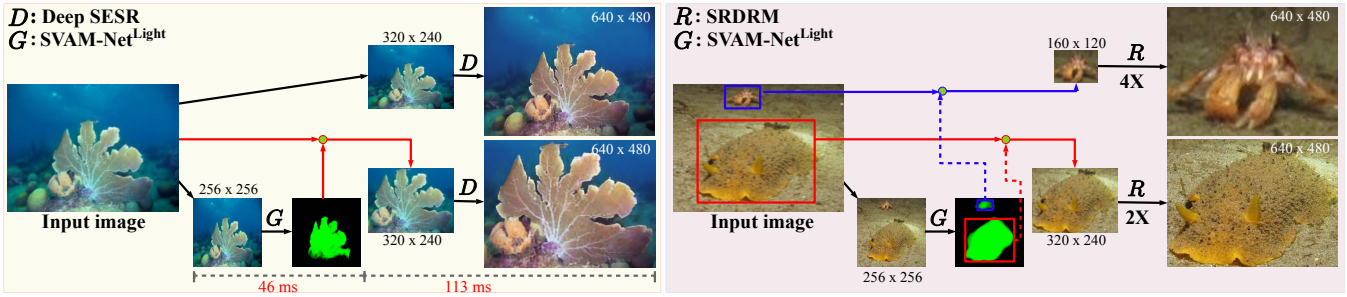
1) Salient RoI Enhancement: AUVs and ROVs operating in noisy visual conditions frequently use various image enhancement models to restore the perceptual image qualities for improved visual perception [54, 51]. However, these models typically have a low-resolution input reception, *e.g.*, 224×224, 256×256, or 320×240. Hence, despite the robustness of SOTA underwater image enhancement models [22, 34], their applicability to high-resolution robotic visual data is limited. For instance, the fastest available model, FUnIE-GAN [22], has an input resolution of 256×256, and it takes 20 ms processing time to generate 256×256 outputs (on AGX Xavier). As a result, it eventually requires 250 ms to enhance and combine all patches of a 1080×768 input image, which is too slow to be useful in near real-time applications.

An effective alternative is to adopt a salient RoI enhancement mechanism to intelligently enhance useful image regions only. As shown in Fig. 8a, SVAM-Net$^{\text{Light}}$-generated saliency maps are used to *pool* salient image RoIs, which are then reshaped to convenient image patches for subsequent enhancement. Although this process requires an additional 46 ms of processing time (by SVAM-Net$^{\text{Light}}$), it is still considerably faster than enhancing the entire image. As demonstrated in Fig. 8a, we can save over 45% processing time even when the salient RoI occupies more than half the input image.

2) Effective Image Super-Resolution: Single image super-resolution (SISR) [20, 41] and simultaneous enhancement and super-resolution (SESR) [21] techniques enable visually-guided robots to *zoom into* interesting image regions for detailed visual perception. Since performing SISR/SESR on



(a) Benefits of salient RoI enhancement are shown for two high-resolution input images. On the left: (i) SVAM-Net$^{\text{Light}}$-generated saliency maps are used for RoI pooling, (ii) the salient RoIs are reshaped based on their area, and then (iii) FUnIE-GAN [22] is applied on all 256×256 patches; the total processing time is 88 ms for a 512×256 RoI (top image) and 131 ms for a 512×512 RoI (bottom image). In comparison, as shown on the right, it takes 250 ms to enhance the entire image at 1024×768 resolution.



(b) Utility of SVAM-Net$^{\text{Light}}$ for effective image super-resolution is illustrated by two examples. As shown on the left, Deep SESR [21] on the salient image RoI is potentially more useful for detailed perception rather than SESR on the entire image. Moreover, as seen on the right, SVAM-Net$^{\text{Light}}$-generated saliency maps can also be used to determine the scale for super-resolution; here, we use $2\times$ and $4\times$ SRDRM [20] on two salient RoIs based on their respective resolutions.

