STARS: Zero-shot Sim-to-Real Transfer for Segmentation of Shipwrecks in Sonar Imagery

Advaith Venkatramanan Sethuraman
advaiths@umich.edu
Katherine A. Skinner
kskin@umich.edu

Department of Robotics
University of Michigan
Ann Arbor, Michigan, USA

Abstract

In this paper, we address the problem of sim-to-real transfer for object segmentation when there is no access to real examples of an object of interest during training, i.e. *zero-shot sim-to-real transfer for segmentation*. We focus on the application of shipwreck segmentation in side scan sonar imagery. Our novel segmentation network, STARS, addresses this challenge by fusing a predicted deformation field and anomaly volume, allowing it to generalize better to real sonar images and achieve more effective *zero-shot sim-to-real transfer* for image segmentation. We evaluate the sim-to-real transfer capabilities of our method on a real, expert-labeled side scan sonar dataset of shipwrecks collected from field work surveys with an autonomous underwater vehicle (AUV). STARS is trained entirely in simulation and performs zero-shot shipwreck segmentation with no additional fine-tuning on real data. Our method provides a significant **20%** increase in segmentation performance for the targeted shipwreck class compared to the best baseline.

1 Introduction

Autonomous underwater vehicles (AUVs) equipped with sonar systems have demonstrated potential to carry out efficient, cost-effective large area surveys of marine environments to locate submerged objects, such as shipwrecks, downed airplanes, and underwater mines. Still, identification of submerged objects from sonar data is currently a manual process relying on expert knowledge for interpretation, which can take many months to complete.

On land, deep learning algorithms are able to leverage large, publicly available training datasets to achieve impressive performance on automated computer vision tasks such as object detection and segmentation from optical imagery [9, 40]. Unfortunately, domain-specific barriers like security concerns, cost of collection, and difficulty labeling sonar data prevent the creation of large, publicly available training datasets for side scan sonar. Furthermore, even when data is available, real examples of specific target objects may not be present due to limited frequency of appearance. This motivates the need to develop new machine learning methods capable of performing accurate object detection and segmentation in spite of significant data restrictions for network training.

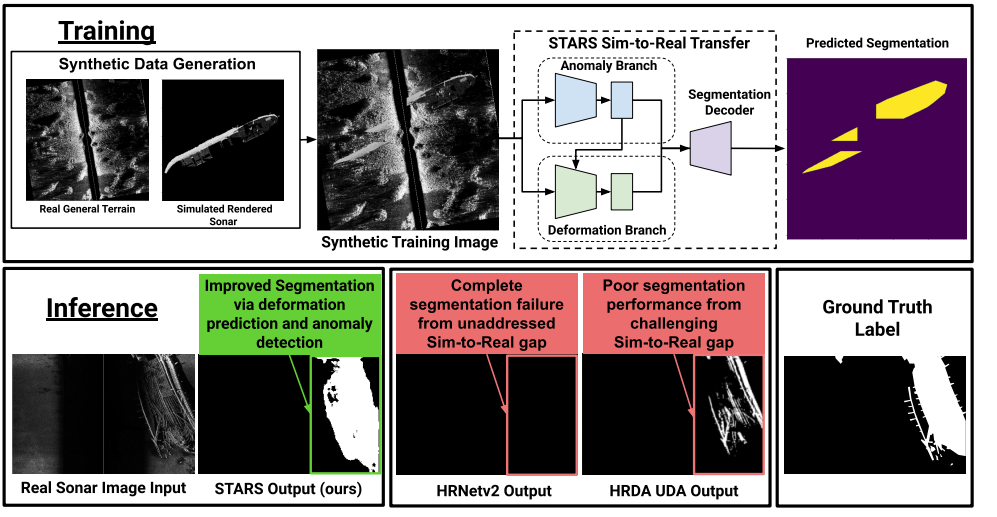

Figure 1: The primary problem this work addresses is the sim-to-real gap between synthetic training data and real testing data for side scan sonar images. We find that naively training segmentation (HRNetv2) and domain adaptation models (HRDA) on synthetic data does not transfer effectively to diverse and unstructured real sonar data. Our method STARS addresses this research gap with the novel fusion of a predicted deformation field and anomaly volume. STARS takes a step towards more accurate segmentation in never-before-seen environments.

Simulation is useful for generating data that is difficult and expensive to collect and it can be leveraged to generate additional examples of rare objects or events. However, as shown in Fig. 1, naively training state-of-the-art supervised segmentation methods on simulated data fails due to the *sim-to-real* gap between the synthetic data and real data, which hinders model performance at test time [27, 34]. Techniques like domain adaptation have been developed to improve the performance of networks trained in simulation and tested on real data. Still, many domain adaptation methods require examples of objects of interest (i.e. shipwrecks) in the target domain.

