Supplementary Materials for **Paper ID: 52**

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

1 Evaluation Data Preparation

We conduct benchmark evaluation on the test sets of three publicly available datasets: SUIM [5], UFO-120 [7], and MUED [9]. SVAM-Net is jointly supervised on 3025 training instances of SUIM and UFO-120; their test sets contain an additional 110 and 120 instances, respectively. These datasets contain a diverse collection of natural underwater images with important object categories such as fish, coral reefs, humans, robots, wrecks/ruins, etc. Besides, MUED dataset contains 8600 images in 430 groups of conspicuous objects; although it includes a wide variety of complex backgrounds, the images lack diversity in terms of object categories and water-body types. Moreover, MUED provides bounding-box annotations only. Hence, to maintain consistency in our quantitative evaluation, we select 300 diverse groups and perform pixel-level annotations on those images.

Figure 1: There are 300 test images in the proposed USOD dataset (resolution: 640×480); a few sample images and their ground truth saliency maps are shown on the top and bottom row, respectively. The dataset can be downloaded from here: `http://irvlab.cs.umn.edu/resources/usod-dataset`.

In addition to the existing datasets, 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. We collect these images from two major sources:

- **Existing unlabeled datasets**: we utilize benchmark datasets that are generally used for underwater image enhancement and super-resolution tasks; specifically, we select subsets of images from datasets named USR-248 [6], UIEB [10], and EUVP [8].
- **Field trials**: we have collected data from several oceanic trials and field explorations in multiple open water sites. The selected images include diverse underwater scenes and various setups for human-robot cooperative experiments.

Once the images are compiled, four human participants independently annotated the salient pixels to generate ground truth labels; a few samples are provided in Fig. 1.

2 SOD Performance Metrics

We evaluate the performance of SVAM-Net and other existing SOD methods based on four widely-used evaluation criteria [2, 4, 11, 13]:

- **F-measure** ($\mathbf{F}_\beta$) is an overall performance measurement that is computed by the weighted harmonic mean of the precision and recall as:

$$\mathbf{F}_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall}. \tag{1}$$

  Here, β^2 is set to 0.3 as per the SOD literature to weight precision more than recall. Also, the maximum scores ($\mathbf{F}_\beta^{max}$) are reported for quantitative comparison.

- **S-measure** ($\mathbf{S}_m$) is a recently proposed metric [3] that simultaneously evaluates region-aware ($\mathbf{S}_o$) and object-aware ($\mathbf{S}_r$) structural similarities between the predicted and ground truth saliency maps as follows:

$$\mathbf{S}_m = \alpha \times \mathbf{S}_o + (1 - \alpha) \times \mathbf{S}_r. \tag{2}$$

  Here, we set $\alpha = 0.5$ as suggested in [3].

- **Mean absolute error (MAE)** is a generic metric that measures the average pixel-wise differences between the predicted ($\hat{s}^{m \times n}$) and ground truth ($s^{m \times n}$) saliency maps as follows:

$$MAE(\mathbf{s}, \hat{\mathbf{s}}) = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} |\mathbf{s}_{i,j} - \hat{\mathbf{s}}_{i,j}|. \tag{3}$$

- **Precision-recall (PR) curve** is a standard performance metric and is complementary to MAE. It is evaluated by *binarizing* the predicted saliency maps with a threshold sliding from 0 to 255 and then performing bin-wise comparison with the ground truth values. Then, the relationship between *precision* and *recall* is plotted for every possible cut-off values.

3 Generalization Performance of SVAM

Underwater imagery suffers from a wide range of non-linear distortions caused by the waterbody-specific properties of light propagation [1, 8]. The image quality and statistics also vary depending on visibility conditions, background patterns, and the presence of artificial light sources and unknown objects in a scene. Consequently, learning-based SOD solutions oftentimes fail to generalize beyond supervised data. To address this issue, SVAM-Net adopts a two-step training pipeline that includes supervision by (i) a large collection of samples with diverse scenes and object categories to learn a generalizable SOD function, and (ii) a wide variety of natural underwater images to learn to capture the inherent optical distortions. In Fig. 2 through Fig. 5, we demonstrate the robustness of SVAM-Net with a series of challenging test cases.

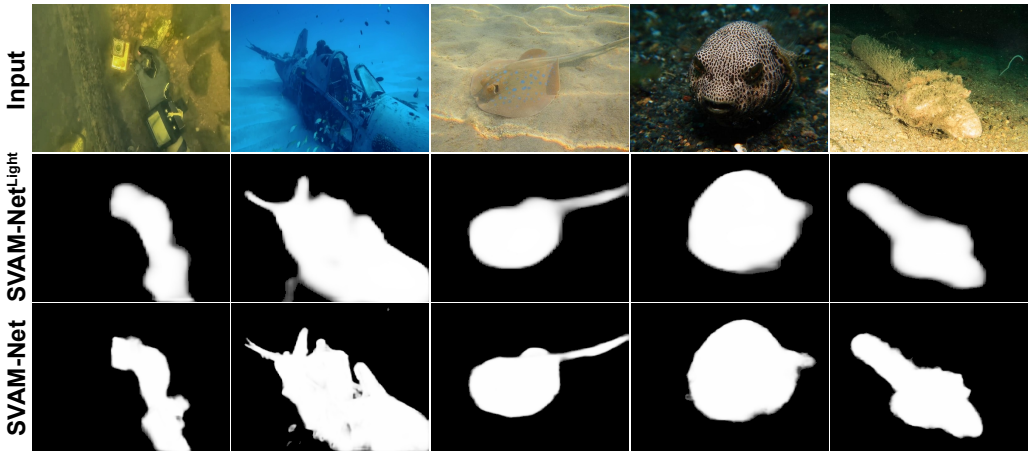

Figure 2: SVAM-Net performance for challenging test cases: with low contrast and/or color distortions.