Fig. 8: Demonstrations for two important use cases of fast SOD by SVAM-Net$^{\text{Light}}$: salient RoI enhancement and image super-resolution. The saliency maps are shown as green intensity values; all evaluations are performed on a single-board NVIDIA™ GTX Xavier device.

the entire input image is not computationally feasible, the challenge here is to determine which image regions are salient. As shown in Fig. 8b, SVAM-Net$^{\text{Light}}$ can be used to find the salient image RoIs for effective SISR/SESR. Moreover, the super-resolution scale (*e.g.*, $2\times$, $3\times$, or $4\times$) can be readily determined based on the shape/pixel-area of a salient RoI. Hence, a class-agnostic SOD module is of paramount importance to gain the operational benefits of image super-resolution, especially in vision-based tasks such as tracking/following fast-moving targets [67, 52] and surveying distant coral reefs [44]. For its computational efficiency and robustness, SVAM-Net$^{\text{Light}}$ is an ideal choice to be used alongside a SISR/SESR module in practical applications.

3) Fast Visual Search and Attention Modeling: In §II-B, we discussed various saliency-guided approaches for fast visual search [30, 26] and spatial attention modeling [14]. Robust identification of salient pixels is the most essential first step in these approaches irrespective of the high-level application-specific tasks, *e.g.*, enhanced object detection [67, 50, 47], place recognition [43], coral reef monitoring [44], autonomous exploration [14, 49], etc. SVAM-Net$^{\text{Light}}$ offers a general-purpose solution to this, while ensuring fast inference rates on single-board devices. As shown in Fig. 9, SVAM-Net$^{\text{Light}}$ reliably detects humans, robots, wrecks/ruins, instruments, and other salient objects in a scene. Additionally, it accurately discards all background (waterbody) pixels and focuses on salient foreground pixels only. Such precise segmentation of salient image regions enables fast and effective spatial attention modeling, which is key to the operational success of visually-guided underwater robots.

VI. Concluding Remarks

"Where to look" is a challenging and open problem in underwater robot vision. An essential capability of visually-guided AUVs is to identify interesting and salient objects in the scene to accurately make important operational decisions. In this paper, we present a novel deep visual model named SVAM-Net, which combines the power of bottom-up and top-down SOD learning in a holistic encoder-decoder architecture. We design dedicated spatial attention modules to effectively exploit the coarse-level and fine-level semantic features along the two learning pathways. In particular, we configure the bottom-up pipeline to extract semantically rich hierarchical features from early encoding layers, which facilitates an abstract yet accurate saliency prediction at a fast rate; we denote this decoupled bottom-up pipeline as SVAM-Net$^{\text{Light}}$. On the other hand, we design a residual refinement module that

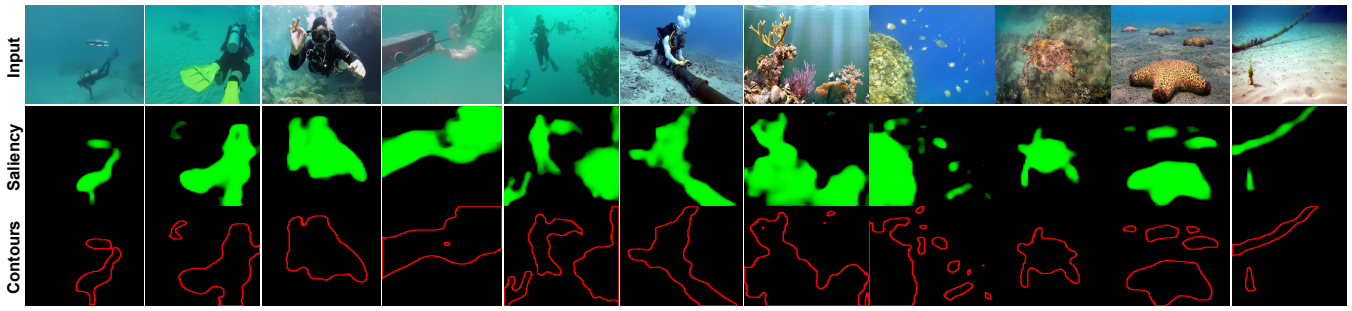


Fig. 9: SVAM-Net$^{\text{Light}}$-generated saliency maps and respective object contours are shown for a variety of snapshots taken during human-robot cooperative experiments and oceanic explorations. Further experimental results demonstrating its generalization performance are provided in the supplementary material; a video demonstration can be seen here: `https://youtu.be/SxJcsoQw7KI`.

ensures fine-grained saliency estimation through the deeper top-down pipeline.

