Real-time Underwater 3D Reconstruction Using Global Context and Active Labeling

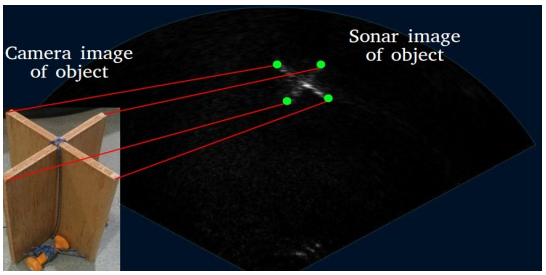
Robert DeBortoli*, Austin Nicolai*, Fuxin Li, Geoffrey A. Hollinger

***Abstract*— In this work we develop a novel framework that enables the real-time 3D reconstruction of underwater environments using features from 2D sonar images. Due to noisy and low-resolution imagery as compared with standard cameras, automatic feature extractors for sonar images are not reliable in many scenarios. Thus, a human often needs to hand-select features in sonar imagery for environment reconstructions. Given the high data capture rates of standard imaging sonars (on the order of 20Hz), hand-annotating the features in every frame cannot be done in real-time. To address this we use a Convolutional Neural Network (CNN) that analyzes incoming imagery in real-time and proposes only a small subset of high-quality frames to the user for feature annotation. We demonstrate that our approach provides real-time reconstruction capability without loss in classification performance on datasets captured onboard our underwater vehicle while operating in a variety of environments.**

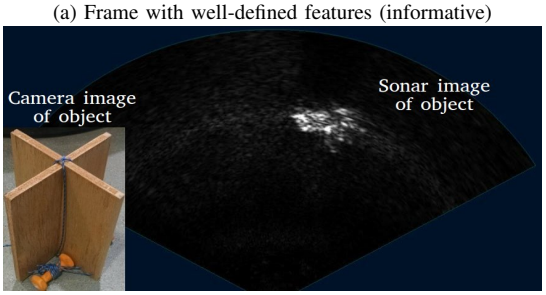
I. INTRODUCTION

The 3D reconstruction of underwater environments has proven useful in a variety of applications, including ship hull inspection and underwater surveying tasks [1], [2]. Oftentimes in these and similar applications, 2D imaging sonars are the choice for exteroceptive sensing, as standard cameras have extremely limited visibility in turbid waters. While sonar has superior range, its imagery is often corrupted by noise as well as the non-diffuse reflection of the acoustic wave off of the object of interest [3]. Multipath returns are a large producer of noise. Such returns occur when the projected beam completes a sequence of reflections other than simply to the object and back to the sonar (e.g., from the sonar to the seafloor, to the object, and then back to the sonar). Due to the increased length of time for the reflected wave to arrive back at the sonar, these reflections are often imaged at a range beyond the object. This can result in an image similar to Fig. 1b, where the horizontal component of the X appears thicker, due to delayed reflections from multipath reflections. Non-diffuse or specular reflection occurs when an object is smooth and fails to reflect the beam back in the direction it was emitted from. An example of this can be seen in Fig. 8i, where only the corners of the triangle are shown and the flat top surface does not return energy back to the sonar.

The substantial amount of noise present in sonar images is problematic for two reasons. First, noise will often corrupt an image so much that analyzing it further would waste time and computational resources. Our method provides a way to spend these resources more efficiently by only providing quality imagery to perception algorithms. Second, noise combined with the low-resolution of sonar imagery hampers the development of an automatic feature extractor for sonar images. For example, feature extractors that use image gradients would incorrectly attempt to extract features from the gradients present in Fig. 1b.



(a) Frame with well-defined features (informative)



(b) Frame with object lacking well-defined features (non-informative)

Fig. 1: Examples of informative (a human can confidently identify features) and non-informative frames while inspecting a target in the shape of an X. The sub-image in both figures is a camera image of the X shaped target that is insonified. In (a) feature correspondences (red) can be made between the camera image and sonar features (green). In our experiments we found on average only about 39% of the captured frames to be informative.

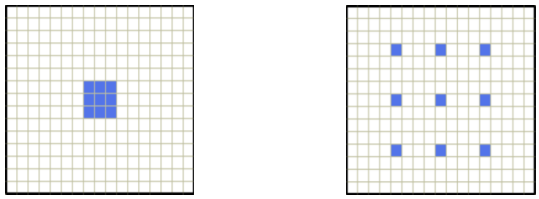
The lack of an automatic feature extractor for sonar images necessitates human annotation of the features for 3D reconstructions [4]. Due to the high data rate of imaging sonars, humans cannot annotate the frames in real-time. We show that naively sub-sampling the data (to allow for real-time annotation) does not result in high-quality reconstructions because noise corrupts many of the images sampled.

