Underwater Object Detection Under Domain Shift

Joseph L. Walker, Zheng Zeng, Chengchen L. Wu, Jules S. Jaffe, Kaitlin E. Frasier, and Stuart S. Sandin

***Abstract*—There is increasing interest in using deep learning–based object recognition algorithms to automate the labeling of image data collected from marine surveys. However, underwater object detection is a particularly challenging problem due to changes in scattering and absorption of light, and spotty data collection efforts, which rarely capture the broad variability. Using deep learning–based object detection systems for long-term or multisite marine surveying is further complicated by shifting data distributions between training and testing stages. Using data from the 100 Island Challenge, we investigate how object detection performance is impacted by changes in site characteristics and imaging conditions. We demonstrate that the combined use of data augmentation and unsupervised domain adaptation techniques can mitigate performance drops in the presence of domain shift. The proposed methodologies are broadly applicable to observational data sets in marine and terrestrial environments where a single algorithm needs to adapt to and perform comparably across changing conditions.**

***Index Terms*—Aquatic ecosystems, digital images, image processing, unsupervised learning.**

I. Introduction

OPTICAL imaging has remained an indispensable tool in oceanographic studies, as it offers detailed descriptions that are easily interpreted by humans. As a result, a myriad of systems have been developed for acquiring optical images in almost every oceanographic context. Autonomous underwater vehicles and unmanned underwater vehicles equipped with optical cameras have been used for the exploration and mapping of the seafloor [1], [2], monitoring invasive species [3], and fisheries management [4]. Imaging systems for in situ studies of plankton and other marine particles have also been developed [5], [6], [7], [8]. The rising popularity of these tools has led to an explosion in underwater optical data collection [9]. This increase in data has driven the need to develop object detection systems that can automate the analysis of underwater digital imagery.

Manuscript received 4 October 2022; revised 19 July 2023 and 8 January 2024; accepted 1 July 2024. Date of publication 12 September 2024; date of current version 15 October 2024. This work was supported by the Office of Naval Research and Task Force Ocean (Robert Headrick) under Grant N00014-19-1-2851. *(Corresponding author: Joseph L. Walker.)*

Associate Editor: C. Li.

Joseph L. Walker, Jules S. Jaffe, Kaitlin E. Frasier, and Stuart S. Sandin are with the Scripps Institution of Oceanography, University of California San Diego, San Diego, CA 92093-0238 USA (e-mail: jlwalker@ucsd.edu; jjaffe@ucsd.edu; kfrasier@ucsd.edu; ssandin@ucsd.edu).

Zheng Zeng is with the Department of Electrical and Computer Engineering, University of California San Diego, San Diego, CA 92093-0238 USA (e-mail: zhz396@ucsd.edu).

Chengchen L. Wu is with the Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA 92093 USA (e-mail: c8wu@ucsd.edu).

Digital Object Identifier 10.1109/JOE.2024.3425453

Object detection is a computer vision task concerned with locating and classifying objects in images or videos. The most significant advancements in object detection can be attributed to the use of deep convolutional neural networks. Currently, one of the most popular architectures for object detection is the faster region-based convolutional network (Faster R-CNN) [10]. The Faster R-CNN consists of three modules.

1) A feature extractor convolutional neural network to extract features from the entire image.
2) A region proposal network, which is trained end-to-end with the rest of the detection network to propose regions of interest in the feature map produced by 1).
3) Two fully connected networks for classification and bounding box regression.

Underwater object detection is a particularly challenging problem as images are typically of lower quality compared to out-of-water images due to light scattering and absorption. The lack of precise control over the relative imaging depth and orientation of objects in underwater environments can produce high variability in their features. Despite these challenges, numerous applications of underwater object detection exist, ranging from the estimation of plankton and fish population densities [11], biodiversity monitoring of coral reefs [12], unexploded ordnance detection [13], and detection of other man-made objects [14], [15].

A fundamental challenge in incorporating deep learning technology in oceanography (and most other real-world applications) arises from the fact that models tend to overfit the training data distribution. Differences in the training and testing data set distributions, referred to as data set shifts, have been shown to contribute to diminished model performance [16], [17], [18]. Changes in the sampling location or methodology can produce data set shifts by altering image appearance (shifts in illumination, color, noise, etc.), background features, or statistical differences in object class features (e.g., new phenotypes or morphologies of species of interest). The specific term used to describe the statistical changes in input features is known as domain shift [18]. Another type of data set shift is prior probability shift, characterized by differences in the prior probabilities of predicted variables between the training and testing phases [19], [20], [21], [22].

For applications of deep learning in oceanography, model deployment is almost always limited to the same study site and data collection protocol as the training data. However, oceanographic data are most often collected in multiple locations with varying environmental conditions, making the application of a model built in a single context insufficient to achieve high performance across use cases. One solution for producing a more

generalizable model is to annotate data from all environments in which the model is deployed. However, data annotation is extremely costly and serves as the primary bottleneck in incorporating deep learning in long-term or multisite studies. It is therefore desirable to develop adaptive deep learning models that can scale to many study sites even when annotated data are limited to a single site.

Data augmentation (DA) is often used to artificially increase the volume of available training data [23]. This process involves defining a set of augmentation functions that alter the appearance of the training data while preserving the class label, effectively synthesizing new data examples from existing data. Commonly used augmentation functions include image flipping, cropping, translation, and noise addition. These augmentations are often treated as universal, as they are used across a range of image recognition applications. In some cases, it may be possible to design specialized augmentation functions that address known sources of variability to improve model generalizability. For example, if future data are expected to be collected using an imaging system with higher illumination intensity, then illumination synthesis could be used to simulate the difference between collected and future data. However, this approach may require a priori knowledge of the variability and well-defined augmentation functions that accurately model the variability.

The problem of domain shift has received a significant amount of attention in recent years primarily for the development of autonomous driving. In this context, collecting and annotating data from enough environments, weather conditions, and sensor configurations to ensure that future data are not outside the training distribution may be prohibitively costly or impossible. In practice, it may only be possible to collect labeled data from a single source domain; however, acquiring *unlabeled* data from the testing, i.e., target, domain may be more attainable. In such circumstances, it has been shown that leveraging both labeled data from a source domain and unlabeled data from a target domain can help achieve better performance on data from the target domain [24], [25], [26], [27]. This technique is referred to as unsupervised domain adaptation (UDA).

Prior works in underwater object recognition under domain shift have largely focused on using domain generalization techniques [28], [29], [30]. Unlike UDA, domain generalization methods aim to build models that can generalize well across multiple target domains without accessing unlabeled target domain data during training [31]. Domain generalization techniques are particularly beneficial in object detection applications where the data arrive continuously and need to be processed in real time. However, in cases where the real-time annotation is not required and unlabeled target data are available during training, UDA has been shown to outperform domain generalization [32], [33]. We describe three commonly used classes of UDA methods in Section I-A–C.

A. Adversarial Feature Learning

First introduced in the context of image classification, adversarial feature learning (AFL) involves the use of a domain classifier to adversarially train the model to learn domain invariant features [25]. Source images are assigned one domain label, whereas target images are assigned another. The domain classifier's objective is to categorize images based on their respective domains. During backpropagation, the weights of the domain classifier are updated and then the gradients are pushed through a gradient reversal layer before being applied to the weights of the convolutional layers. This results in the learning of features that fool the domain classifier that are domain-independent.

