2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) | 978-1-7281-6212-6/20/$31.00 ©2020 IEEE | DOI: 10.1109/IROS45743.2020.9340821

# Semantic Segmentation of Underwater Imagery: Dataset and Benchmark

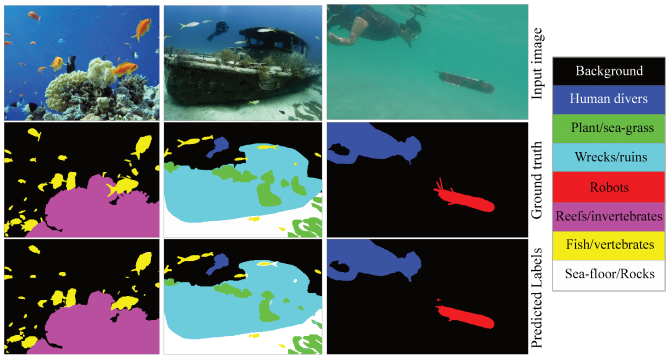
Md Jahidul Islam[1], Chelsey Edge[2], Yuyang Xiao[3], Peigen Luo[4], Muntaqim Mehtaz[5], Christopher Morse[6], Sadman Sakib Enan[7] and Junaed Sattar[8]

***Abstract*— In this paper, we present the first large-scale dataset for semantic Segmentation of Underwater IMagery (SUIM). It contains over 1500 images with pixel annotations for eight object categories: fish (vertebrates), reefs (invertebrates), aquatic plants, wrecks/ruins, human divers, robots, and sea-floor. The images have been rigorously collected during oceanic explorations and human-robot collaborative experiments, and annotated by human participants. We also present a comprehensive benchmark evaluation of several state-of-the-art semantic segmentation approaches based on standard performance metrics. Additionally, we present SUIM-Net, a fully-convolutional deep residual model that balances the trade-off between performance and computational efficiency. It offers competitive performance while ensuring fast end-to-end inference, which is essential for its use in the autonomy pipeline by visually-guided underwater robots. In particular, we demonstrate its usability benefits for visual servoing, saliency prediction, and detailed scene understanding. With a variety of use cases, the proposed model and benchmark dataset open up promising opportunities for future research in underwater robot vision.**

## I. Introduction

Semantic segmentation is a well-studied problem in the domains of robot vision and deep learning [1], [2], [3] for its usefulness in estimating scene geometry, inferring interactions and spatial relationships among objects, salient object identification, and more. It is particularly important for detailed scene understanding in autonomous driving by visually-guided robots. Over the last decade, substantial contributions from both industrial and academic researchers have led to remarkable advancements of the state-of-the-art (SOTA) methodologies for semantic segmentation [1], [4]. This success is largely propelled by various genres of deep convolutional neural network (CNN)-based models that learn from large collections of annotated data [5], [6]. Several large-scale benchmark datasets of terrestrial imagery [7], [8] and videos [9] provide a standard platform for such research and fuel the rapid development of the relevant literature. To date, the SOTA semantic segmentation models are core elements in the visual perception pipelines of most terrestrial robots and systems.

For visually-guided underwater robots, however, the existing solutions for semantic segmentation and scene parsing are significantly less advanced. The practicalities and limitations are twofold. First, the visual content of underwater imagery is entirely different because of the domain-specific object categories, background patterns, and optical distortion artifacts [10]; hence, the learning-based SOTA models trained on terrestrial data are not directly applicable. Secondly, there are no underwater datasets to facilitate large-scale training and benchmark evaluation of semantic segmentation models for general-purpose use. The existing large-scale annotated data and relevant methodologies are tied to specific applications such as coral-reef classification and coverage estimation [11], [12], [13], fish detection and segmentation [14], [15], etc. Other datasets contain either binary annotations for salient foreground pixels [16] or semantic labels for very few object categories (*e.g.*, sea-grass, rocks/sand, etc.) [17]. Therefore, the large-scale learning-based semantic segmentation methodologies for underwater imagery are not explored in depth in the literature. Besides,




(a) Three instances of semantic segmentation by SUIM-Net and respective ground truth labels are shown; the object categories and color codes are provided on the right.

(b) Pixel-level detection for specific object categories: humans, robots, fish, and wrecks/ruins are shown (the segmentation masks are overlayed on the bottom row images); such fine-grained object localization is useful for visual attention modeling and servoing.

Fig. 1: Demonstrations for semantic segmentation of underwater scenes and other use cases of the proposed SUIM dataset and SUIM-Net model.

The authors are with the Interactive Robotics and Vision Laboratory (IRVLab), Department of Computer Science and Engineering (CSE), Minnesota Robotics Institute (MnRI), University of Minnesota, Twin Cities, US. {[1]islam034, [2]edge0037, [3]xiao0153, [4]luo00034, [5]mehta216, [6]morse164, [7]enan0001, [8]junaed}@umn.edu

the traditional class-agnostic approaches are only suitable for simple tasks such as foreground segmentation [18], [19], obstacle detection [20], saliency prediction [21], etc.; they are not generalizable for multi-object semantic segmentation.

We attempt to address these limitations by presenting a large-scale annotated dataset for semantic segmentation of underwater scenes in general-purpose robotic applications. As shown in Figure 1, the proposed SUIM dataset considers object categories for fish, reefs, aquatic plants, and wrecks/ruins, which are of primary interest in many underwater exploration and surveying applications [22], [23], [24]. Additionally, it contains pixel annotations for human divers, robots/instruments, and sea-floor/rocks; these are major objects of interest in human-robot cooperative applications [25], [26]. The SUIM dataset contains 1525 natural underwater images and their ground truth semantic labels; it also includes a test set of 110 images. The dataset and relevant resources are available at https://irvlab.cs.umn.edu/resources/suim-dataset.

Moreover, we conduct a thorough experimental evaluation of several SOTA semantic segmentation models and present their performance benchmark on the SUIM dataset. We also design a computationally efficient novel model named SUIM-Net based on fully-convolutional deep residual learning [27]. In addition to presenting the conceptual model and detailed network architecture of SUIM-Net, we analyze its performance in quantitative and qualitative terms based on standard metrics. SUIM-Net offers significantly faster run-time than the SOTA models while achieving competitive semantic segmentation performance. Furthermore, we demonstrate various use cases of SUIM-NET and specify corresponding training configurations for the SUIM dataset. The model and associated training pipelines are released for academic research at http://irvlab.cs.umn.edu/image-segmentation/suim-and-suim-net.