In the implementation, we incorporate comprehensive end-to-end supervision of SVAM-Net by large-scale diverse training data consisting of both terrestrial and underwater imagery. Subsequently, we validate the effectiveness of its learning components and various loss functions by extensive ablation experiments. In addition to using existing datasets, we release a new challenging test set named USOD for the benchmark evaluation of SVAM-Net and other underwater SOD models. By a series of qualitative and quantitative analyses, we show that SVAM-Net provides SOTA performance for SOD on underwater imagery, exhibits significantly better generalization performance on challenging test cases than existing solutions, and achieves fast end-to-end inference on single-board devices. Moreover, we demonstrate that a delicate balance between robust performance and computational efficiency makes SVAM-Net$^{\text{Light}}$ suitable for real-time use by visually-guided underwater robots. In the near future, we plan to optimize the end-to-end SVAM-Net architecture further to achieve a faster runtime. The subsequent pursuit will be to analyze its feasibility in online learning pipelines for task-specific model adaptation.

Appendix A
Dataset and Code Repository Pointers

- The SUIM [19], UFO-120 [21], EUVP [22], and USR-248 [20] datasets: `http://irvlab.cs.umn.edu/resources/`
- The UIEB dataset [34]: `https://li-chongyi.github.io/proj_benchmark.html`
- Other underwater datasets: `https://github.com/xahidbuffon/underwater_datasets`
- BASNet [46] (PyTorch): `https://github.com/NathanUA/BASNet`
- PAGE-Net [62] (Keras): `https://github.com/wenguanwang/PAGE-Net`
- ASNet [59] (TensorFlow): `https://github.com/wenguanwang/ASNet`
- CPD [63] (PyTorch): `https://github.com/wuzhe71/CPD`
- SOD evaluation (Python): `https://github.com/xahidbuffon/SOD-Evaluation-Tool-Python`

Acknowledgment

This work was supported by the National Science Foundation (NSF) grant `IIS-#1637875`. We are grateful to the Bellairs Research Institute (`https:/www.mcgill.ca/bellairs/`) of Barbados for providing us with the facilities for field experiments.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, et al. TensorFlow: A System for Large-scale Machine Learning. In *USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, pages 265–283, 2016.

[2] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned Salient Region Detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1597–1604. IEEE, 2009.

[3] Derya Akkaynak and Tali Treibitz. A Revised Underwater Image Formation Model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6723–6732, 2018.

[4] Iñigo Alonso, Matan Yuval, Gal Eyal, Tali Treibitz, and Ana C Murillo. CoralSeg: Learning Coral Segmentation from Sparse Annotations. *Journal of Field Robotics (JFR)*, 36(8):1456–1477, 2019.

[5] L. Bazzani, H. Larochelle, and L. Torresani. Recurrent Mixture Density Network for Spatiotemporal Visual Attention. In *International Conference on Learning Representations (ICLR)*, 2017.

[6] Oscar Beijbom, Peter J Edmunds, David I Kline, B Greg Mitchell, and David Kriegman. Automated Annotation of Coral Reef Survey Images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1170–1177. IEEE, 2012.

[7] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient Object Detection: A Benchmark. *IEEE Transactions on Image Processing (TIP)*, 24(12):5706–5722, 2015.

[8] Ali Borji, Ming-Ming Cheng, Qibin Hou, Huaizu Jiang,

and Jia Li. Salient Object Detection: A Survey. *Computational Visual Media*, pages 1–34, 2019.

[9] Zhe Chen, Yang Sun, Yupeng Gu, Huibin Wang, Hao Qian, and Hao Zheng. Underwater Object Segmentation Integrating Transmission and Saliency Features. *IEEE Access*, 7:72420–72430, 2019.