In this paper, we address the problem of sim-to-real transfer for object segmentation when there is no access to real examples of an object of interest during training, i.e. *zero-shot sim-to-real transfer for segmentation*. Our main contributions are:

- We introduce a novel zero-shot sim-to-real segmentation framework, **STARS**: *Sonar Target Anomaly Recognition for Segmentation* [1], that exploits cues from anomaly detection and deformation prediction for improved segmentation performance.
- We propose a simple but effective synthetic side scan sonar image generation method that produces diverse and randomized debris fields to support training for sim-to-real transfer for side scan sonar imagery.
- We present extensive qualitative and quantitative evaluation on a real side scan sonar dataset consisting of 220 scans of 14 distinct shipwreck sites collected with an AUV in Thunder Bay National Marine Sanctuary (TBNMS), which includes detailed and high-resolution labels generated by an expert marine archaeologist.

Through experiments, we show that our STARS provides a significant **20%** improvement in segmentation performance for the shipwreck class over state-of-the-art baselines.

[1] https://umfieldrobotics.github.io/STARS.github.io/

2 Related Work

2.1 Object Detection and Segmentation in Sonar Imagery

Recent work on object detection and segmentation from sonar imagery has leveraged machine learning. The majority of work in object detection for sonar imagery involves fine-tuning existing object detection algorithms on real, labeled sonar imagery [7, 9, 22, 37]. The unique nature of side scan sonar imagery has also motivated the development of specialized network architectures [7, 36], but these networks still require access to labeled datasets. Due to data restrictions for target objects, our work instead aims to perform segmentation of shipwreck sites without access to real labeled examples during training.

Recent work has focused on leveraging simulation to generate realistic side scan sonar imagery to overcome data limitations. The image formation process for side scan sonars can be approximated using ray-tracing and does not require access to real, labeled sonar data [16]. However, the lack of diverse terrains and consideration of environmental factors leads to a *sim-to-real* gap between the synthetic sonar data and real sonar data that can hinder model performance on real data at test time [27, 28, 34]. Current approaches to improving realism in synthetic sonar data generation include Generative Adversarial Networks [15, 19], style transfer [20, 21], ray-tracing [16], and image composition techniques [6, 11, 32]. For example, [11] uses style transfer networks and cutting and pasting to produce synthetic side scan sonar images. Although effective, this method also requires real examples of objects in side scan sonar imagery to train the style transfer network, whereas our method does not require real examples of our object of interest. Our work leverages image synthesis techniques and physics-based rendering concepts but adapts them for shipwreck segmentation in side scan sonar imagery. We also contribute a novel method for simulating shipwreck debris using deformation fields, which we use as a proxy learning task for our method.

2.2 Domain Adaptation

While simulation allows us to generate datasets with more examples of our target class, models trained on simulated or synthetic data will still suffer at test time due to the sim-to-real gap. Domain adaptation methods have been proposed to modify a network trained in a source domain (e.g. simulated data) to perform inference in a target domain (e.g. real data) [10, 13, 14]. Relevant to this work is Unsupervised Domain Adaptation (UDA), where there are no object labels in the target domain [14, 35]. Recently, Hoyer et. al proposed HRDA, a state-of-the-art UDA method for adaptation across different object sizes [14]. Still, unsupervised domain adaptation requires a representative set of target objects to be present in both synthetic and real data for training, which may not be possible for rare targets.

2.3 Zero-shot Segmentation

Recently, zero-shot segmentation has been proposed to perform segmentation of novel objects at test time that were unseen during training [1, 12, 18]. Zero-shot segmentation generalizes segmentation learned on seen classes from a training dataset to unseen classes during test time. However, in our situation, we have access to labeled objects in a specific *domain* and wish to generalize to another domain. The classes remain the same. A relatively new field called zero-shot unsupervised domain adaptation explores the problem of adapting a model learned in a specific domain to another domain without any examples from the target

domain. Models like Prompt-driven Zero-shot Domain Adaptation (PODA) are prompted by text descriptions of the target domain and are able to adapt features learned in the source domain [8]. Our method does not require prompts or text description, and instead leverages synthetic data for sim-to-real transfer. Thus, our method addresses a unique problem of zero-shot sim-to-real transfer for segmentation.