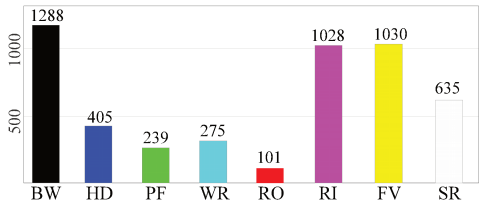
## II. The SUIM Dataset

We consider the following object categories for semantic labeling in the SUIM dataset: *a)* waterbody background (BW), *b)* human divers (HD), *c)* aquatic plants/flora (PF), *d)* wrecks/ruins (WR), *e)* robots and instruments (RO), *f)* reefs and other invertebrates (RI), *g)* fish and other vertebrates (FV), and *h)* sea-floor and rocks (SR). As depicted in Table I, we use 3-bit binary RGB colors to represent these eight object categories in the image space.

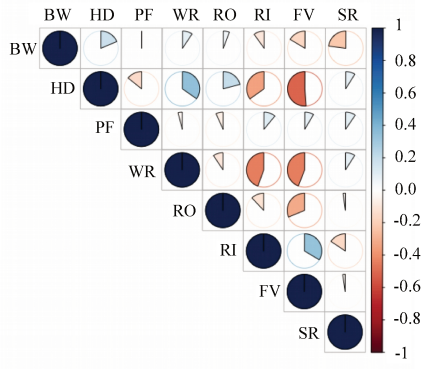
TABLE I: The object categories and corresponding color codes for pixel annotations in the SUIM dataset.

| Object category | RGB color | Code |
|---|---|---|
| Background (waterbody) | 000 | BW |
| Human divers | 001 | HD |
| Aquatic plants and sea-grass | 010 | PF |
| Wrecks or ruins | 011 | WR |
| Robots (AUVs/ROVs/instruments) | 100 | RO |
| Reefs and invertebrates | 101 | RI |
| Fish and vertebrates | 110 | FV |
| Sea-floor and rocks | 111 | SR |

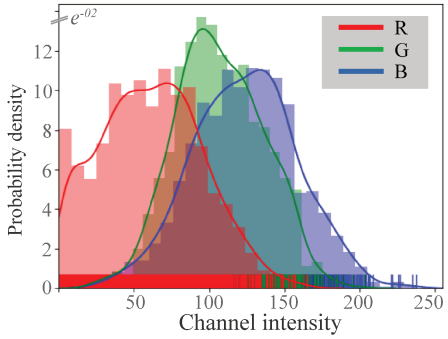
The SUIM dataset has 1525 RGB images for training and validation; another 110 test images are provided for benchmark evaluation of semantic segmentation models. The images are of various spatial resolutions, *e.g.*, $1906 \times 1080$, $1280 \times 720$, $640 \times 480$, and $256 \times 256$. These images are carefully chosen from a large pool of samples collected during oceanic explorations and human-robot cooperative experiments in several locations of various water types. We also utilize a few images from large-scale datasets named EUVP [10], USR-248 [28], and UFO-120 [16], which we previously proposed for underwater image enhancement and super-resolution problems. The images are chosen to accommodate a diverse set of natural underwater scenes and various setups for human-robot collaborative experiments. Figure 2 demonstrates the population of each object category, their pairwise correlations, and the distributions of RGB channel intensity values in the SUIM dataset.




(a) The number of images containing each object category.




(b) Pairwise correlations of the object categories' occurrences.




(c) Distributions of averaged pixel intensity values.

Fig. 2: Statistics concerning various object categories and image intensity values in the SUIM dataset.

All images of the SUIM dataset are pixel-annotated by seven human participants; a few samples are shown in Fig-

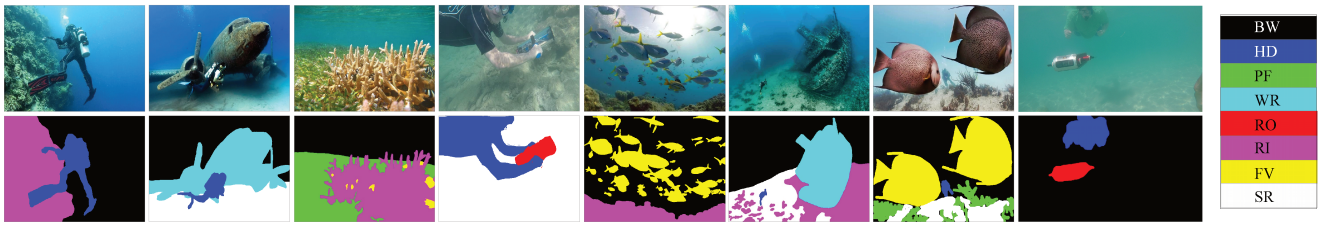



Fig. 3: A few sample images and corresponding pixel-annotations are shown on the top, and bottom row, respectively.

ure 3. We followed the guidelines discussed in [29] and [30] for classifying potentially confusing objects of interest such as plants/reefs, vertebrates/invertebrates, etc.

## III. Use Cases and Related Work

### A. Semantic Segmentation

The learning-based semantic segmentation methodologies have made remarkable progress over the last decade due to the advent of powerful deep models [1] and large-scale annotated datasets (of mostly terrestrial images [7], [8] and videos [9]). The forerunner approaches are based on fully convolutional networks (FCNs) [4] that utilize the seminal CNN-based models (*e.g.*, VGG [31], GoogLeNet [32], ResNet [27]) for hierarchical feature extraction. The *encoded* semantic information is then exploited by a *decoder* network that learns to classify each pixel; it gradually up-samples the low-dimensional features by a series of de-convolution layers [33] and eventually generates the pixel-wise labels. More effective learning pipelines are later proposed for integrating global contextual information and instance awareness. For instance, SegNet architectures [3] accommodate the mapping of deep encoder layers' output features into input dimensions rather than performing ad hoc up-sampling. Moreover, UNet architectures [5] reuse each encoder layers' output by skip-connections to mirrored decoder layers, which significantly improves performance. Several other contemporary approaches incorporate capabilities such as global feature fusion [34], [35], spatio-temporal learning [36], multi-scale learning [37], etc. The notion of *atrous convolution* [2] (aka dilated convolution [38]) and conditional random field (CRF)-based post-processing stages (introduced in the DeepLab models [2], [6]) further ensure multi-scale context awareness and fine-grained localization of object boundaries.