B. Image-to-Image Translation

Another more intuitive approach to adaptation is to match the appearance of the source domain images to that of the target domain (or vice versa). Image-to-image translation reduces the domain discrepancy in the pixel domain, which has the advantage of utilizing human visual inspection for quality assessment. Many contemporary image-to-image translation techniques borrow directly from or use ideas similar to the CycleGAN model [34].

C. Pseudolabeling

Pseudolabeling (PL) leverages the model's own predictions on unlabeled target data to create pseudolabels, essentially "teaching" itself by treating its own outputs as ground truth. This process generally follows a three-step process.

1) The model is trained on labeled source data, establishing a foundation for the task.
2) The model uses this knowledge to predict labels for the unlabeled target data, creating initial guesses for the target domain.
3) To refine these guesses, the model retrains, incorporating both source data and unlabeled target data with its own pseudolabels added as additional examples.

This iterative process improves the quality of pseudolabels and boosts the model's performance on the target domain.

D. Contributions

The goal of this study is to train a Faster R-CNN model using labeled data from one source environment that can scale to many target environments. To do this, we use a combination of DA and UDA techniques to minimize domain shifts between environments. Our primary contributions are summarized as follows.

1) We present a new underwater object detection data set for domain adaptation experimentation.
2) We present a framework for developing robust underwater object detectors that are more resilient to data set shifts.
3) For our task, we show that existing UDA techniques can be improved by incorporating DA.
4) In the case of limited training data (as in our case), we show that the HM-multivariate Gaussian distribution (MVGD)-HM [34] image-translation algorithm can produce better image-to-image translation results than more sophisticated methods such as CycleGAN.
5) Our data set and code are publicly available and will be used to augment data collection for the 100 Island Challenge (100IC) project (described in Section II-A) and other projects requiring robust underwater object detection.

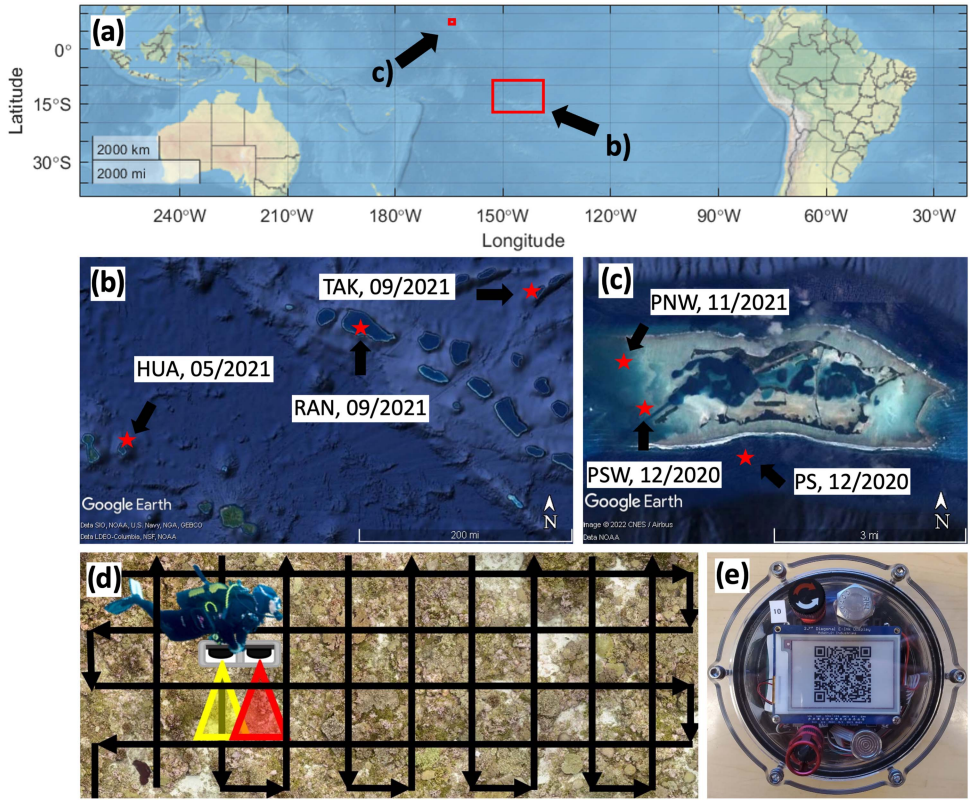


Fig. 1. Sampling locations, dates, and methodology. (a) Data were collected from two regions (red boxes) in the Tropical Pacific Ocean. (b) Study sites from the Tuamotu Archipelago region, which include the islands Takapoto (TAK), Rangiroa (RAN), and Huahine (HUA). The sampling date for each study site is reported as MM/YYYY. (c) Study sites from the Palmyra Atoll, which include sites in the southern (PS), southwest (PSW), and northwest (PNW) parts of the island. (d) For each study site, a survey plot (100 or 200 m^2) is defined and imaged by divers in a grid pattern. (e) Photograph of the SUIT.

II. Material and Methods

A. Data Set

The 100IC is an ongoing collaborative effort based at Scripps Institution of Oceanography, University of California San Diego, San Diego, CA, USA, to digitally archive and monitor coral reefs across the globe. Using tools of large-area imaging, high-resolution images have been collected and collated to form comprehensive digital mosaics and 3-D reconstructions from multiple coral reef sites at each of over 100 islands across the globe. These detailed maps enable the study of benthic dynamics at an unprecedented spatial scale. The 100IC has incorporated the use of smart underwater imaging telemeters (SUITs) to facilitate in situ environmental data collection to complement visual surveys of coral reefs [36]. The 100IC has produced a unique data set of this standardized object (the SUIT) that has been imaged across multiple study sites and imaging conditions. This data set, consisting of a single annotated class, is therefore particularly well suited for the study of binary underwater object detection under domain shift.

In this study, we consider the subset of 100IC image data that were collected from two regions in the Tropical Pacific Ocean. The data come from the islands Huahine (HUA), Takapoto (TAK), and Rangiroa (RAN) in the Tuamotu Archipelago region as well as three sites around the Palmyra Atoll, which include sites in the southern (PS), southwest (PSW), and northwest (PNW) parts of the island. The location and sampling date of these sites and an illustration of the data collection procedure and SUIT are shown in Fig. 1. Example images from each of the six sites are shown in Fig. 2. Table I presents the count of images collected and bounding box annotations per study site, along with statistics describing the distribution of bounding box sizes. In the following sections, we identify two types of variability across the images which make detection of the SUITs challenging.