2.4 Deep Anomaly Detection and Salient Object Detection

Salient object detection (SOD) aims to segment the primary object in an image from the scene. Although SOD networks like InSPyReNet do not explicitly consider sim-to-real gaps, they are designed to perform segmentation on novel objects and scenes [17]. Prior works on anomaly detection also have a natural utility for this problem: they can generalize to the diverse set of anomalies encountered during test-time. Unsupervised anomaly detection methods train only on a normal set of images and identify anomalies in test images by producing an anomaly score at the image or pixel level [23, 25, 39]. DeSTSeg is an anomaly detection method that uses a student-teacher paradigm to segment anomalous regions of images [39]. DeSTSeg first generates synthetic anomalies then trains a supervised segmentation network. However, the synthetic anomalies are restricted to small cracks and defects often found in an industrial setting, not the complex shapes of shipwrecks as in our problem. The current state-of-the-art method, PatchCore, uses a memory bank to store normal features [25]. Finally, PatchCore provides an anomaly score by reporting a distance metric between test features and a subsampled memory bank. Our method, STARS, leverages anomalous features but refines them in a manner relevant to the segmentation task. We show that this ability to identify anomalous features during test-time is useful for minimizing the sim-to-real gap in a segmentation task.

3 Technical Approach

STARS has two main components: synthetic data generation (Fig. 2) and network development and training (Fig. 3).

3.1 Synthetic Sonar Image Generation

We generate synthetic side scan sonar data for training using physics-based rendering. We extended BLAINDER to develop a side scan sonar simulation system within the Blender graphics environment [3, 24], modeled after the side scan sonar used for real experiments. Our simulation environment consists of 3D meshes of ships downloaded from TurboSquid [33]. We randomized acoustic reflectance, ship location, ship scale, ship orientation.

To calculate the intensity of the sonar return for a given ray r_i, we account for the propagation of sound in decibels underwater using the SONAR equation [26]:

$$RI_i = SL_i - 2TL_i + TS_i \tag{1}$$

where RI_i is the intensity of sound returned to the sensor, SL_i is the source level of the emitted acoustic pulse, and TL_i is the transmission loss from sound propagation through water such that $TL_i = 10\log_{10}(d_i)$, where d_i is the distance of the return. TS is the target strength, or sonar cross section of the object imaged. We are able to produce both side scan sonar images and segmentation masks from our simulation environment.

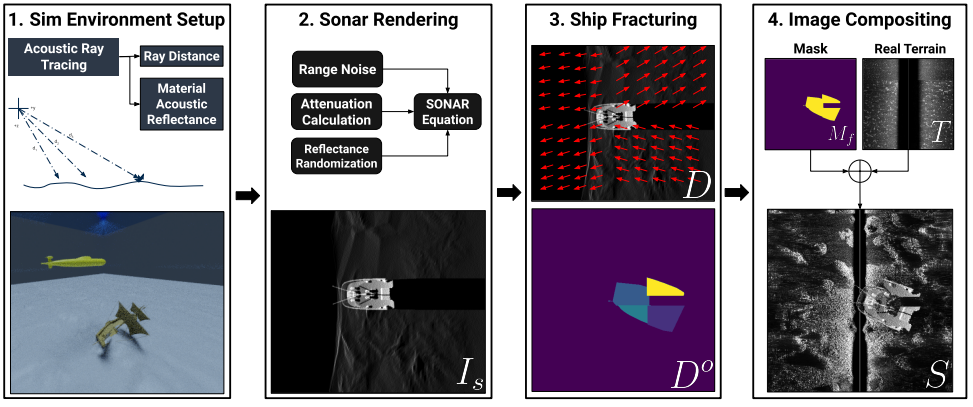

Figure 2: Shipwrecks are generated in a graphics environment then rendered into sonar images based on the SONAR equation. Then, they are fractured using deformation fields and composited onto real terrain images.

We introduce a novel method for simulating diverse debris fields in side scan sonar images. This process is illustrated in Fig. 2. A deformation field, or optical flow field, dictates how pixels are translated in relation to the original image. The origin is at the center of the ship and the ship is split into four pieces. The field is randomly generated, but all pixels within a given quadrant experience the same field. The field $D \in \mathbb{R}^{H\times W\times 2}$ is parameterized by magnitude and direction (r,θ) for each pixel. The magnitude $r \in [0, r_{max}]$ is divided into $N_r = 10$ discrete values, whereas the direction $\theta \in [0, 2\pi]$ is divided into $N_\theta = 20$ discrete values. This is used to create a one-hot representation of the magnitude and direction $D^{mag} \in \mathbb{R}^{H\times W\times N_r}, D^{ang} \in \mathbb{R}^{H\times W\times N_\theta}$. These are concatenated to create $D^o = D^{mag} \oplus D^{ang} \in \mathbb{R}^{H\times W\times D_{def}}$, where $D_{def} = N_r + N_\theta = 30$. Then, the field is applied to split the ship image I_s in an arbitrary pattern and produce the fractured image I_f. Let $I_s(u,v)$ represent the image value at pixel location (u,v) and $(r,\theta) = D(u,v)$ represent the deformation field at the same location, then