Despite the advancements, semantic segmentation of underwater imagery is considerably less studied. Moreover, the existing solutions for terrestrial imagery are not directly applicable because the object categories and image statistics are entirely different. A unique set of underwater image distortion artifacts and the unavailability of large-scale annotated datasets further influence a significant lack of research attempts. Several important contributions address the problems of coral-reef classification and segmentation [12], [11], coral-reefs' coverage estimation [11], fish detection and segmentation [13], [14], [15], etc. However, these application-specific models are not feasible for general use in robotic applications. Other classical approaches use fuzzy C-means clustering and stochastic optimization methods for foreground segmentation [18], [20], [19]. Such class-agnostic approaches group image regions by evaluating local salient features [21], [39]; hence, these cannot be generalized for multi-object semantic segmentation.

Nevertheless, a few recent work [17], [16] explored the performance of contemporary deep CNN-based semantic segmentation models such as VGG-based encoder-decoders [31], [1], UNet [5], and SegNet [3] for underwater imagery. Although they report inspiring results, they only consider sea-grass, sand, and rock as object categories. Moreover, performance evaluations of many important SOTA models are still unexplored. We attempt to address these limitations by considering a more comprehensive set of object categories in the SUIM dataset, and provide a thorough benchmark evaluation of the SOTA semantic segmentation models (see Section V).

### B. Visual Attention Modeling and Servoing

'Where to look'- is an important open problem for autonomous underwater exploration and surverying [23], [24]. In particular, the most essential capability of visually-guided AUVs is to analyze image-based features for modeling relative attention [22], [40] of various regions of interest (RoIs). Such visual attention-based cues are eventually exploited to make important navigational and other operational decisions. The classical approaches utilize features such as luminance, color, texture, and often depth information to extract salient features for enhanced object detection or template identification [41], [42], [43]. In recent years, the standard one-shot object detection models based on large-scale supervised learning are effectively applied for vision-based tracking and following [44], [25]. These general-purpose object detectors are then coupled with application-specific bounding box (BBox)-reactive controllers for visual servoing [44], [45]. Nevertheless, semantic segmentation provides pixel-level detection accuracy and tighter object boundaries than a BBox (see Figure 4), which are useful for more robust tracking.

Another genre of approaches for visual attention modeling focuses on spatial *saliency* computation, *i.e.*, predicting relevance/importance of each pixel in the image [16], [46]. The salient image regions can be exploited for the detection and tracking of specific (known) objects, or for finding new objects of interest in exploration tasks [22]. For human-robot collaborative applications, in particular, an AUV needs to keep its companion humans/robots within the field-of-view, and pay attention to other objects in the scene, *e.g.*, fish,

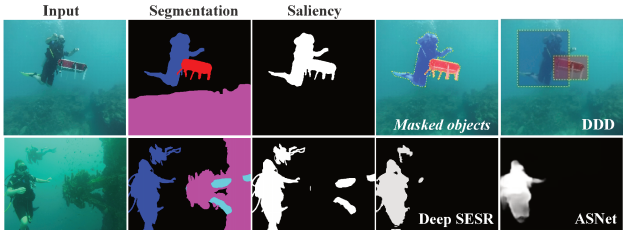



Fig. 4: Generation of semantic saliency maps from segmentation masks: the intensities of HD, RO, FV, and WR pixels are set to 1.0, and the rest are set to 0.0. In comparison: output of a CNN-based object detector named DDD [45] (top); and two class-agnostic saliency predictors named Deep SESR [16] and ASNet [46] (bottom).

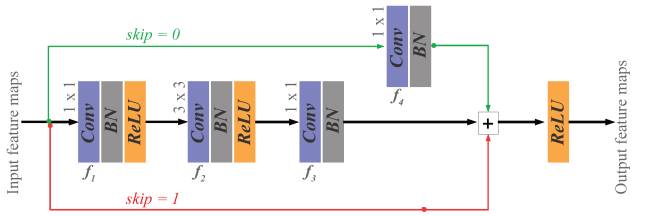
reefs, wrecks, etc. A particular instance of such semantic saliency computation is demonstrated in Figure 4. While a class-agnostic saliency map provides interesting foreground regions, the semantic saliency map embeds additional information about the spatial distribution and interaction among the objects in the scene. Moreover, this semantic knowledge is potentially useful for learning spatio-temporal attention modeling and visual question answering [47], [48], *i.e.*, finding image regions that are relevant to a given query. These exciting research problems have not been explored in depth for underwater robotic applications.

## IV. The SUIM-Net Model

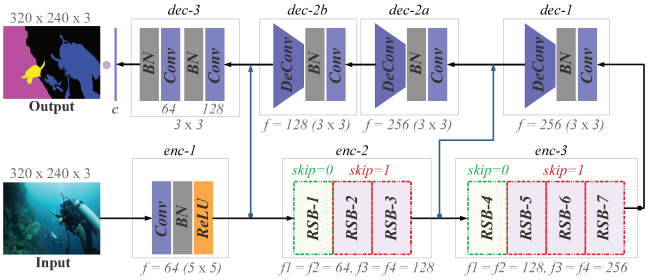
### A. Network Architecture

The proposed SUIM-Net model incorporates a fully convolutional encoder-decoder architecture with skip connections between three mirrored composite layers. As shown in Figure 5a, it embodies residual learning [27] with an optional layered skip-connection at its core building block which we refer to as RSB (residual skip block). Each RSB consists of three sets of convolutional (conv) layers, each followed by Batch Normalization (BN) [49] and ReLU non-linearity [50]. As shown in Figure 5b, two sets of RSBs are used sequentially in the second and third encoder layers; the number of filters, feature dimensions, and other parameters are marked in the figure. The encoder network extracts 256 feature maps from input RGB images (of dimension $320 \times 240 \times 3$); the encoded feature maps are then exploited by three sequential decoder layers. Each decoder layer consists of a conv layer that receives skip-connections from their respective conjugate encoder layer; it is followed by BN and a de-convolutional (deconv) layer [33] for spatial up-sampling. The final conv layer subsequently generates the per-channel binary pixel labels for each object category, which can be post-processed for visualizing in the RGB space.