B. Variability in Low-Level Features

In the context of image processing, low-level visual information may include brightness, texture, color distribution, and noise. In the 100IC imagery, these features can change according to various physical phenomena that influence the image formation process. Due to the large spatiotemporal range of the sampling, it is likely that the inherent optical properties of the seawater are inconsistent across sampling periods. This can lead to different degrees of color distortion and contrast loss. The ambient light field is also subject to change according to weather conditions leading to inconsistent scene illumination. Caustic patterns on the seafloor, especially visible in Fig. 2(d), can create bright white regions that are similar to the white pixels of the SUIT display. Finally, some images have been color-corrected, whereas others have not. In cases where color

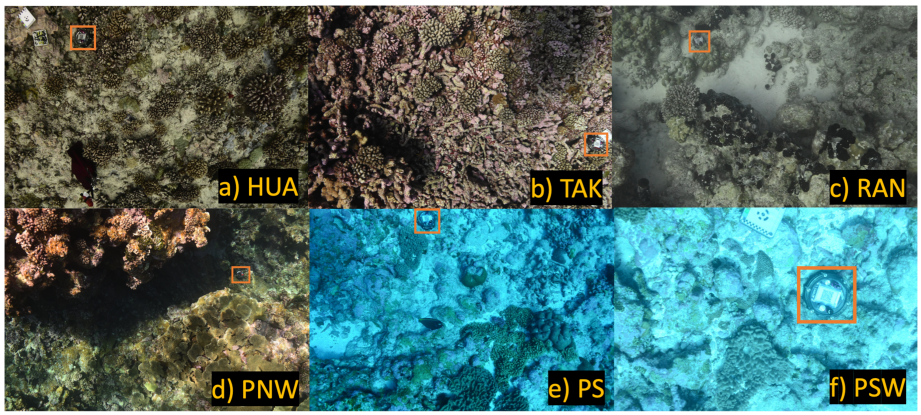


Fig. 2. Example images collected from the six study sites. The bounding boxes containing the SUITs are shown in orange.

TABLE I
DATA SET STATISTICS FOR EACH STUDY SITE

Site name	Abbreviation	Num. images	Num. boxes	Avg. box area	Std. Dev. box area
Huahine Island, French Polynesia	HUA	4550	351	0.8%	0.7%
Rangiroa Atoll, French Polynesia	RAN	3321	153	1.0%	0.5%
Southwest Palmyra atoll, USA	PSW	1596	239	1.0%	1.0%
South Palmyra Atoll, USA	PS	212	81	4.8%	3.0%
Northwest Palmyra Atoll, USA	PNW	3728	450	0.2%	0.1%
Takapoto Atoll, French Polynesia	TAK	6866	835	0.7%	0.4%

The average and standard deviation of bounding box sizes are reported and expressed in terms of a fraction of the total image area, where all images are 500 × 751 pixels. Note that most of the collected images from the study sites do not contain SUITs.

TABLE II
IMAGE FEATURE SIMILARITY BETWEEN STUDY SITES

	HUA	RAN	PSW	PS	PNW	TAK
HUA	1.0	0.94	0.88	0.88	0.95	0.96
RAN	0.94	1.0	0.9	0.84	0.9	0.91
PSW	0.88	0.9	1.0	0.94	0.88	0.85
PS	0.88	0.84	0.94	1.0	0.86	0.85
PNW	0.95	0.9	0.88	0.86	1.0	0.88
TAK	0.96	0.91	0.85	0.85	0.88	1.0

Similarity scores have been rescaled linearly between [0,1] where 1 indicates mean size is identical, and 0 represents the largest observed dissimilarity.

TABLE III
SUIT SIZE SIMILARITY BETWEEN STUDY SITES

	HUA	RAN	PSW	PS	PNW	TAK
HUA	1.0	0.94	0.11	0.96	0.89	0.99
RAN	0.94	1.0	0.17	0.98	0.83	0.94
PSW	0.11	0.17	1.0	0.16	0.0	0.11
PS	0.96	0.98	0.16	1.0	0.89	0.95
PNW	0.89	0.83	0.0	0.89	1.0	0.9
TAK	0.99	0.94	0.11	0.95	0.9	1.0

Similarity scores have been rescaled linearly between [0,1] where 1 indicates mean size is identical, and 0 represents the largest observed dissimilarity.

correction is applied, broad assumptions (such as constant scene depth) are made. Therefore, there is high variability among the color-corrected images including noticeable depth-related artifacts.

We provide a quantitative measure for low-level feature similarity between the environments by comparing the image features extracted from a VGG16 [37] encoder. For each of the eight study sites, a centroid is calculated by averaging the extracted image features. The pairwise distances between the centroids are used to calculate the similarity between the study sites using a cosine similarity measure. The pairwise similarity values are reported in Table II.

C. Variability in SUIT Scale

Another source of variability is created by changes in structural features associated with the SUITs themselves, caused by changes in scale and orientation. The data used in this study were collected using a Nikon D780 or Nikon D7000 camera used in combination with a 24 mm and 18 mm wide-angle lens, respectively. This affects the apparent size of the SUITs in the images. Variability in the distance between the camera and the seafloor can make the SUITs appear to be differently sized. Topography is also highly variable across the environments. In environments with highly textured benthic surfaces, the SUITs are more likely to be imaged at an angle. For each of the eight sampling locations, we calculate the average bounding box size and compute the magnitude of the pairwise differences between each of the averages. We divide these differences by the largest difference to scale the values to be between 0 and 1 and then subtract each of the values from 1 to calculate the similarity. All pairwise similarity scores are reported in Table III.

TABLE IV
LEAVE-ONE-IN CROSS-VALIDATION RESULTS

Model	Adaptation (source→ target)					
	TAK → HUA	TAK → RAN	TAK → PSW	TAK → PS	TAK → PNW	TAK → TAK
baseline	87.0	87.5	45.5	69.5	65.9	-
DA	85.2	85.5	85.2	58.4	72.2	-
CGAN	86.7	87.6	60.7	80.2	71.0	-
AFL_{CGAN}	86.2	87.6	62.3	86.0	77.2	-
AFL_{CM}	**87.9**	87.7	62.0	86.3	79.1	-
AFL_{CM}+*DA*	87.7	**88.1**	**89.1**	87.9	**82.5**	-
PL_{CGAN}	86.6	87.6	59.3	85.3	74.5	-
PL_{CM}	87.0	86.2	62.4	88.4	75.5	-
PL_{CM}+*DA*	86.1	87.5	85.7	**90.3**	80.7	-
	PNW → HUA	PNW → RAN	PNW → PSW	PNW → PS	PNW → PNW	PNW → TAK
baseline	68.6	26.6	1.7	1.2	-	42.0
DA	84.0	86.1	62.7	40.0	-	88.4
CGAN	76.2	58.4	7.4	62.1	-	57.1
AFL_{CGAN}	73.2	61.9	0.3	64.8	-	64.8
AFL_{CM}	75.9	62.3	0.3	65.5	-	65.7
AFL_{CM}+*DA*	**85.9**	**86.9**	**76.2**	86.6	-	**89.1**
PL_{CGAN}	71.5	59.4	1.1	68.7	-	61
PL_{CM}	74.2	59.1	2.4	70.4	-	62.4
PL_{CM}+*DA*	83.8	80.4	72.9	**89**	-	88.2

Each model is trained using a single source study site (X) where all other study sites are individually treated as a target domain (Y). The adaptation of a source study site to a target study site is expressed as X→Y. Two source study sites, TAK and PNW, are considered individually. The five primary models and their variants are separated by the dashed lines. Results are reported in terms of classification accuracy (higher is better). The best-performing model for each adaptation is shown in bold.

1) Progressive Domain Adaptation: We use UDA techniques to mitigate performance drops caused by image feature differences. Specifically, we compare two different approaches that follow the progressive domain adaptation (PDA) method proposed by Hsu et al. [27]. The authors proposed PDA as a method for gradually aligning the features of a model using a multistage feature alignment approach. First, a synthetic image data set is generated by mapping the source images to the target domain using a CycleGAN. In the first stage, the features of the source and synthetic domains are aligned using AFL. In the second stage, the features of the synthetic and target domains are aligned using AFL.