$$I_f(u + rcos(\theta), v + rsin(\theta)) = I_s(u,v) \tag{2}$$

The synthetic segmentation mask M is also fractured with D to produce fractured segmentation map M_f. The final synthetic image S is created by compositing onto a real terrain image T. It is reasonable to assume access to unlabeled, real terrain images T because they are publicly available and collected during routine surveys of bodies of water.

3.2 Network Architecture

Our key insight with STARS lies in identifying the root causes of our sim-to-real gap and explicitly developing architectural inductive biases that mitigate them. First, we observe that any object underwater (especially shipwrecks) can experience extreme environmental degradation and destruction, resulting in fracturing and deformation. To address this, we introduce a deformation prediction proxy task supervised by our synthetic data that makes our network aware of fracturing and the spatial relationship between fragments and the whole ship. Secondly, we find that the appearance of real shipwrecks may differ from synthetic training data

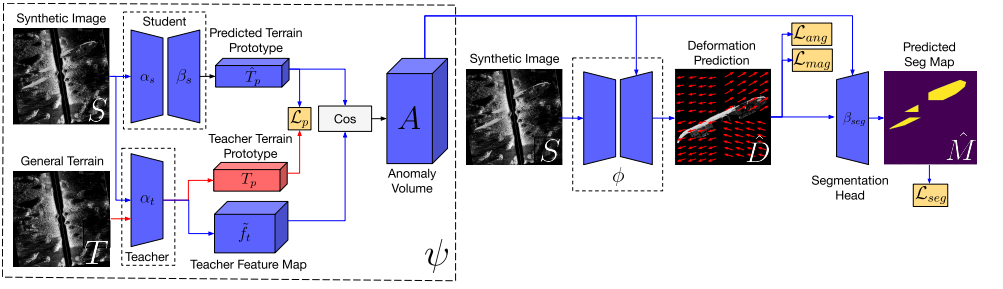

Figure 3: STARS network architecture. The anomaly prediction network ψ produces anomaly volume A by taking the cosine distance between the student's predicted terrain prototype $\hat{T}_p$ and the teacher's feature map of the synthetic image $\tilde{f}_t$. Then, the anomaly volume A is used within the deformation prediction network ϕ to produce deformation field $\hat{D}$. The outputs of ϕ and ψ are fused by β_{seg} to produce a segmentation prediction. During inference, the teacher produces $\tilde{f}_t$ and the anomaly volume can be computed. All modules in blue are used for training and inference, while modules in red are needed only for training. Best viewed in color and zoomed in.

due to their high natural diversity. To address this challenge, our anomaly volume informs STARS of generally anomalous features in the scene that may aid the segmentation task.

The detailed network architecture for STARS is shown in Fig. 3. Given an image $I \in \mathbb{R}^{H\times W\times 3}$, we wish to produce a segmentation map $\hat{M} \in \mathbb{R}^{H\times W\times 1}$ that segments two classes: $\{terrain, shipwreck\}$. We use a deformation prediction network $\phi : \mathbb{R}^{H\times W\times 3} \rightarrow \mathbb{R}^{H\times W\times D_{def}}$ and anomaly prediction network $\psi : \mathbb{R}^{H\times W\times 3} \rightarrow \mathbb{R}^{H\times W\times D_l}$. The outputs are fused by a lightweight decoder β_{seg} to produce segmentation prediction $\hat{M}$. Note that although deformation prediction, optical flow estimation [6, 30] and anomaly detection [23, 25] have been independently explored in the vision community, they have not been combined in this manner to produce segmentation outputs in a zero-shot sim-to-real transfer setting for sonar data. Please refer to supplementary material for layer-level network architecture details.

3.2.1 Anomaly Prediction Network ψ

Inspired by [39], our method uses a student-teacher paradigm to produce an anomaly volume. [39] detects anomalies by computing the cosine distance between student and teacher features in the same spatial location. Our student network instead produces a single terrain prototype $\mathbb{R}^{1\times 1\times C}$ that summarizes the general terrain in an image. By forcing the network to summarize the entire terrain with a single terrain prototype, we create a natural bottleneck for more efficient representation of terrain patterns. Then, our method computes the cosine distance between a terrain prototype and the teacher's entire feature map.