As evident from Figure 5, we attempt to combine the benefits of skip-connections [5] and residual learning [27], [3] into a fully convolutional model for fast inference. Instead of attempting to surpass the SOTA performance with denser architecture, our motive for such design is to achieve real-time inference while ensuring competitive semantic segmentation performance. We present the performance analysis of SUIM-Net and other SOTA models in the following sections.




(a) Architecture of an RSB: the skip-connection can be either fed from an intermediate conv layer (by setting *skip=0*) or from the input (by setting *skip=1*) for local residual learning.




(b) The end-to-end network architecture: three layers of encoding is performed by a total of seven RSBs, followed by three layers of decoding with mirrored skip-connections.

Fig. 5: Detailed specification of the SUIM-Net model.

### B. Training Pipeline and Implementation Details

We formulate the problem as learning a mapping from an input domain $X$ (of natural underwater images) to its target semantic labeling $Y$ in RGB space. We consider eight object categories of the SUIM dataset (see Table I) and its paired annotated data to learn the underlying function $G : X \rightarrow Y$. The end-to-end training is supervised by the standard cross-entropy loss [51], [3] which evaluates the discrepancy between the predicted and ground truth pixel labels. We use TensorFlow libraries [52] to implement the optimization pipeline; a Linux host with one Nvidia™ GTX 1080 graphics card is used for training. Adam optimizer [53] is used for the global iterative learning with a rate of $10^{-4}$ and a momentum of 0.5. Moreover, we apply several standard image transformations for data augmentation during training; the specifications are provided in Appendix I.

TABLE II: The input dimensions and training parameters for all the models in comparison. [$e$: number of epochs; $s$: steps per epoch; $b$: batch size]

| **Model** | Input dimension | $e \odot s$ | $b$ |
|---|---|---|---|
| FCN8$_{CNN}$ | $320 \times 240 \times 3$ | $50 \odot 4000$ | 2 |
| FCN8$_{VGG}$ | $320 \times 240 \times 3$ | $40 \odot 4000$ | 2 |
| SegNet$_{CNN}$ | $320 \times 256 \times 3$ | $40 \odot 4000$ | 8 |
| SegNet$_{ResNet}$ | $320 \times 256 \times 3$ | $50 \odot 5000$ | 4 |
| UNet$_{GREY}$ | $320 \times 240 \times 1$ | $20 \odot 4000$ | 4 |
| UNet$_{RGB}$ | $320 \times 240 \times 3$ | $30 \odot 5000$ | 2 |
| PSPNet$_{MobileNet}$ | $384 \times 384 \times 3$ | $60 \odot 5000$ | 2 |
| DeepLab$_{V3}$ | $320 \times 320 \times 3$ | $50 \odot 4000$ | 2 |
| SUIM-Net | $320 \times 240 \times 3$ | $45 \odot 5000$ | 4 |

TABLE III: Quantitative performance comparison for semantic segmentation and saliency prediction: scores are shown as *mean* $\pm$ $\sqrt{variance}$; the best score and next top three scores for each comparison are colored red and blue, respectively.

| | Model | HD | WR | RO | RI | FV | Combined | Saliency Pred. |
|---|---|---|---|---|---|---|---|---|
| $\mathcal{F}$ (↑) | FCN8$_{CNN}$ | 76.34 ± 2.24 | 70.24 ± 2.26 | 39.83 ± 3.87 | 61.65 ± 2.36 | 76.24 ± 1.87 | 64.86 ± 2.52 | 75.62 ± 1.79 |
| | FCN8$_{VGG}$ | 89.10 ± 1.50 | 82.03 ± 1.94 | 74.01 ± 3.23 | 79.19 ± 2.27 | 90.46 ± 1.18 | 82.96 ± 2.02 | 89.63 ± 1.24 |
| | SegNet$_{CNN}$ | 59.60 ± 2.02 | 41.60 ± 1.65 | 31.77 ± 3.03 | 41.88 ± 2.66 | 60.08 ± 1.91 | 46.97 ± 2.25 | 56.96 ± 1.58 |
| | SegNet$_{ResNet}$ | 80.52 ± 3.26 | 77.65 ± 3.15 | 62.45 ± 3.90 | 82.30 ± 1.96 | 91.47 ± 1.01 | 76.88 ± 2.66 | 86.88 ± 1.83 |
| | UNet$_{GRAY}$ | 85.47 ± 2.21 | 79.77 ± 2.01 | 60.95 ± 3.31 | 69.95 ± 2.57 | 84.47 ± 1.39 | 75.12 ± 2.30 | 83.96 ± 1.40 |
| | UNet$_{RGB}$ | 89.60 ± 1.84 | 86.17 ± 1.73 | 68.87 ± 3.30 | 79.24 ± 2.70 | 91.35 ± 1.14 | 83.05 ± 2.14 | 89.99 ± 1.29 |
| | PSPNet$_{MobileNet}$ | 80.21 ± 1.19 | 70.94 ± 1.61 | 72.04 ± 2.21 | 72.65 ± 1.62 | 79.19 ± 1.74 | 76.01 ± 1.67 | 78.42 ± 1.59 |
| | DeepLab$_{V3}$ | 89.68 ± 2.09 | 77.73 ± 2.18 | 72.72 ± 3.35 | 78.28 ± 2.70 | 87.95 ± 1.59 | 81.27 ± 2.30 | 85.94 ± 1.72 |
| | **SUIM-Net** | 89.04 ± 1.31 | 65.37 ± 2.22 | 74.18 ± 2.11 | 71.92 ± 1.80 | 84.36 ± 1.37 | 78.86 ± 1.79 | 81.36 ± 1.72 |
| *mIOU* (↑) | FCN8$_{CNN}$ | 67.27 ± 2.50 | 81.64 ± 2.16 | 36.44 ± 3.67 | 78.72 ± 2.50 | 70.25 ± 2.28 | 66.86 ± 2.62 | 75.63 ± 1.89 |
| | FCN8$_{VGG}$ | 79.86 ± 1.50 | 85.77 ± 2.09 | 65.05 ± 3.00 | 85.23 ± 2.07 | 81.18 ± 1.46 | 79.42 ± 2.02 | 85.22 ± 1.24 |
| | SegNet$_{CNN}$ | 62.76 ± 2.35 | 66.75 ± 2.57 | 36.63 ± 3.12 | 63.46 ± 3.18 | 62.48 ± 2.32 | 58.42 ± 2.71 | 65.90 ± 2.12 |
| | SegNet$_{ResNet}$ | 74.00 ± 2.88 | 82.68 ± 2.94 | 58.63 ± 3.61 | 89.61 ± 1.15 | 82.96 ± 1.38 | 77.58 ± 2.39 | 83.09 ± 1.96 |
| | UNet$_{GRAY}$ | 78.33 ± 2.34 | 85.14 ± 2.14 | 57.25 ± 3.00 | 79.96 ± 2.55 | 78.00 ± 1.90 | 75.74 ± 2.38 | 82.77 ± 1.59 |
| | UNet$_{RGB}$ | 81.17 ± 2.02 | 87.54 ± 2.00 | 62.07 ± 3.12 | 83.69 ± 2.58 | 83.83 ± 1.47 | 79.66 ± 2.24 | 85.85 ± 1.54 |
| | PSPNet$_{MobileNet}$ | 75.76 ± 1.47 | 86.82 ± 1.26 | 72.66 ± 1.47 | 85.16 ± 1.65 | 74.67 ± 1.90 | 77.41 ± 1.56 | 80.87 ± 1.56 |
| | DeepLab$_{V3}$ | 80.78 ± 2.07 | 85.17 ± 2.08 | 66.03 ± 3.16 | 83.96 ± 2.52 | 79.62 ± 1.85 | 79.10 ± 2.34 | 83.55 ± 1.65 |
| | **SUIM-Net** | 81.12 ± 1.76 | 80.68 ± 1.74 | 65.79 ± 2.10 | 84.90 ± 1.77 | 76.81 ± 1.82 | 77.77 ± 1.64 | 80.86 ± 1.64 |