The original PDA method uses a CycleGAN to produce the synthetic data set [27]. However, CycleGANs typically require large amounts of training data to produce quality image mappings. For the experiments in [27], Hsu et al. used between 3475–41 986 images to train the CycleGAN models. This amount of data is often not available for many oceanographic applications where data collection and annotation is difficult. For this study, only 81–835 instances of the SUIT were collected from each study site. In our initial CycleGAN experiments, we observed that nearly all the translated images featuring SUITs exhibited substantial distortions and failed to retain their characteristic features. For this reason, we replaced the CycleGAN with the HM-MVGD-HM color-matching algorithm which requires no training [35]. The HM-MVGD-HM algorithm uses an analytical solution to an MVGD color transfer equation in addition to classical histogram matching. Example synthetic images using CycleGAN and HM-MVGD-HM are shown in Fig. 6. We report object detection performance using CycleGAN and HM-MVGD-HM in Table IV.

Since the original paper by Hsu et al. [27] was published, other multistage adaptation approaches have been proposed. In this study, we compare the performance of the original method with a semisupervised version of PDA similar to the method by Inoue et al. [38], which uses PL instead of AFL. Unlike Inoue et al. [38], we do not use image-level annotations from the target domain. We compare the results of this PL approach using both CycleGAN and HM-MVGD-HM and report results in Table IV.

2) Data Augmentation: To address variability in SUIT object size, we implement five DA techniques, each designed to simulate a potential source of variation. The (x,y) pixel coordinate of the SUIT center in an image is arbitrary and is determined only by the SUIT's placement relative to the transect during image collection. To prevent the models from learning irrelevant patterns related to the position of the SUITs, we simulate different SUIT placements by applying random image translations and rotations. We define this set of placement transformations as $T_P = \{\text{translation, rotation}\}$. The distance between the camera and the bottom will affect the apparent size of the SUIT. Simulating imaging at a closer range can be approximated by using random cropping. However, imaging at greater distances involves simulating the effects of resolution and contrast loss. To simulate these effects, we created an augmentation function that performs downsampling followed by contrast reduction. The subsequent image is then zero-padded to the original image size. Because symmetrically padding the image would bias SUIT placement toward the image center, padding is followed by a random translation. We refer to this augmentation as *distance*. To simulate different imaging angles, we adopt an approach similar to that presented by Huang et al. [39] by applying perspective transformations to the images. Because perspective

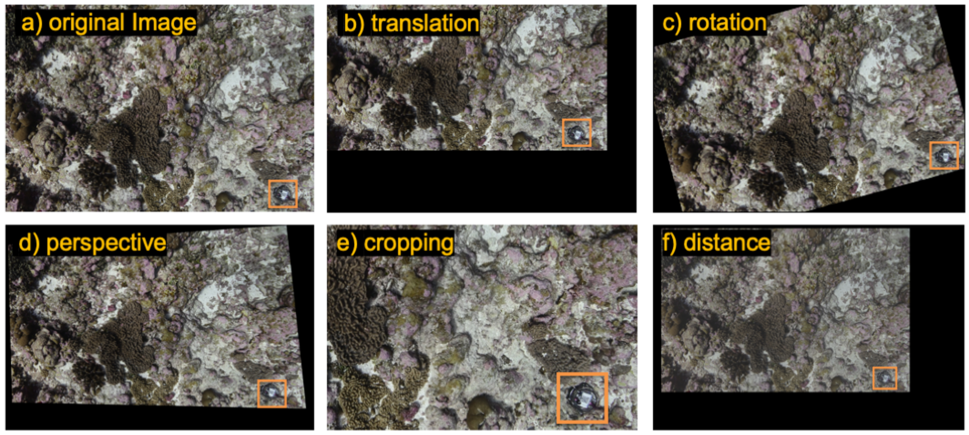


Fig. 3. Data augmentations. (a) Example image from the TAK study site. (b)–(f) Output of the translation, rotation, perspective transformation, cropping, and distance image augmentation functions, respectively, using the image in (a) as input.

transformation, cropping, and distance augmentations can distort the apparent size of the SUIT, we refer to this set of transformations as $T_S = \{\text{perspective, cropping, distance}\}$. Fig. 3 shows examples of all five augmentations.

3) Models: All models use a Faster R-CNN architecture with a VGG16 [37] backbone. We consider five primary models.

1) *Baseline*: Faster R-CNN trained without UDA or DA. Models are trained on data from a single study site and applied directly to a target site.
2) DA: Same as *baseline* but trained using DA.
3) CGAN: Same as *baseline* but source images are first translated to the target domain using a CycleGAN [34].
4) AFL_{CGAN}*:* Faster R-CNN trained with PDA using AFL and CycleGAN, as proposed by Hsu et al. [27].
5) PL_{CGAN}*:* Faster R-CNN trained with PDA using PL and CycleGAN, similar to that presented in [38].

In addition to these five primary models, we consider variants of the AFL and PL models by incorporating DA and the HM-MVGD-HM color-matching algorithm. AFL_{CM} and PL_{CM} indicate that the CycleGAN has been replaced with the HM-MVGD-HM color-matching algorithm. AFL_{CM}+DA and PL_{CM}+DA indicate that the CycleGAN has been replaced with the HM-MVGD-HM color-matching algorithm and DA is used during the alignment stages.

4) Experimental Setup: The models undergo training and testing following a leave-one-in cross-validation methodology. Specifically, they are trained using labeled data from a single source study site while each remaining study site is considered individually as the target domain. All study sites with at least 400 bounding box annotations are used as source and target environments. Study sites with fewer examples are used as target domains only. For models using DA, one augmentation is selected randomly from both T_P and T_S (defined in Section II-C). The transformations are applied only during training to both source and target images. During testing, no augmentations are applied. The process of incorporating DA into the AFL process is shown in Fig. 4. Models using PL follow a similar process to Fig. 4, except pseudolabels are used during the alignment stages. A batch size of one is used during training and the images are resized to 500 × 751. All experiments were run using a Tesla P100 GPU and Intel Xeon 6126 CPU.

III. Results

Table IV shows that in all adaptation scenarios, AFL_{CM} outperformed or performed very similarly to the best-performing model. The two most notable exceptions are the cases where the model is adapted to the PS target domain. This case suggests that PL approaches may have an advantage when there are fewer available images in the target domain (see Table I). The limited data likely hampered the AFL approaches because a larger data set is required to train the domain discriminator network. Additionally, the higher ratio of bounding box annotations to available images reduces the potential for pseudolabels to introduce false positives.

The performance of *baseline* varied significantly across the adaptation experiments. Tables II and III indicate that the TAK→HUA and TAK→RAN adaptations have relatively high image features and SUIT size similarity. For both adaptations, *baseline* performed comparatively well with the other models providing little improvement. For adaptation instances with relatively low image feature similarity but similar SUIT size similarity, which includes TAK→PS, TAK→PNW, and PNW→PS, models utilizing some form of UDA outperformed DA and *baseline*. This supports the hypothesis that UDA techniques are most effective for bridging differences in low-level features.