Our anomaly prediction network ψ is composed of student encoder α_s, student decoder β_s, and teacher encoder α_t. The teacher network is frozen and does not receive gradient updates. First, a synthetic image S is passed through the student encoder α_s and decoder β_s. The decoder produces feature maps $f_s(i), i \in [1, D_l]$ at varying resolutions. Then, a global average pooling (GAP) layer is used on the spatial dimensions of $f_s(i)$ to produce a student terrain prototype $\hat{T}_p(i)$. This terrain prototype is supervised by the teacher encoder α_t. Note that the teacher is fed the regular terrain image T *without any objects*. This is easily taken from the synthetic image generation phase. $\alpha_t(T)$ produces feature maps $f_t(i), i \in [1, D_l]$. Similarly, global average pooling is used to construct a teacher terrain prototype,

$T_p(i)$. Finally, an MSE loss is used to supervise the student:

$$\mathcal{L}_p = \sum_{i=1}^{D_l} ||\hat{T}_p(i) - T_p(i)||_2^2 \tag{3}$$

In order to compute the anomaly map $A(i)$ at a given depth i, the teacher encoder is fed the synthetic image S. This produces a teacher feature map $\tilde{f}_t(i)$. The cosine distance between each $\tilde{f}_t(i)$ and $\hat{T}_p(i)$ is used as an anomaly score:

$$A(i) = 1 - \frac{\hat{T}_p(i) \cdot \tilde{f}_t(i)}{|\hat{T}_p(i)||\tilde{f}_t(i)|} \tag{4}$$

Intuitively, the terrain prototype will have smaller cosine distance when compared to terrain features but higher distance when compared to anomalous debris features. Since the input to the student is S and supervised by $T_p(i)$, it will learn to ignore any objects and focus on summarizing the terrain effectively. Note that T_p is not needed for inference and is only used for supervised training of the student.

3.2.2 Deformation Prediction Network ψ

Forcing STARS to predict the deformation field $\hat{D}$ that turned intact image I_s into fractured image I_f implicitly teaches STARS to identify parts of a broken ship, then learn the spatial relation between the different pieces. The decoder also concatenates the anomaly maps at varying resolutions $A(i), i \in [1, D_l]$ to the skip connections. This allows the anomaly prediction to aid the deformation prediction.

We pose deformation prediction as a classification task, with discretized magnitude and phase components. $\hat{D} \in \mathbb{R}^{H \times W \times D_{def}}$ is composed of a magnitude $\hat{D}^{mag} \in \mathbb{R}^{H \times W \times D_r}$ and angle $\hat{D}^{ang} \in \mathbb{R}^{H \times W \times D_\theta}$ prediction for each pixel. These predictions are supervised with one-hot ground truth D^o using a cross entropy losses $\mathcal{L}_{mag}$ and $\mathcal{L}_{ang}$.

3.2.3 Segmentation Decoder β_{seg}

The two branches are fused with a 1×1 convolutional decoder β_{seg}. All layers $A(i)$ are bilinearly interpolated to a resolution of (H, W) and concatenated along the channel dimension to produce feature volume $F \in \mathbb{R}^{H \times W \times D_l}$. Finally, the segmentation map is given by $\hat{M} = \beta_{seg}(F)$. The segmentation output is supervised by a binary cross entropy loss, $\mathcal{L}_{seg}$. The final loss becomes

$$\mathcal{L} = \mathcal{L}_{mag} + \mathcal{L}_{ang} + \mathcal{L}_p + \mathcal{L}_{seg} \tag{5}$$

4 Experiments & Results

4.1 Datasets

To collect a real dataset, field tests were carried out using an Iver-3 AUV in Thunder Bay National Marine Sanctuary (TBNMS) in Lake Huron, Michigan. TBNMS has almost 100 known shipwreck sites of varying size, type, and wreck condition, making this an ideal location for dataset collection to validate the proposed approach for the application of shipwreck segmentation [30]. Surveys were conducted over a two week period and produced images