## V. Benchmark Evaluation

We consider the following SOTA models for performance evaluation on the SUIM dataset: *i)* FCN8 [4] with two variants of base model: vanilla CNN (FCN8$_{CNN}$) and VGG-16 (FCN8$_{VGG}$), *ii)* SegNet [3] with two variants of base model: vanilla CNN (SegNet$_{CNN}$) and ResNet-50 (SegNet$_{ResNet}$), *iii)* UNet [5] with two variants of input: grayscale images (UNet$_{GRAY}$) and RGB images (UNet$_{RGB}$), iv) pyramid scene parsing network [35] with MobileNet [54] as the base model (PSPNet$_{MobileNet}$), and *v)* DeepLab$_{V3}$ [6]. We use TensorFlow implementations of all these models and train them on the SUIM datasest using the same hardware setup (as of SUIM-Net). A few important training parameters are listed in Table II; further information can be found in their source repositories which are provided in Appendix II.

We mentioned various use cases of the SUIM dataset for semantic segmentation and saliency prediction in Section III. In our evaluation, we conduct performance comparison of the SOTA models for the following two training configurations:

- Semantic segmentation with the five major object categories (see Table I): HD, WR, RO, RI, and FV; the rest are considered as background, *i.e.*, BW=PF=SR=(000)$_{RGB}$. Each model is configured for five channels of output, one for each category. The predicted separate pixel masks are combined to RGB masks for visualization.
- Single-channel saliency prediction: the ground truth intensities of HD, RO, FV, and WR pixels are set to 1.0, and the rest are set to 0.0. The output is thresholded and visualized as binary images.

Detailed performance analysis for these two setups is presented in the following sections.

### A. Evaluation Criteria

We compare the performance of all the models based on standard metrics that evaluate region similarity and contour accuracy [1], [9]. The region similarity metric quantifies the correctness of predicted pixel labels compared to ground truth by using the notion of 'dice coefficient' aka $\mathcal{F}$ score. It is calculated using the precision ($\mathcal{P}$) and recall ($\mathcal{R}$) as $\mathcal{F} = \frac{2\times\mathcal{P}\times\mathcal{R}}{\mathcal{P}+\mathcal{R}}$. On the other hand, contour accuracy represents the object boundary localization performance; it is quantified by the mean IOU (intersection over union) scores, where $IOU = \frac{Area\ of\ overlap}{Area\ of\ union}$.

### B. Quantitative and Qualitative Analysis

We present the quantitative results in Table III. It compares the $\mathcal{F}$ and *mIOU* scores for semantic segmentation of each object category; it also compares the respective scores for saliency prediction. The results suggest that UNet$_{RGB}$, FCN8$_{VGG}$, and DeepLab$_{V3}$ generally perform better than other models. In particular, they achieve the top three $\mathcal{F}$ and *mIOU* scores for both semantic segmentation and saliency prediction. SegNet$_{ResNet}$ and PSPNet$_{MobileNet}$ also provide competitive results; however, their performances are slightly inconsistent over various object categories. Moreover, significantly better scores of SegNet$_{ResNet}$ (FCN8$_{VGG}$) over SegNet$_{CNN}$ (FCN8$_{CNN}$) validate the benefits of using a powerful feature extractor. As Table IV shows, SegNet$_{ResNet}$ (FCN8$_{VGG}$) has about twice (five times) the number of network parameters than SegNet$_{CNN}$ (FCN8$_{CNN}$). On the other hand, consistently better performance of UNet$_{RGB}$ over UNet$_{GRAY}$ validates the utility of learning on RGB image space (rather than using grayscale images as input).

TABLE IV: Comparison for the number of parameters and computational overhead of each model; the frame rates are computed on a single Nvidia™ GTX 1080 GPU.

| Model | # of parameters | Frame rate |
|---|---|---|
| FCN8$_{CNN}$ | 69.744 M | 17.11 FPS |
| FCN8$_{VGG}$ | 134.286 M | 8.79 FPS |
| SegNet$_{CNN}$ | 2.845 M | 17.52 FPS |
| SegNet$_{ResNet}$ | 15.012 M | 10.86 FPS |
| UNet$_{GRAY}$ | 31.032 M | 20.13 FPS |
| UNet$_{RGB}$ | 31.033 M | 19.98 FPS |
| PSPNet$_{MobileNet}$ | 63.967 M | 6.65 FPS |
| DeepLab$_{V3}$ | 41.254 M | 16.00 FPS |
| **SUIM-Net** | **3.864 M** | **28.65 FPS** |