For adaptation instances with low SUIT size similarity, including every scenario in which study site PNW is used as a source, DA outperformed all models that did not use DA, except in the case of PNW→PS. We note that this case also exhibits low image feature similarity and that DA still brought significant improvements compared to *baseline*. As is shown in Fig. 5, *baseline* is restricted to predicting regions that are of a similar size to the bounding box annotations of the source data set. The added augmentation functions allow the model to consider a greater range of bounding box predictions. This

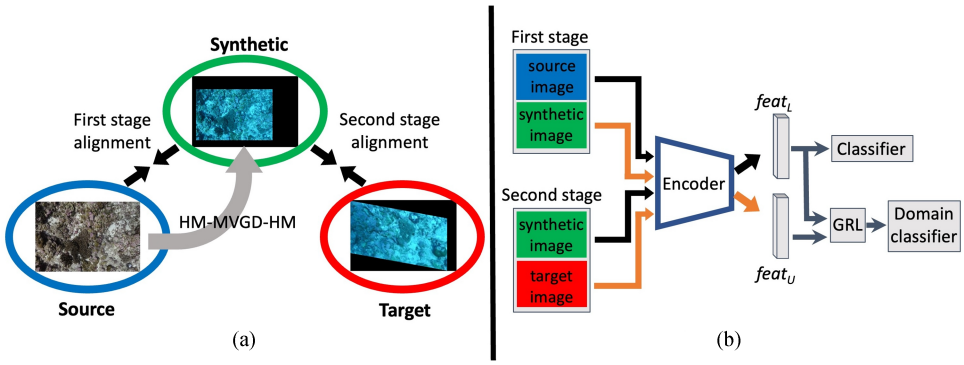


Fig. 4. (a) Overview of the augmented version of the AFL progressive domain adaptation method by Hsu et al. [27] using the HM-MVGD-HM color-matching algorithm and data augmentation, referred to as AFL_{CM}+DA. A source image (blue oval) and a target image are drawn from the source and target study sites, respectively. A synthetic image is generated by color-matching the source image to the target image using the HM-MVGD-HM algorithm. The synthetic and target images are then augmented, producing the images seen in the green and red ovals, respectively. Black arrows represent the feature alignment steps. (b) Illustration of the adversarial feature alignment process. In the first stage of training, features are extracted from the labeled source image and unlabeled synthetic image, denoted as feat_{L} and feat_{U}, respectively. Supervised object detection is performed using only feat_{L}. Adversarial feature learning is performed by passing both feat_{L} and feat_{U} to the domain classifier whose gradients are reversed during backpropagation when passed through the gradient reversal layer (GRL). In the second stage of training, features are extracted from the labeled synthetic image and unlabeled target image, where labels for the synthetic image are inherited from the source image.

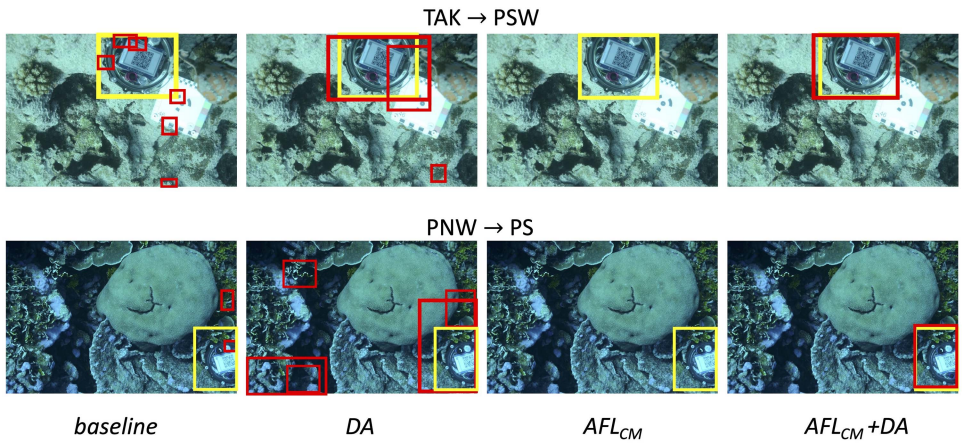


Fig. 5. Region proposals (red) from four different models for an example image from the target study site of two adaptation experiments. (a) Using TAK as a source and PSW as a target. (b) Using PNW as a source and PS as a target. Ground truth bounding boxes are shown in yellow.

supports the hypothesis that DA techniques may be more effective for bridging apparent structural differences in the objects of interest.

The PNW→PS and PNW→PSW adaptations are assumed to be the most difficult as they exhibit low image features and SUIT size similarity. This difficulty is evident by the very low performance of *baseline* in both cases. Despite the large shift in data distributions, the best-performing model was able to improve performance on the target domains dramatically.

Transforming images from the source domain to the target domain using the HM-MVGD-HM algorithm requires no additional training. Generating all ten synthetic data sets for the ten adaptation experiments took approximately 40 min or approximately 0.5 s per image. Generating the synthetic images using a CycleGAN required significantly more memory allocation and increased training time. CycleGAN training took about 5–6 min per epoch or 16.6–20 h in total for each source/target pair. As seen in Table IV, the trainable CycleGAN generally did not provide improved performance compared to the HM-MVGD-HM algorithm. We believe that the relatively small number of SUIT annotations per study site was insufficient for training a CycleGAN and resulted in poor target domain rendering of the SUITs (see Fig. 6).

We performed ablation on the five DA functions for adaptations TAK→HUA and PNW→PSW and show the results in Fig. 7. These two adaptations were selected for ablation due to their representation of extreme cases, where SUIT size similarity is either very small or very large (see Table III). The results of the ablation reveal that the performance contribution of each of the five augmentation functions is strongly dependent on the variability in apparent size. In cases where there is little to no difference in apparent object size between the source and target study sites, the incorporation of any amount of DA can negatively impact performance [see Fig. 7(a)]. However, if the difference