[10] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global Contrast-based Salient Region Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 37(3):569–582, 2014.

[11] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. A Tutorial on the Cross-entropy Method. *Annals of Operations Research*, 134(1):19–67, 2005.

[12] Duane R Edgington, Karen A Salamy, Michael Risi, RE Sherlock, Dirk Walther, and Christof Koch. Automated Event Detection in Underwater Video. In *Oceans*, volume 5, pages 2749–2753. IEEE, 2003.

[13] Y. Girdhar, P. Giguere, and G. Dudek. Autonomous Adaptive Exploration using Realtime Online Spatiotemporal Topic Modeling. *International Journal of Robotics Research (IJRR)*, 33(4):645–657, 2014.

[14] Yogesh Girdhar and Gregory Dudek. Modeling Curiosity in a Mobile Robot for Long-term Autonomous Exploration and Monitoring. *Autonomous Robots*, 40(7):1267–1278, 2016.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE, 2016.

[16] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply Supervised Salient Object Detection with Short Connections. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3203–3212. IEEE, 2017.

[17] Ping Hu, Bing Shuai, Jun Liu, and Gang Wang. Deep Level Sets for Salient Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2300–2309. IEEE, 2017.

[18] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *CoRR, abs/1502.03167*, 2015.

[19] Md Jahidul Islam, Chelsey Edge, Yuyang Xiao, Peigen Luo, Muntaqim Mehtaz, Christopher Morse, Sadman Sakib Enan, and Junaed Sattar. Semantic Segmentation of Underwater Imagery: Dataset and Benchmark. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1769–1776, 2020.

[20] Md Jahidul Islam, Sadman Sakib Enan, Peigen Luo, and Junaed Sattar. Underwater Image Super-Resolution using Deep Residual Multipliers. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 900–906, 2020.

[21] Md Jahidul Islam, Peigen Luo, and Junaed Sattar. Simultaneous Enhancement and Super-Resolution of Underwater Imagery for Improved Visual Perception. In *Robotics: Science and Systems (RSS)*, 2020. doi: 10.15607/RSS.2020.XVI.018.

[22] Md Jahidul Islam, Youya Xia, and Junaed Sattar. Fast Underwater Image Enhancement for Improved Visual Perception. *IEEE Robotics and Automation Letters (RA-L)*, 5(2):3227–3234, 2020.

[23] Laurent Itti, Christof Koch, and Ernst Niebur. A Model of Saliency-based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 20(11):1254–1259, 1998.

[24] Muwei Jian, Qiang Qi, Junyu Dong, Yilong Yin, and Kin-Man Lam. Integrating QDWD with Pattern Distinctness and Local Contrast for Underwater Saliency Detection. *Journal of Visual Communication and Image Representation*, 53:31–41, 2018.

[25] Muwei Jian, Qiang Qi, Hui Yu, Junyu Dong, Chaoran Cui, Xiushan Nie, Huaxiang Zhang, Yilong Yin, and Kin-Man Lam. The Extended Marine Underwater Environment Database and Baseline Evaluations. *Applied Soft Computing*, 80:425–437, 2019.

[26] Matthew Johnson-Roberson, Oscar Pizarro, and Stefan Williams. Saliency Ranking for Benthic Survey using Underwater Images. In *International Conference on Control Automation Robotics & Vision*, pages 459–466. IEEE, 2010.

[27] Ayoung Kim and Ryan M Eustice. Real-time Visual SLAM for Autonomous Underwater Hull Inspection using Visual Saliency. *IEEE Transactions on Robotics (TRO)*, 29(3):719–733, 2013.

[28] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. In *International Conference for Learning Representations (ICLR)*, 2015.

[29] Dominik A Klein and Simone Frintrop. Center-surround Divergence of Feature Statistics for Salient Object Detection. In *International Conference on Computer Vision (ICCV)*, pages 2214–2219. IEEE, 2011.