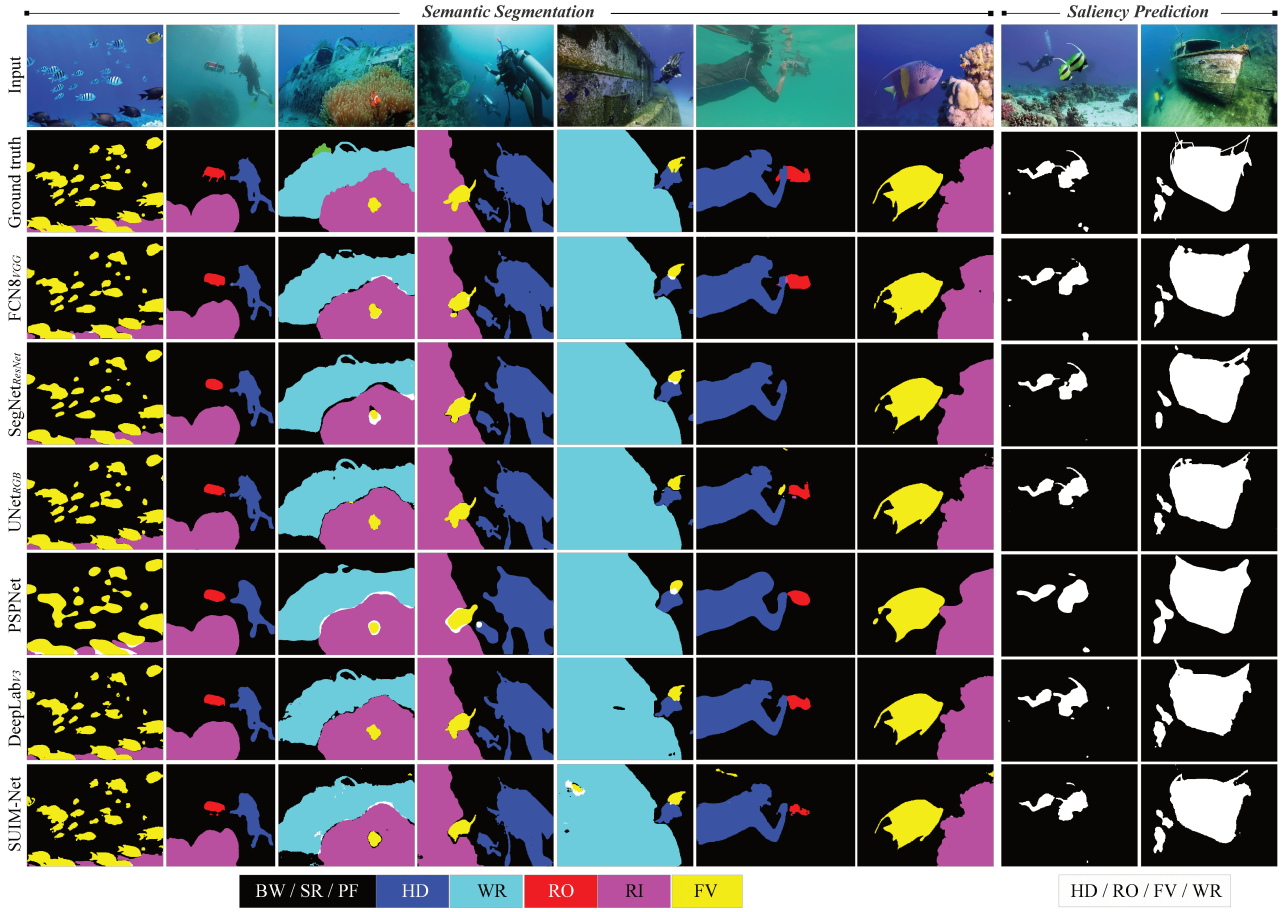



Fig. 6: A few qualitative comparison for the experiment of Table III: (left) semantic segmentation with `HD`, `WR`, `RO`, `RI`, and `FV` as object categories; (right) saliency prediction with `HD=RO=FV=WR=1` and `RI=PF=SR=BW=0`. Results for the top performing models are shown; best viewed digitally by zoom for details.

SUIM-Net provides consistent and competitive performance for both region similarity and object localization. As Table III suggests, the $\mathcal{F}$ and *mIOU* scores for SUIM-Net are within 5% margins of the respective top scores. The accuracy of semantic labeling and object localization can be further visualized in Figure 6, which shows that the SUIM-Net-generated segmentation masks are qualitatively comparable to the ground truth labels. Although UNet$_{RGB}$, FCN8$_{VGG}$, and DeepLab$_{V3}$ achieve much fine-grained object contours, the loss is not perceptually significant. Moreover, as shown in Table IV, SUIM-Net operates at a rate of 28.65 frames-per-second (FPS) on a Nvidia$^{TM}$ GTX 1080 GPU, which is significantly faster than other models in comparison. Also, it is over 10 times more memory efficient than UNet$_{RGB}$, FCN8$_{VGG}$, and DeepLab$_{V3}$. These computational aspects are ideal for its use in near real-time applications (which typically require close to 30 FPS inference rates).

Figure 7 further demonstrates the effectiveness of SUIM-Net-generated segmentation masks for fine-grained object localization in the image space. In particular, it compares the SUIM-Net's pixel-level detection of human divers and robots with the standard object detectors such as DDD [45]. In addition to providing more precise object localization, SUIM-Net incurs considerably fewer cases of missed detection, particularly in occluded or low-contrast regions in the image. Moreover, the additional semantic information facilitates much-improved saliency prediction, whereas the class-agnostic models such as Deep SESR [16] and AS-Net [46] concentrate on the high-contrast foreground regions only. These, among many others, are important use cases of the proposed SUIM dataset. Further research efforts will be useful to explore the design and feasibility of various deep visual models for other application-specific attention modeling tasks.