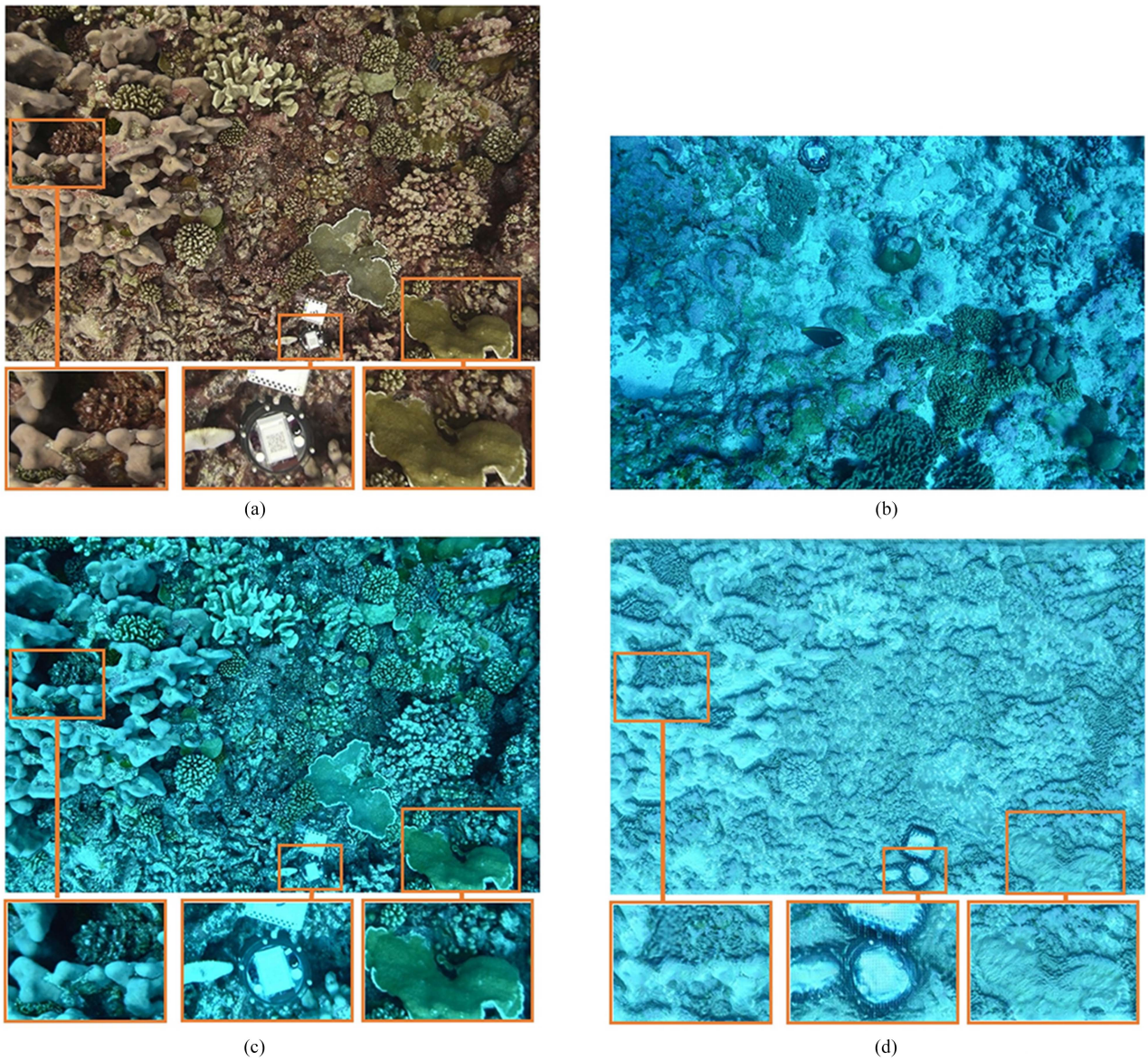


Fig. 6. Qualitative comparison of image-to-image translation methods using images from TAK and PS as source and target, respectively. (a) Random image drawn from TAK to be translated to PS. (b) Example image from PS. (c) TAK image is color-matched to the target image using the HM-MVGD-HM algorithm. (d) TAK image is translated to the PS environment using a CycleGAN model.

in apparent object size is large, the choice of augmentation functions can have a substantial impact on performance [see Fig. 7(b)].

The results suggest that in the case where it is known a priori that the target domain objects will appear much larger/smaller, then the best results may be achieved by limiting the set of augmentations to a set of function(s) that exclusively model this difference. Fig. 7(a) and (b) suggests that incorporating augmentation functions that do not directly relate to the sources of variability may negatively impact performance. However, we note that the cases studied in the ablation represent the extreme cases and that in the absence of a priori knowledge of the object variability, AFL_{CM}+DA (trained using all augmentations) still performs the best on average and therefore we conclude that using the entire set of augmentations is a strong default choice.

IV. Discussion

Developing generalizable object detection models is complicated by shifts in data distribution. We have shown that domain shift can greatly impact detection performance. However, we demonstrate that by combining DA with existing UDA techniques, performance drops can be significantly reduced. This is a significant finding, as the results provide the possibility for broad spatiotemporal surveying even when annotated data is limited to one study site. We further show that models can be trained to be robust against other sources of variability including color correction and object scale.

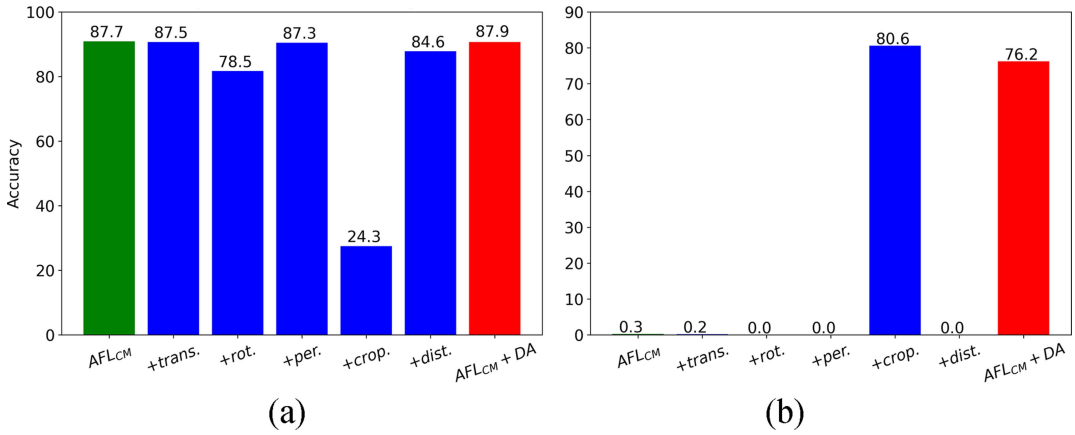


Fig. 7. Ablation on the five augmentation functions: translation (trans.), rotation (rot.), perspective (per.), cropping (crop.), and distance (dist.). (a) Ablation results using TAK as source and HUA as target. (b) Ablation results using PNW as source and PSW as target. The performance of $\mathrm{AFL}_{\mathrm{CM}}$, which uses no data augmentation, is shown in green. The performance of $\mathrm{AFL}_{\mathrm{CM}}$+DA, which uses all five augmentations, is shown in red. The ablation applies each one of the five augmentations individually with $\mathrm{AFL}_{\mathrm{CM}}$. (a) TAK→HUA (similar mean SUIT size). (b) PNW→PSW (dissimilar mean SUIT size).

Overall, the results of the cross-validation experiments are intuitive - source and target study sites with low visual differences produced higher baseline performance and, in these cases, more sophisticated models produced marginal improvements. However, in many cases, study sites with significant visual differences benefited tremendously from the combined use of UDA and DA. An alternative approach to bridging differences in data distributions could involve the use of light attenuation models that are specific to each environment. However, this would require accurate measurement and prediction of the ambient light field and inherent optical properties of the water column. This approach is likely most appropriate when target domain image data is unavailable during training, which may include real-time detection tasks. In these cases, a priori knowledge of future imaging conditions should be leveraged to synthetically generate the training data set. If target data is available during model training, we propose that the main advantage of using the techniques developed in this study is that they require no prior knowledge of the light field or water column properties. Instead, environment-agnostic features can be learned through the combined use of image translation and AFL. This data-first approach also has the advantage of scaling to many sources of visual variability beyond water column properties that can be difficult to model, including different cameras or lenses or illumination patterns (e.g., shadows or caustics).

There is growing interest in using video to conduct oceanographic surveys [40], [41]; however, directly applying still image-based object detection models to video presents unique challenges. These challenges include increased computational costs and motion blur and video defocus. In addition, the methods outlined in this study assume that target domain data are available during training; however, applications of real-time object detection and live target searching may be incompatible with this assumption. In these cases, the models must be able to scale to multiple target domains using source data alone. We conclude that domain randomization techniques remain the best possible solution when target domain data are unavailable [42], [43], [44].

In many real-world applications of machine learning, including in oceanography, the available annotated data are insufficient for training models with large parameter spaces and could result in overfitting [45]. Fig. 6 shows that the SUITs translated by the CycleGAN exhibit lower visual quality compared to those produced by the HM-MVGD-HM algorithm. This discrepancy is likely attributed to the sparse representation of SUIT instances within the image data set, with only approximately 10% of the images containing a SUIT. The scarcity of SUIT instances during CycleGAN training led to the blurring of SUIT features, causing the loss of their characteristic geometric details during the translation process.