As shown in Fig. 2, underwater images tend to have a dominating green or blue hue because red wavelengths get absorbed in deep water (as light travels further). Such wavelength dependent attenuation, scattering, and other optical properties of the waterbodies cause irregular and non-linear distortions which result in low-contrast, often blurred, and color-degraded images. We notice that both SVAM-Net and SVAM-Net$^{\text{Light}}$ can overcome the noise and image distortions and successfully localize the salient objects. They are also robust to other pervasive issues such as occlusion and cluttered backgrounds with confusing textures. As Fig. 3 demonstrates, the salient objects are mostly well-segmented from the confusing background pixels having similar colors and textures. Here, we observe that although SVAM-Net$^{\text{Light}}$ introduces a few false-positive pixels, SVAM-Net's predictions are rather accurate and fine-grained.

Another important feature of general-purpose SOD models is the ability to identify novel salient objects, particularly with complicated shapes. As shown in Fig. 4, objects such as wrecked/submerged cars, planes, statues, and cages are accurately segmented by both SVAM-Net and SVAM-Net$^{\text{Light}}$. Their SOD performance is also invariant to the scale and orientation of salient objects. We postulate that the large-scale supervised pre-training step contributes to this robustness as the terrestrial datasets include a variety of object categories. In fact, we

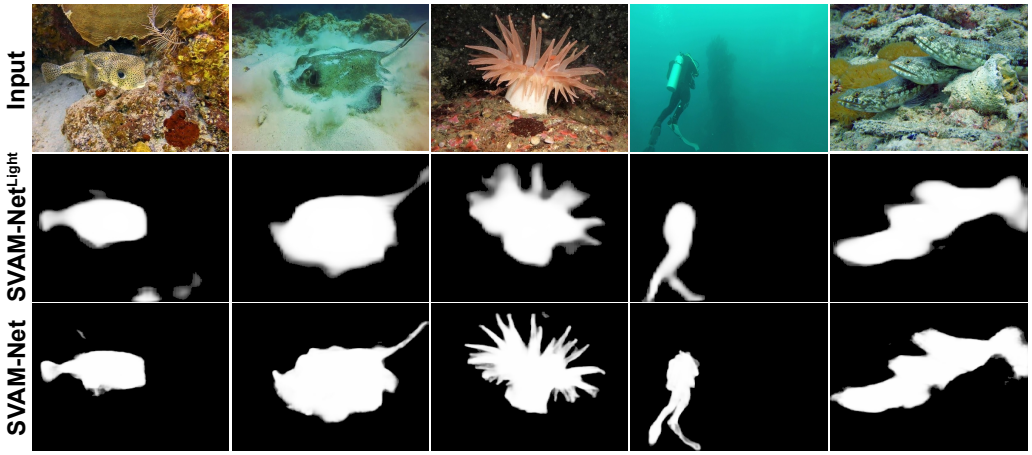

Figure 3: SVAM-Net performance for challenging test cases: with cluttered background and/or confusing textures.

Figure 4: SVAM-Net performance for challenging test cases: with unseen objects/shapes and variations in scale.

find that they also perform reasonably well on arbitrary terrestrial images [12]; a few results are shown in Fig. 5. This suggests that with domain-specific end-to-end training, SVAM-Net could be effectively used in terrestrial applications as well.

References

[1] 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.

[2] Ali Borji, Ming-Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li. Salient Object Detection: A Survey. *Computational Visual Media*, pages 1–34, 2019.

[3] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A New Way to Evaluate Foreground Maps. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4548–4557, 2017.

[4] Mengyang Feng, Huchuan Lu, and Errui Ding. Attentive Feedback Network for Boundary-aware Salient Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1623–1632. IEEE, 2019.

[5] 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.

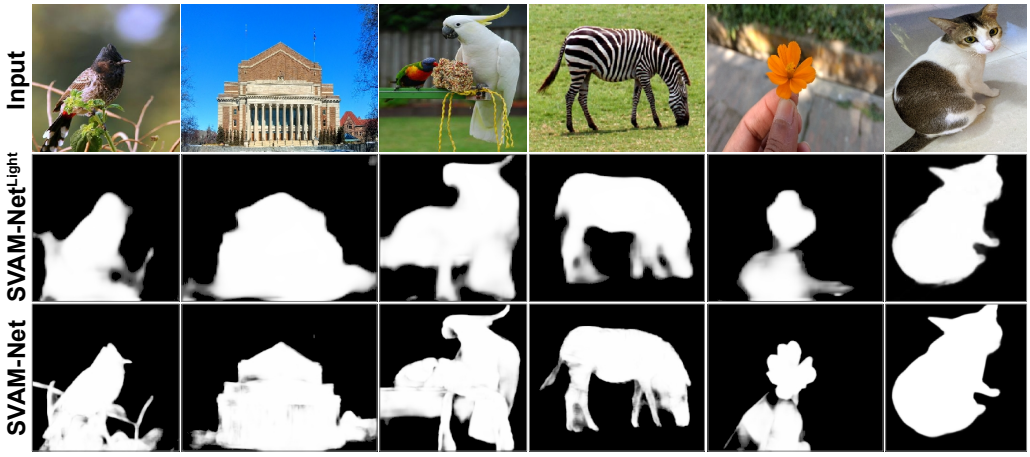

Figure 5: SVAM-Net performance for challenging test cases: on unseen terrestrial images with arbitrary objects.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.