[30] Karim Koreitem, Florian Shkurti, Travis Manderson, Wei-Di Chang, Juan Camilo Gamboa Higuera, and Gregory Dudek. One-Shot Informed Robotic Visual Search in the Wild. *ArXiv preprint arXiv:2003.10010*, 2020.

[31] Srinivas SS Kruthiventi, Kumar Ayush, and R Venkatesh Babu. DeepFix: A Fully Convolutional Neural Network for Predicting Human Eye Fixations. *IEEE Transactions on Image Processing (TIP)*, 26(9):4446–4456, 2017.

[32] Nitin Kumar, Harish Kumar Sardana, and SN Shome. Saliency-based Shape Extraction of Objects in Unconstrained Underwater Environment. *Multimedia Tools and Applications*, 78(11):15121–15139, 2019.

[33] Olivier Le Meur, Patrick Le Callet, Dominique Barba, and Dominique Thoreau. A Coherent Computational Approach to Model Bottom-Up Visual Attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 28(5):802–817, 2006.

[34] C. Li, C. Guo, W. Ren, R. Cong, J. Hou, S. Kwong, and D. Tao. An Underwater Image Enhancement Benchmark

Dataset and Beyond. In *IEEE Transactions on Image Processing (TIP)*, pages 1–1. IEEE, 2019.

[35] Guanbin Li and Yizhou Yu. Deep Contrast Learning for Salient Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 478–487. IEEE, 2016.

[36] Xin Li, Fan Yang, Hong Cheng, Wei Liu, and Dinggang Shen. Contour Knowledge Transfer for Salient Object Detection. In *European Conference on Computer Vision (ECCV)*, pages 355–370. Springer, 2018.

[37] Xiu Li, Jing Hao, Min Shang, and Zhixiong Yang. Saliency Segmentation and Foreground Extraction of Underwater Image based on Localization. In *OCEANS*, pages 1–4. IEEE, 2016.

[38] Nian Liu and Junwei Han. DHSNet: Deep Hierarchical Saliency Network for Salient Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 678–686. IEEE, 2016.

[39] Nian Liu, Junwei Han, and Ming-Hsuan Yang. PiCAet: Learning Pixel-wise Contextual Attention for Saliency Detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3089–3098. IEEE, 2018.

[40] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to Detect a Salient Object. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(2):353–367, 2010.

[41] H. Lu, Y. Li, S. Nakashima, H. Kim, and S. Serikawa. Underwater Image Super-resolution by Descattering and Fusion. *IEEE Access*, 5:670–679, 2017.

[42] Alejandro Maldonado-Ramírez and L Abril Torres-Méndez. Robotic Visual Tracking of Relevant Cues in Underwater Environments with Poor Visibility Conditions. *Journal of Sensors*, 2016(1), 2016.

[43] Alejandro Maldonado-Ramírez and L Abril Torres-Mendez. Learning Ad-hoc Compact Representations from Salient Landmarks for Visual Place Recognition in Underwater Environments. In *International Conference on Robotics and Automation (ICRA)*, pages 5739–5745. IEEE, 2019.

[44] Md Modasshir and Ioannis Rekleitis. Enhancing Coral Reef Monitoring Utilizing a Deep Semi-Supervised Learning Approach. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1874–1880. IEEE, 2020.

[45] Vinod Nair and Geoffrey E Hinton. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proc. of the International Conference on Machine Learning (ICML)*, pages 807–814, 2010.

[46] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. BASNet: Boundary-aware Salient Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7479–7489. IEEE, 2019.

[47] Mehdi Ravanbakhsh, Mark R Shortis, Faisal Shafait, Ajmal Mian, Euan S Harvey, and James W Seager. Automated Fish Detection in Underwater Images Using Shape-Based Level Sets. *Photogrammetric Record*, 30 (149):46–62, 2015.

[48] Zeeshan Ur Rehman. *Salient Object Detection from Underwater Image*. PhD thesis, CAPITAL UNIVERSITY, 2019.

[49] Ioannis Rekleitis, Gregory Dudek, and Evangelos Milios. Multi-robot Collaboration for Robust Exploration. *Annals of Mathematics and Artificial Intelligence*, 31(1-4):7–40, 2001.