## VI. Conclusion

Semantic segmentation of underwater scenes and pixel-level detection of salient objects are critically important features for visually-guided AUVs. The existing solutions are either too application-specific or outdated, despite the rapid advancements of relevant literature in the terrestrial domain. In this paper, we attempt to address these limitations by presenting the first large-scale annotated dataset for general-purpose semantic segmentation of underwater scenes. The proposed SUIM dataset contains 1525 images with pixel annotations for eight object categories: fish, reefs, plants,

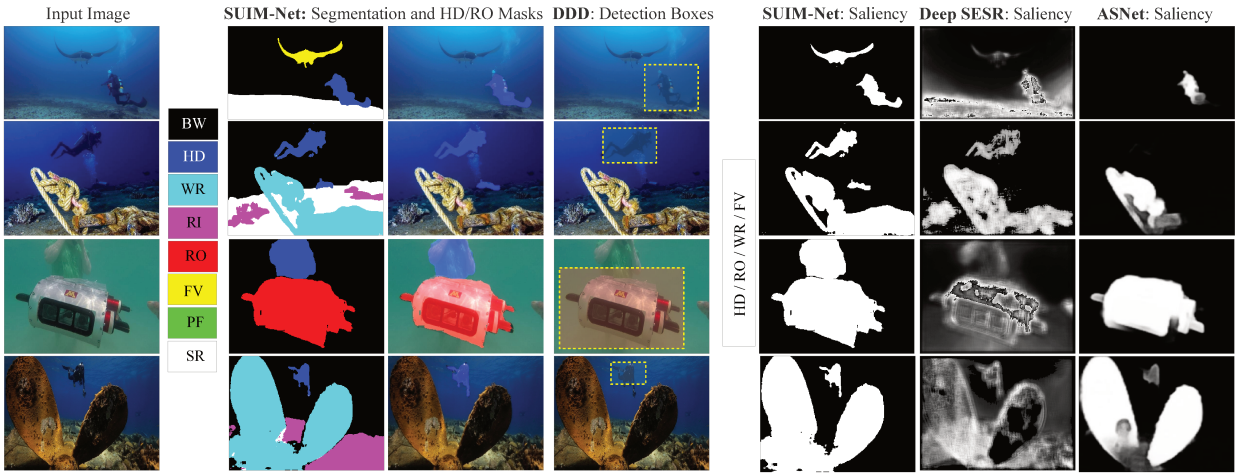



Fig. 7: SUIM-Net-generated segmentation masks are shown alongside the bounding box (BBox) outputs of DDD [45] for the detection of human divers and robots; also, their corresponding saliency masks are compared with the class-agnostic predictions of Deep SESR [16] and ASNet [46].

wrecks/ruins, humans, robots, sea-floor/sand, and waterbody background. We also provide a benchmark evaluation of the SOTA semantic segmentation approaches on its test set. Moreover, we present SUIM-Net, a fully-convolutional encoder-decoder model that offers a considerably faster runtime than the SOTA approaches while achieving competitive semantic segmentation performance. The delicate balance of robust performance and computational efficiency make SUIM-Net suitable for near real-time use by visually-guided underwater robots in attention modeling and servoing tasks. In near future, we plan to further utilize the SUIM dataset and explore various learning-based models for visual question answering and guided search; the subsequent pursuit will be to analyze their feasibility in underwater human-robot cooperative applications.

## Appendix I: Data Augmentation Parameters

We used the standard Keras libraries (https://keras.io/preprocessing/image/) for data augmentation in our work. The specific parameters are as follows: rotation range of 0.2; width shift, height shift, shear, and zoom range of 0.05; horizontal flip is enabled; and the rest of the parameters are left as default.

## Appendix II: Relevant Source Code

- FCN variants: https://github.com/divamgupta/image-segmentation-keras
- BilinearUpsampling for FCN: https://github.com/aurora95/Keras-FCN
- SegNet variants and PSPNet: https://github.com/divamgupta/image-segmentation-keras
- UNet: https://github.com/zhixuhao/unet
- DeepLabv3: https://github.com/MLearing/Keras-Deeplab-v3-plus/
- Deep SESR: https://github.com/xahidbuffon/Deep-SESR
- ASNet: https://github.com/wenguanwang/ASNet

## Acknowledgement

This work was supported by the National Science Foundation grant IIS-#1845364, the Doctoral Dissertation Fellowship (DDF) at the University of Minnesota and the Minnesota Robotics Institute (MnRI). We are grateful to the Bellairs Research Institute of Barbados for the field trial venue, and to the Mobile Robotics Lab of McGill University for data and resources.

## References

[1] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A Review on Deep Learning Techniques Applied to Semantic Segmentation," *arXiv preprint arXiv:1704.06857*, 2017.

[2] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[3] V. Badrinarayanan, A. Handa, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Robust Semantic Pixel-wise Labelling," *arXiv preprint arXiv:1505.07293*, 2015.

[4] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer, 2015, pp. 234–241.

[6] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation," *ArXiv preprint arXiv:1706.05587*, 2017.

[7] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and Transferring Mid-level Image Representations using Convolutional Neural Networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1717–1724.

[8] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 740–755.

[9] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 724–732.

[10] M. J. Islam, Y. Xia, and J. Sattar, "Fast underwater image enhancement for improved visual perception," *IEEE Robotics and Automation Letters (RA-L)*, vol. 5, no. 2, pp. 3227–3234, 2020.

[11] O. Beijbom, P. J. Edmunds, D. I. Kline, B. G. Mitchell, and D. Kriegman, "Automated Annotation of Coral Reef Survey Images," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 1170–1177.

[12] I. Alonso, M. Yuval, G. Eyal, T. Treibitz, and A. C. Murillo, "CoralSeg: Learning Coral Segmentation from Sparse Annotations," *Journal of Field Robotics (JFR)*, vol. 36, no. 8, pp. 1456–1477, 2019.

[13] NOAA, "VIAME Datasets and Challenges," https://www.viametoolkit.org/cvpr-2018-workshop-data-challenge/challenge-data-description/, 2018, accessed: 7-22-2020.

[14] M. Ravanbakhsh, M. R. Shortis, F. Shafait, A. Mian, E. S. Harvey, and J. W. Seager, "Automated Fish Detection in Underwater Images Using Shape-Based Level Sets," *Photogrammetric Record*, vol. 30, no. 149, pp. 46–62, 2015.

[15] M.-C. Chuang, J.-N. Hwang, K. Williams, and R. Towler, "Automatic Fish Segmentation via Double Local Thresholding for Trawl-based Underwater Camera Systems," in *IEEE International Conference on Image Processing*. IEEE, 2011, pp. 3145–3148.

[16] M. J. Islam, P. Luo, and J. Sattar, "Simultaneous Enhancement and Super-Resolution of Underwater Imagery for Improved Visual Perception," 2020.

[17] Y. Zhou, J. Wang, B. Li, Q. Meng, E. Rocco, and A. Saiani, "Underwater Scene Segmentation by Deep Neural Network," *UK Robotics and Autonomous Systems*, 2019.