V. Conclusion

Two major present-day obstacles hindering advances in the analysis of oceanographic data include 1) challenges in developing analysis tools that are robust across different conditions, equipment, and locations; and 2) costs associated with trying to collect and annotate variable data sets from which effective models can be trained. Our results indicate that the procedures developed in this study may be a viable solution for improving model robustness while reducing the human data annotation effort. We view this as a critical step for maximizing the utility and cost-effectiveness of oceanographic field campaigns.

Data Availability: All data used in this study are available on the project's GitHub repository (https://github.com/JosephLWalker96/underwater-object-detection).

Acknowledgment

The authors would like to thank NVIDIA Corporation for the donation of the Titan V GPU used for this research.

References

[1] O. Pizarro and H. Singh, "Toward large-area mosaicing for underwater scientific applications," *IEEE J. Ocean. Eng.*, vol. 28, no. 4, pp. 651–672, Oct. 2003, doi: 10.1109/JOE.2003.819154.

[2] N. R. Gracias, S. van der Zwaan, A. Bernardino, and J. Santos-Victor, "Mosaic-based navigation for autonomous underwater vehicles," *IEEE J. Ocean. Eng.*, vol. 28, no. 4, pp. 609–624, Oct. 2003, doi: 10.1109/JOE.2003.819156.

[3] N. Barrett, J. Seiler, T. Anderson, S. Williams, S. Nichol, and S. Nicole Hill, "Autonomous underwater vehicle (AUV) for mapping marine biodiversity in coastal and shelf waters: Implications for marine management," in *Proc. OCEANS'10 IEEE SYDNEY*, May 2010, pp. 1–6, doi: 10.1109/OCEANSSYD.2010.5603860.

[4] D. A. Smale et al., "Regional-scale benthic monitoring for ecosystem-based fisheries management (EBFM) using an autonomous underwater vehicle (AUV)," *ICES J. Mar. Sci.*, vol. 69, no. 6, pp. 1108–1118, Jul. 2012, doi: 10.1093/icesjms/fss082.

[5] E. C. Orenstein et al., "The Scripps Plankton Camera system: A framework and platform for in situ microscopy," *Limnol. Oceanogr., Methods*, vol. 18, no. 11, pp. 681–695, 2020, doi: 10.1002/lom3.10394.

[6] R. J. Olson and H. M. Sosik, "A submersible imaging-in-flow instrument to analyze nano-and microplankton: Imaging FlowCytobot," *Limnol. Oceanogr., Methods*, vol. 5, no. 6, pp. 195–203, 2007, doi: 10.4319/lom.2007.5.195.

[7] R. K. Cowen and C. M. Guigand, "In situ ichthyoplankton imaging system (*I*SIIS): System design and preliminary results," *Limnol. Oceanogr, Methods*, vol. 6, no. 2, pp. 126–132, 2008, doi: 10.4319/lom.2008.6.126.

[8] R. W. Campbell, P. L. Roberts, and J. Jaffe, "The Prince William Sound Plankton Camera: A profiling in situ observatory of plankton and particulates," *ICES J. Mar. Sci.*, vol. 77, no. 4, pp. 1440–1455, Jul./Aug. 2020, doi: 10.1093/icesjms/fsaa029.

[9] H. Lu, Y. Li, Y. Zhang, M. Chen, S. Serikawa, and H. Kim, "Underwater optical image processing: A comprehensive review," *Mobile Netw. Appl.*, vol. 22, no. 6, pp. 1204–1211, Dec. 2017, doi: 10.1007/s11036-017-0863-4.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[11] X. Li, M. Shang, H. Qin, and L. Chen, "Fast accurate fish detection and recognition of underwater images with Fast R-CNN," in *Proc. OCEANS 2015 - MTS/IEEE Washington*, Oct. 2015, pp. 1–5, doi: 10.23919/OCEANS.2015.7404464.

[12] S. M. Jaisakthi, P. Mirunalini, and C. Aravindan, "Coral reef annotation and localization using faster R-CNN," in *Proc. 10th Int. Conf. CLEF Assoc.*, 2019.

[13] M. Czub et al., "Deep sea habitats in the chemical warfare dumping areas of the Baltic Sea," *Sci. Total Environ.*, vol. 616–617, pp. 1485–1497, Mar. 2018, doi: 10.1016/j.scitotenv.2017.10.165.

[14] D. L. Rizzini, F. Kallasi, F. Oleari, and S. Caselli, "Investigation of vision-based underwater object detection with multiple datasets," *Int. J. Adv. Robot. Syst.*, vol. 12, no. 77, pp. 1–13, Jun. 2015, doi: 10.5772/60526.

[15] A. Olmos, E. Trucco, and D. Lane, "Automatic man-made object detection with intensity cameras," in *Proc. OCEANS '02 MTS/IEEE*, Oct. 2002, vol. 3, pp. 1555–1561, doi: 10.1109/OCEANS.2002.1191867.

[16] C. M. Bishop, "Pattern recognition and machine learning," in *Information Science and Statistics*. Berlin, Germany: Springer, 2006.

[17] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *Proc. Comput. Vis. – ECCV*, 2010, pp. 213–226.

[18] J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset Shift in Machine Learning*. Cambridge, MA, USA: MIT Press, 2008.

[19] G. Forman, "Quantifying counts and costs via classification," *Data Mining Knowl. Discov.*, vol. 17, no. 2, pp. 164–206, Oct. 2008, doi: 10.1007/s10618-008-0097-y.

[20] E. C. Orenstein, K. M. Kenitz, P. L. D. Roberts, P. J. S. Franks, J. S. Jaffe, and A. D. Barton, "Semi- and fully supervised quantification techniques to improve population estimates from machine classifiers," *Limnol. Oceanogr., Methods*, vol. 18, no. 12, pp. 739–753, 2020, doi: 10.1002/lom3.10399.

[21] O. Beijbom et al., "Quantification in-the-wild: Data-sets and baselines," Nov. 2015. [Online]. Available: http://arxiv.org/abs/1510.04811

[22] P. González, A. Castaño, E. E. Peacock, J. Díez, J. J. Del Coz, and H. M. Sosik, "Automatic plankton quantification using deep features," *J. Plankton Res.*, vol. 41, no. 4, pp. 449–463, Jul. 2019, doi: 10.1093/plankt/fbz023.

[23] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, Jul. 2019, Art. no. 60, doi: 10.1186/s40537-019-0197-0.

[24] G. Wilson and D. J. Cook, "A survey of unsupervised deep domain adaptation," *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 5, Jul. 2020, Art. no. 51, doi: 10.1145/3400066.

[25] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. 32nd Int. Conf. Mach. Learn.*, Jun. 2015, pp. 1180–1189.

[26] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3723–3732, doi: 10.1109/CVPR.2018.00392.

[27] H.-K. Hsu et al., "Progressive domain adaptation for object detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Mar. 2020, pp. 738–746, doi: 10.1109/WACV45572.2020.9093358.