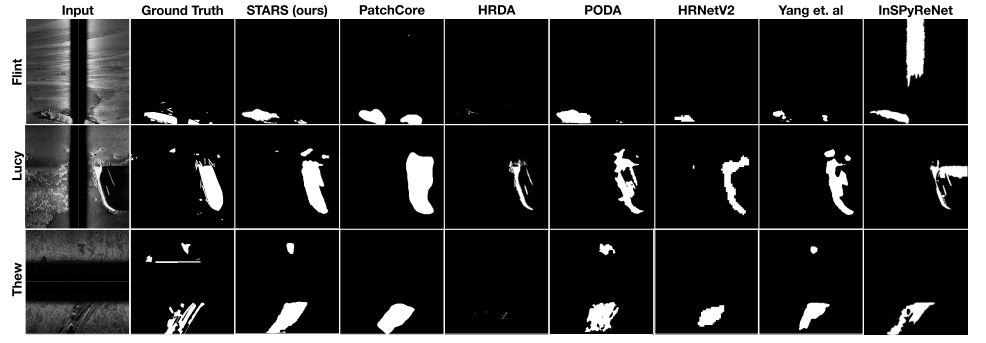

Figure 4: Selected qualitative results from our method compared to baselines. It should be noted that some methods have a tendency to inaccurately over-segment or fail to segment debris from shipwrecks, resulting in lower performance. However, STARS consistently produces more accurate segmentation outputs. Best viewed zoomed in.

from 13 distinct shipwreck sites and 1 artificial reef that serves as a potential distractor object. The field work surveys resulted in 220 scans of terrain and shipwrecks. The raw scans are of variable height, but maintain a horizontal resolution of 1728 pixels. We take a sliding window of size $(1728, 1728)$ and generate images with vertical stride of 100 pixels. The resulting dataset has 861 images, including 312 images of terrain and 549 images of shipwrecks. We do not use the shipwreck images for training, and instead withhold them as a test set for evaluation only.

Labeling side scan sonar images can be challenging because of shadowing effects, view-dependent acoustic artifacts, and complex debris fields. To ensure accurate ground truth, each shipwreck site was labeled with the help of an expert marine archaeologist from TBNMS who has extensively studied and visually confirmed these wrecks by scuba diving.

We wish to emphasize that collecting side scan sonar imagery is extremely expensive and time consuming, even with autonomous systems. We recognize that our dataset, composed of 220 distinct scans, is orders of magnitude smaller than large scale benchmark datasets collected for optical deep learning. However, given the diversity of our dataset, we believe it is representative of sites seen during real deployment. Our TBNMS dataset provides useful insight into how well networks can generalize to real, challenging, and unstructured sonar surveys. Please refer to supplementary material for sample data from our field surveys.

For simulated data, we create a dataset of 10,000 images of resolution $(1728, 1728)$ using our generation process outlined in Section 3.1. An 80/20 split was used to create training and validation sets respectively.

4.2 Baselines

We chose a variety of baselines that perform segmentation while respecting the restriction on real training data. We evaluate STARS against PatchCore unsupervised anomaly detection [25], HRDA unsupervised domain adaptation [14], PODA zero-shot domain adaptation, [8], HRNetv2+OCR [29, 38] as a baseline method for no sim-to-real transfer, Yang et. al as a side scan sonar segmentation method [36], and InSPyReNet [17] as a salient object detection baseline. Please refer to supplementary information for more details of implementation and baselines.

4.3 Segmentation Performance

Qualitative results from selected sites are shown in Figure (4). Note that none of the evaluated networks have seen a real shipwreck site during training, making segmentation at novel sites a very challenging task. All experiments use images of resolution (1728, 1728). For quantitative results, we note that there are two classes of interest for our semantic segmentation task: $\{terrain, shipwreck\}$. We use Intersection over Union (IOU) and F1 score to quantify the performance of the methods tested. We report the IOU of the shipwreck and terrain class, but terrain is considered a background class. Given the class imbalance between terrain pixels and shipwreck pixels (10:1 ratio respectively), we use F1 score instead of per-pixel accuracy. The overall segmentation performance of our method is shown in Table (1). The IOU for both classes and F1 Score were calculated for each site then averaged across all sites. Our method consistently outperforms baselines, achieving **0.42 IOU**$_{ship}$. Additional results including detailed per-site segmentation performance are shown in supplementary material.

Table 1: Summarized segmentation performance of our method compared to baselines averaged across 14 sites.