[18] X. Li, J. Song, F. Zhang, X. Ouyang, and S. U. Khan, "MapReduce-based Fast Fuzzy C-means Algorithm for Large-scale Underwater Image Segmentation," *Future Generation Computer Systems*, vol. 65, pp. 90–101, 2016.

[19] G. Padmavathi, M. Muthukumar, and S. K. Thakur, "Nonlinear Image Segmentation Using Fuzzy C-means Clustering Method with Thresholding for Underwater Images," *International Journal of Computer Science Issues (IJCSI)*, vol. 7, no. 3, p. 35, 2010.

[20] B. Arain, C. McCool, P. Rigby, D. Cagara, and M. Dunbabin, "Improving Underwater Obstacle Detection Using Semantic Image Segmentation," in *International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 9271–9277.

[21] Y. Zhu, B. Hao, B. Jiang, R. Nian, B. He, X. Ren, and A. Lendasse, "Underwater Image Segmentation with Co-saliency Detection and Local Statistical Active Contour Model," in *OCEANS*. IEEE, 2017, pp. 1–5.

[22] Y. Girdhar, P. Giguere, and G. Dudek, "Autonomous Adaptive Exploration using Realtime Online Spatiotemporal Topic Modeling," *International Journal of Robotics Research (IJRR)*, vol. 33, no. 4, pp. 645–657, 2014.

[23] F. Shkurti, A. Xu, M. Meghjani, J. C. G. Higuera, Y. Girdhar, P. Giguere, B. B. Dey, J. Li, A. Kalmbach, C. Prahacs *et al.*, "Multi-domain Monitoring of Marine Environments using a Heterogeneous Robot Team," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2012, pp. 1747–1753.

[24] B. Bingham, B. Foley, H. Singh, R. Camilli, K. Delaporta, R. Eustice *et al.*, "Robotic Tools for Deep Water Archaeology: Surveying an Ancient Shipwreck with an Autonomous Underwater Vehicle," *Journal of Field Robotics (JFR)*, vol. 27, no. 6, pp. 702–717, 2010.

[25] M. J. Islam, M. Ho, and J. Sattar, "Understanding Human Motion and Gestures for Underwater Human-Robot Collaboration," *Journal of Field Robotics (JFR)*, pp. 1–23, 2018.

[26] J. Sattar and G. Dudek, "A Vsion-based Control and Interaction Framework for a Legged Underwater Robot," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2009, pp. 329–336.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[28] M. J. Islam, S. S. Enan, P. Luo, and J. Sattar, "Underwater Image Super-Resolution using Deep Residual Multipliers," *IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

[29] Oceana Inc., "The Ocean Animal Encyclopedia," https://oceana.org/marine-life/, 2001, accessed: 20-2-2020.

[30] ETI BioInformatics, "Marine Species Identification Portal," http://species-identification.org/, 2009, accessed: 20-2-2020.

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[32] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.

[33] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional Networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010, pp. 2528–2535.

[34] W. Liu, A. Rabinovich, and A. C. Berg, "ParseNet: Looking Wider to See Better," *arXiv preprint arXiv:1506.04579*, 2015.

[35] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid Scene Parsing Network," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2881–2890.

[36] F. Visin, M. Ciccone, A. Romero, K. Kastner, K. Cho, Y. Bengio, M. Matteucci, and A. Courville, "ReSeg: A Recurrent Neural Network-based Model for Semantic Segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2016, pp. 41–48.

[37] A. Roy and S. Todorovic, "A Multi-scale CNN for Affordance Segmentation in RGB Images," in *European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 186–201.

[38] F. Yu and V. Koltun, "Multi-scale Context Aggregation by Dilated Convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[39] Z. Ye, "Objective Assessment of Nonlinear Segmentation Approaches to Gray Level Underwater Images," *International Journal on Graphics, Vision, and Image Processing (GVIP)*, vol. 9, no. II), pp. 39–46, 2009.

[40] J. W. Kaeli, J. J. Leonard, and H. Singh, "Visual Summaries for Low-bandwidth Semantic Mapping with Autonomous Underwater Vehicles," in *IEEE/OES Autonomous Underwater Vehicles (AUV)*. IEEE, 2014, pp. 1–7.

[41] A. Maldonado-Ramírez and L. A. Torres-Méndez, "Robotic Visual Tracking of Relevant Cues in Underwater Environments with Poor Visibility Conditions," *Journal of Sensors*, vol. 2016, 2016.

[42] L. Zhang, B. He, Y. Song, and T. Yan, "Underwater Image Feature Extraction and Matching based on Visual Saliency Detection," in *OCEANS*. IEEE, 2016, pp. 1–4.

[43] M. J. Islam and J. Sattar, "Mixed-domain Biological Motion Tracking for Underwater Human-robot Interaction," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 4457–4464.

[44] F. Shkurti, W.-D. Chang, P. Henderson, M. J. Islam, J. C. G. Higuera, J. Li, T. Manderson, A. Xu, G. Dudek, and J. Sattar, "Underwater Multi-Robot Convoying using Visual Tracking by Detection," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017.

[45] M. J. Islam, M. Fulton, and J. Sattar, "Toward a Generic Diver-Following Algorithm: Balancing Robustness and Efficiency in Deep Visual Detection," *IEEE Robotics and Automation Letters (RA-L)*, vol. 4, no. 1, pp. 113–120, 2018.

[46] W. Wang, J. Shen, X. Dong, and A. Borji, "Salient Object Detection Driven by Fixation Prediction," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1711–1720.

[47] D. Yu, J. Fu, T. Mei, and Y. Rui, "Multi-level Attention Networks for Visual Question Answering," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4709–4717.

[48] L. Bazzani, H. Larochelle, and L. Torresani, "Recurrent Mixture Density Network for Spatiotemporal Visual Attention," *arXiv preprint arXiv:1603.08199*, 2016.

[49] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *CoRR, abs/1502.03167*, 2015.

[50] V. Nair and G. E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines," in *Proc. of the International Conference on Machine Learning (ICML)*, 2010, pp. 807–814.

[51] Z. Zhang and M. Sabuncu, "Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels," in *Advances in Neural Information Processing Systems*, 2018, pp. 8778–8788.

[52] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean *et al.*, "TensorFlow: A System for Large-scale Machine Learning," in *USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 2016, pp. 265–283.

[53] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[54] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *arXiv preprint arXiv:1704.04861*, 2017.