[28] H. Liu, P. Song, and R. Ding, "WQT and DG-YOLO: Towards domain generalization in underwater object detection," Apr. 2020, Accessed: Jun. 14, 2023. [Online]. Available: http://arxiv.org/abs/2004.06333

[29] H. Liu, P. Song, and R. Ding, "Towards domain generalization in underwater object detection," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2020, pp. 1971–1975, doi: 10.1109/ICIP40778.2020.9191364.

[30] Y. Chen et al., "Achieving domain generalization in underwater object detection by domain mixup and contrastive learning," vol. 528, pp. 20–34, 2023.

[31] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4396–4415, Apr. 2023, doi: 10.1109/TPAMI.2022.3195549.

[32] F. J. Piva, D. De Geus, and G. Dubbelman, "Empirical generalization study: Unsupervised domain adaptation vs. domain generalization methods for semantic segmentation in the wild," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Jan. 2023, pp. 499–508, doi: 10.1109/WACV56688.2023.00057.

[33] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, Oct. 2018, doi: 10.1016/j.neucom.2018.05.083.

[34] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2242–2251.

[35] C. Hahne and A. Aggoun, "PlenoptiCam v1.0: A light-field imaging framework," *IEEE Trans. Image Process.*, vol. 30, pp. 6757–6771, 2021, doi: 10.1109/TIP.2021.3095671.

[36] D. Ratelle, S. Sandin, B. Zgliczynski, C. Edwards, and J. Jaffe, "Design and prototyping of the smart underwater imaging telemeter (SUIT) for embedding environmental data in imaging surveys of benthic communities," Univ. California San Diego, San Diego, CA, USA, 2022.

[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, Apr. 2015, pp. 1–14.

[38] N. Inoue, R. Furuta, T. Yamasaki, and K. Aizawa, "Cross-domain weakly-supervised object detection through progressive domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., Mar. 2018, pp. 5001–5009.

[39] H. Huang, H. Zhou, X. Yang, L. Zhang, L. Qi, and A.-Y. Zang, "Faster R-CNN for marine organisms detection and recognition using data augmentation," *Neurocomputing*, vol. 337, pp. 372–384, Apr. 2019, doi: 10.1016/j.neucom.2019.01.084.

[40] J. Giddens, A. Turchik, W. Goodell, M. Rodriguez, and D. Delaney, "The National Geographic Society deep-sea camera system: A low-cost remote video survey instrument to advance biodiversity observation in the deep ocean," *Front. Mar. Sci.*, vol. 7, Jan. 2021, Art. no. 601411, doi: 10.3389/fmars.2020.601411.

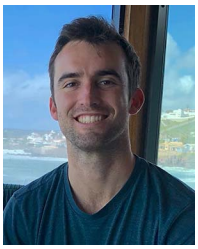
[41] A. M. Friedlander et al., "Marine biodiversity from zero to a thousand meters at Clipperton Atoll (Île de La Passion), Tropical Eastern Pacific," *PeerJ*, vol. 7, Jul. 2019, Art. no. e7279, doi: 10.7717/peerj.7279.

[42] J. Huang, D. Guan, A. Xiao, and S. Lu, "FSDR: Frequency space domain randomization for domain generalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 6887–6898, doi: 10.1109/CVPR46437.2021.00682.

[43] S. Zakharov, W. Kehl, and S. Ilic, "DeceptionNet: Network-driven domain randomization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 532–541, doi: 10.1109/ICCV.2019.00062.

[44] X. Yue, Y. Zhang, S. Zhao, A. Sangiovanni-Vincentelli, K. Keutzer, and B. Gong, "Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 2100–2110, doi: 10.1109/ICCV.2019.00219.

[45] Z. Li, K. Kamnitsas, and B. Glocker, "Overfitting of neural nets under class imbalance: Analysis and improvements for segmentation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019* (Lecture Notes in Computer Science), D. Shen, et al., Eds. Berlin, Germany: Springer, 2019, pp. 402–410, doi: 10.1007/978-3-030-32248-9_45.

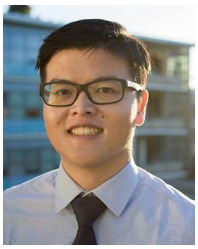


Joseph L. Walker received the Ph.D. degree in applied ocean science from Scripps Institution of Oceanography (SIO), University of California San Diego, San Diego, CA, USA, in 2023.

He is currently a Schmidt AI in Science postdoctoral fellow at the Machine Listening Lab with SIO. He is currently working on developing real-time acoustic processing capabilities for low-power autonomous underwater vehicles. His research focuses on the development of tools for signal detection and classification within underwater image and acoustic data sets.

Zheng Zeng received the B.S. degree in computer engineering in 2021 from the University of California, San Diego, CA, USA, where he is currently working toward the M.S. degree in electrical and computer engineering with a focus on machine learning and data science.

His research interests include deep learning applications in both vision and acoustic tasks.



Chengchen L. Wu received the B.S. degree in data science and the M.S. degree in computer science from the University of California San Diego, San Diego, CA, USA, in 2022 and 2024, respectively.

His research interests include underwater image processing and marine information processing.

Jules S. Jaffe received the B.A. degree in physics from the State University of New York at Buffalo, Buffalo, NY, USA, in 1973, the M.S. degree in biomedical information science from the Georgia Institute of Technology, Atlanta, GA, USA, in 1974, and the Ph.D. degree in biophysics from the University of California, Berkeley, Berkeley, CA, USA, in 1982.

He was a consultant in Silicon Valley for several years, after which he joined the Woods Hole Oceanographic Institution, Woods Hole, MA, USA. He became an Assistant Research Oceanographer with Scripps Institution of Oceanography/UC San Diego (SIO/UCSD), in 1988 and retired as a Full Research Oceanographer in 2020. He was an H. Burr Steinbuck Visiting Scholar with the Woods Hole Oceanographic Institution in 2003 and a visiting Miller Professor with the UC Berkeley in 2006. He is currently a Research Scientist Emeritus with SIO/UCSD. His lab specializes in the development of underwater technologies that range from miniature vehicles to underwater optical and acoustical instruments.

Dr. Jaffe is a Fellow of the Acoustical Society. He was an Associate Editor for the IEEE JOURNAL OF OCEANIC ENGINEERING and the Editor-in-Chief of *Methods in Oceanography* (Elsevier). He was the recipient of the National Science Foundation Creativity Award and the Best Paper Award at the International Ocean Optics Meeting in 2012. His 2002 article on laser sheet microscopy was cited by *Nature* magazine as a "Milestone in Microscopy." His Autonomous Underwater Explorer was featured in the London Museum of Science Exhibit, "Driverless: Who is in Control?" He enjoys public outreach and has been a guest on *CNN*, *ABC*, *NBC*, *Fox News*, and *PBS* to assist the public in understanding underwater technologies in relation to search.

Kaitlin E. Frasier received the Ph.D. degree in biological oceanography from Scripps Institution of Oceanography, University of California San Diego (UC San Diego), San Diego, CA, USA, in 2015.

She is currently an Associate Research Scientist with the Marine Physical Laboratory at Scripps Institution of Oceanography, UC San Diego. Her research interests include autonomous sensing in marine environments and machine learning.

Stuart S. Sandin received the Ph.D. degree in ecology and evolutionary biology from Princeton University, Princeton, NJ, USA, in 2002.

He is currently a Professor and the Director of the Center for Marine Biodiversity and Conservation with Scripps Institution of Oceanography, University of California San Diego, San Diego, CA, USA. His research interests include community ecology, specifically linked with the patterns of organization and change in coral reef ecosystems.