Method	IOU$_{ship}$ ↑	IOU$_{terr}$ ↑	mIOU ↑	F1 Score ↑
PatchCore [ ]	0.28	0.91	0.60	0.43
HRDA [ ]	0.19	0.97	0.58	0.29
PODA [ ]	0.28	0.97	0.63	0.41
HRNetV2 [ ]	0.35	0.97	0.66	0.48
Yang et. al [ ]	0.31	**0.98**	0.65	0.48
Burguera [ ]	0.25	0.97	0.61	0.38
InSPyReNet [ ]	0.33	0.97	0.65	0.45
STARS (ours)	**0.42**	**0.98**	**0.70**	**0.55**

4.4 Ablation Studies

Our method relies on synthetic data generation and a specialized network architecture to exploit the generated data. To investigate the importance of each factor, we ablate the deformation branch (DB), anomaly branch (AB), composition using real terrain (RT), and ship fracturing (SF) in Table (2). After adding the deformation branch (DB) and anomaly branch (AB) in isolation, performance increases by 0.10 and 0.13 IOU$_{ship}$, respectively. With respect to synthetic data generation, we found that real terrain (RT) is essential to improved performance, and ship fracturing (SF) without real terrain decreases performance.

Table 2: Network and Synthetic Data Ablation Studies

Model	DB	AB	RT	SF	IOU$_{ship}$ ↑	F1 Score ↑
No DB, No AB	✗	✗	✓	✓	0.26	0.38
AB Only	✗	✓	✓	✓	0.39	0.45
DB Only	✓	✗	✓	✓	0.36	0.45
No RT, No SF	✓	✓	✗	✗	0.24	0.33
SF Only	✓	✓	✗	✓	0.12	0.19
RT Only	✓	✓	✓	✗	0.36	0.49
STARS (ours)	✓	✓	✓	✓	**0.42**	**0.55**

5 Conclusion & Future Work

We propose a synthetic data generation framework and novel network architecture termed STARS for zero-shot sim-to-real segmentation of shipwrecks in side scan sonar imagery. Our model shows a significant **20%** improvement in IOU_{ship} from state-of-the-art baselines with access to the same data. Both UDA and zero-shot UDA methods performed reasonably well, but struggled to adapt features learned from synthetic data to the target domain. Our work has the potential to significantly reduce the cost and effort needed to train deep learning models for sonar segmentation by removing reliance on real, labeled data.

Future work will focus on improving performance on hard sites, including challenging

distractor sites, such as the artificial reef site. We note that although our method is applied to shipwreck detection, many objects underwater experience the same environmental degradation and fracturing process that our network is designed to handle. This motivates future work in expanding to multi-class segmentation involving other underwater objects.

6 Acknowledgements

We would like to thank Professor Timothy Havens, Professor Guy Meadows, Jamey Anderson and Chris Pinnow of the Great Lakes Research Center at Michigan Technological University for IVER-3 AUV data collection, and Thunder Bay National Marine Sanctuary for supporting field experiments. This work is supported by a University of Michigan Robotics Department Fellowship and by the NOAA Ocean Exploration program under Award #NA21OAR0110196.

References

[1] Maxime Bucher, Tuan-Hung VU, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[2] Antoni Burguera and Francisco Bonin-Font. On-line multi-class segmentation of side-scan sonar imagery using an autonomous underwater vehicle. *Journal of Marine Science and Engineering*, 8(8):557, 2020.

[3] Blender Online Community. Blender - a 3d modelling and rendering package, 2018. URL http://www.blender.org.

[4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, June 2016.

[5] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 764–773, 2017.

[6] D. Dwibedi, I. Misra, and M. Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1310–1319, 2017.

[7] Dylan Einsidler, Manhar Dhanak, and Pierre-Philippe Beaujean. A deep learning approach to target recognition in side-scan sonar imagery. In *OCEANS 2018 MTS/IEEE Charleston*, pages 1–4, 2018.

[8] Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc, Patrick Pérez, and Raoul de Charette. Pøda: Prompt-driven zero-shot domain adaptation. In *ICCV*, 2023.

[9] Peter Feldens, Alexander Darr, Agata Feldens, and Franz Tauber. Detection of boulders in side scan sonar mosaics by a neural network. *Geosciences*, 9(4):159, 2019.

[10] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(1):2096–2030, Jan 2016.

[11] Qiang Ge, Fengxue Ruan, Baojun Qiao, Qian Zhang, Xianyu Zuo, and Lanxue Dang. Side-scan sonar image classification based on style transfer and pre-trained convolutional neural networks. *Electronics*, 10(15):1823, 2021.

[12] Zhangxuan Gu, Siyuan Zhou, Li Niu, Zihan Zhao, and Liqing Zhang. From pixel to patch: Synthesize context-aware features for zero-shot semantic segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–15, 2022.

[13] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. CyCADA: Cycle-consistent adversarial domain adaptation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 1989–1998. PMLR, 10–15 Jul 2018.

[14] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, page 372–391, Berlin, Heidelberg, 2022.