[50] Dario Lodi Rizzini, Fabjan Kallasi, Fabio Oleari, and Stefano Caselli. Investigation of Vision-based Underwater Object Detection with Multiple Datasets. *International Journal of Advanced Robotic Systems*, 12(6):77, 2015.

[51] M. Roznere and A. Q. Li. Real-time Model-based Image Color Correction for Underwater Robots. *arXiv preprint arXiv:1904.06437*, 2019.

[52] Florian Shkurti, Wei-Di Chang, Peter Henderson, Md Jahidul Islam, Juan Camilo Gamboa Higuera, Jimmy Li, Travis Manderson, Anqi Xu, Gregory Dudek, and Junaed Sattar. Underwater Multi-Robot Convoying using Visual Tracking by Detection. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4189–4196, 2017.

[53] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-scale Image Recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[54] Luz A Torres-Méndez and Gregory Dudek. Color Correction of Underwater Images for Aquatic Robot Inspection. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 60–73. Springer, 2005.

[55] Hui Bin Wang, Xin Dong, Jie Shen, Xue Wen Wu, and Zhe Chen. Saliency-based Adaptive Object Extraction for Color Underwater Images. In *Applied Mechanics and Materials*, volume 347, pages 3964–3970. Trans Tech Publ, 2013.

[56] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to Detect Salient Objects with Image-level Supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 136–145. IEEE, 2017.

[57] Linzhao Wang, Lijun Wang, Huchuan Lu, Pingping Zhang, and Xiang Ruan. Saliency Detection with Recurrent Fully Convolutional Networks. In *European Conference on Computer Vision (ECCV)*, pages 825–841. Springer, 2016.

[58] Tiantian Wang, Lihe Zhang, Shuo Wang, Huchuan Lu, Gang Yang, Xiang Ruan, and Ali Borji. Detect Globally, Refine Locally: A Novel Approach to Saliency Detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3127–3135. IEEE, 2018.

[59] Wenguan Wang, Jianbing Shen, Xingping Dong, and Ali Borji. Salient Object Detection Driven by Fixation Prediction. In *IEEE Conference on Computer Vision and*

Pattern Recognition (CVPR), pages 1711–1720, 2018.

[60] Wenguan Wang, Qiuxia Lai, Huazhu Fu, Jianbing Shen, Haibin Ling, and Ruigang Yang. Salient Object Detection in the Deep Learning Era: An In-Depth Survey. *arXiv preprint arXiv:1904.09146*, 2019.

[61] Wenguan Wang, Jianbing Shen, Ming-Ming Cheng, and Ling Shao. An Iterative and Cooperative Top-down and Bottom-up Inference Network for Salient Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5968–5977, 2019.

[62] Wenguan Wang, Shuyang Zhao, Jianbing Shen, Steven CH Hoi, and Ali Borji. Salient Object Detection with Pyramid Attention and Salient Edges. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1448–1457. IEEE, 2019.

[63] Zhe Wu, Li Su, and Qingming Huang. Cascaded Partial Decoder for Fast and Accurate Salient Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3907–3916. IEEE, 2019.

[64] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency Detection via Graph-based Manifold Ranking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3166–3173. IEEE, 2013.

[65] Rui Zhao, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Saliency Detection by Multi-Context Deep Learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1265–1274. IEEE, 2015.

[66] Ting Zhao and Xiangqian Wu. Pyramid Feature Attention Network for Saliency Detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3085–3094. IEEE, 2019.

[67] Jianjiang Zhu, Siquan Yu, Lei Gao, Zhi Han, and Yandong Tang. Saliency-Based Diver Target Detection and Localization Method. *Mathematical Problems in Engineering*, 2020.

[68] Yue Zhu, Baochen Hao, Baohua Jiang, Rui Nian, Bo He, Xinmin Ren, and Amaury Lendasse. Underwater Image Segmentation with Co-saliency Detection and Local Statistical Active Contour Model. In *OCEANS*. IEEE, 2017.