[15] Yifan Jiang, Bonhwa Ku, Wanjin Kim, and Hanseok Ko. Side-scan sonar image synthesis based on generative adversarial network for images in multiple frequencies. *IEEE Geoscience and Remote Sensing Letters*, 18(9):1505–1509, 2021.

[16] L.M. Linnett J.M. Bell. Simulation and analysis of synthetic sidescan sonar images. *IEE Proc. - Radar, Sonar and Navigation*, 144:219–226(7), August 1997.

[17] Taehun Kim, Kunhee Kim, Joonyeong Lee, Dongmin Cha, Jiho Lee, and Daijin Kim. Revisiting image pyramid structure for high resolution salient object detection. In *Proceedings of the Asian Conference on Computer Vision*, pages 108–124, 2022.

[18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[19] Eon-ho Lee, Byungjae Park, Myung-Hwan Jeon, Hyesu Jang, Ayoung Kim, and Sejin Lee. Data augmentation using image translation for underwater sonar image segmentation. *PLOS ONE*, 17(8):1–15, 08 2022.

[20] Sejin Lee, Byungjae Park, and Ayoung Kim. Deep learning from shallow dives: Sonar image generation and training for underwater object detection. *CoRR*, abs/1810.07990, 2018.

[21] Sejin Lee, Byungjae Park, and Ayoung Kim. Deep learning based object detection via style-transferred underwater sonar images. *IFAC-PapersOnLine*, 52(21):152–155, 2019.

[22] Nandeeka Nayak, Makoto Nara, Timmy Gambin, Zoë Wood, and Christopher M. Clark. Machine learning techniques for auv side-scan sonar data feature extraction as applied to intelligent search for underwater archaeological sites. In *Field and Service Robotics*, pages 219–233, Singapore, 2021.

[23] Tal Reiss, Niv Cohen, Liron Bergman, and Yedid Hoshen. Panda: Adapting pretrained features for anomaly detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2806–2814, 2021.

[24] Stefan Reitmann, Lorenzo Neumann, and Bernhard Jung. Blainder—a blender ai add-on for generation of semantically labeled depth-sensing data. *Sensors*, 21(6), 2021.

[25] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14318–14328, June 2022.

[26] James V. Sanders and Alan B. Coppens. An introduction to the sonar equations with applications. 1976.

[27] Advaith Sethuraman and Katherine A. Skinner. Towards sim2real for shipwreck detection in side scan sonar imagery. *3rd Workshop on Closing the Reality Gap in Sim2Real Transfer, Robotics: Science and Systems*, 2022.

[28] Jane Shin, Shi Chang, Matthew J. Bays, Joshua Weaver, Thomas A. Wettergren, and Silvia Ferrari. Synthetic sonar image simulation with various seabed conditions for automatic target recognition. In *OCEANS 2022, Hampton Roads*, pages 1–8, 2022.

[29] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5686–5696, 2019.

[30] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 402–419, 2020.

[31] Thunder Bay National Marine Sanctuary. Thunder Bay National Marine Sanctuary. https://thunderbay.noaa.gov/, Accessed online: 2023.

[32] J. M. Topple and J. A. Fawcett. Minet: Efficient deep learning automatic target recognition for small autonomous vehicles. *IEEE Geoscience and Remote Sensing Letters*, 18(6):1014–1018, 2021.

[33] TURBOSQUID. 3D Models for Professionals. https://www.turbosquid.com, Accessed online: 2022.

[34] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2962–2971, Los Alamitos, CA, USA, Jul 2017.

[35] B. Xie, S. Li, M. Li, C. Liu, G. Huang, and G. Wang. Sepico: Semantic-guided pixel contrast for domain adaptive semantic segmentation. *IEEE Transactions on Pattern Analysis &; Machine Intelligence*, (01):1–17, Jan 2023.

[36] Dianyu Yang, Chensheng Cheng, Can Wang, Guang Pan, and Feihu Zhang. Side-scan sonar image segmentation based on multi-channel cnn for auv navigation. *Frontiers in Neurorobotics*, 16:928206, 2022.

[37] Dianyu Yang, Can Wang, Chensheng Cheng, Guang Pan, and Feihu Zhang. Semantic segmentation of side-scan sonar images with few samples. *Electronics*, 11(19), 2022.

[38] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, page 173–190, Berlin, Heidelberg, 2020.

[39] Xuan Zhang, Shiyu Li, Xi Li, Ping Huang, Jiulong Shan, and Ting Chen. Destseg: Segmentation guided denoising student-teacher for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3914–3923, 2023.

[40] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5122–5130, 2017.