The need for human annotation creates a long delay between data collection and environment reconstruction. To eliminate this delay and enable real-time reconstruction, we develop a method to recognize and propose only informative frames to the human for annotation. We define an informative

This work was funded in part by U.S. Department of Energy contract DE-EE-0006816.

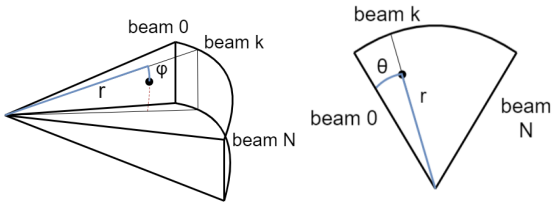
* Robert DeBortoli and Austin Nicolai contributed equally to this work

The authors are with the Collaborative Robotics and Intelligent Systems (CoRIS) Institute, Oregon State University, Corvallis, Oregon 97331 {debortor, nicolaia, fuxin.li, geoff.hollinger}@oregonstate.edu



(a) Dilation rate of 1 (b) Dilation rate of 4

Fig. 2: Example of traditional 3x3 filter (left) and a 3x3 dilated filter (right). The dilated filter uses a larger neighborhood, which compensates for the lack of strong local features in sonar imagery.



(a) Computation of range and elevation angle (b) The acquisition of bearing by emitting multiple beams

Fig. 3: Mapping from Euclidean to polar coordinates.

image as one that contains a set of clear and distinguishing features for environment reconstruction (as seen in Fig. 1a). This automatic identification, which is done in real-time, allows for a small set of frames to be presented to the operator. This set is small enough to be annotated in real-time.

To identify informative images in noisy and low-resolution sonar imagery we use dilated filters (e.g. Fig. 2b) in the atrous convolution architecture, previously used in the computer vision community for image segmentation tasks [5], [6]. Such filters use a large neighborhood of context in analyzing low-resolution and noise-filled sonar images.

In this work we also focus on an approach that is flexible. Given the difficulty in obtaining and labeling sonar imagery, we pay special attention to not overfitting to the objects we train with. This is done by using the context afforded by dilated filters. To demonstrate this, we conduct experiments on imagery of objects not seen in the training set. We thus show that our approach is useful in the mapping or inspecting of objects not recorded in sonar previously.

In summary, we present a novel framework for underwater 3D reconstruction that:

- Automatically selects informative frames for an operator to annotate features, enabling the real-time reconstruction of underwater environments where the features must be manually selected.
- Utilizes the atrous convolution architecture to classify sonar images where features are not well-defined or well-localized.
- Identifies objects not seen in the training set with a higher average precision than previous approaches.

We validate this framework on real sonar imagery containing a variety of objects in different environments.

The remainder of this paper is organized as follows. We first discuss background information and related work in the areas of Structure From Motion and sonar image analysis in Section II. We then discuss the reconstruction formulation in Section III and our method in Section IV. We next present our experiments and their results in Section V. We conclude and propose areas for future work in Section VI.

II. BACKGROUND

A. Sonar Imaging

As seen in Fig. 3, multi-beam sonar creates images by sending out a series of acoustic beams and measuring the energy returned from a reflecting object. Typically the sonar return of a point in Euclidean space (X, Y, Z) is mapped to polar space (r, θ, ϕ). The range r to the object is obtained by measuring the time to return for the reflecting energy, the bearing angle θ to the reflector is characterized by the beam number k, and the pixel value (0-255) is mapped from the amount of energy received back. The elevation angle ϕ is lost in the imaging process.

B. Analyzing Sonar Imagery

Previous work in the area of sonar image analysis have used object shadows or strong image gradients to automatically identify features [7], [8]. While they report impressive results, we note that these features are not robust in our application. In the data we collected, oftentimes an object shadow was not present in the image (e.g., Fig. 1a). Additionally, in imaging moored objects, an object's shadow would not be present (due to the lack of reflection back from the seafloor around the object). During our deployments we also found that strong image gradients can lead to the appearance of features where there are none (as seen in Fig. 1b).

Recently, CNNs have been used to analyze sonar imagery. Kim, et al. use CNNs to analyze sonar imagery and track the trajectory of another underwater vehicle [9]. While they are able to track this vehicle accurately, we compare to this method and show that the atrous architecture achieves a higher precision when classifying objects not in the training set. Williams and Dugelay address the problem of noise in sonar imagery by fusing together multiple views of the object of interest and using a deep network to classify images as either containing a man-made object or a naturally occurring rock [10]. We are able to complete our classification without the need for multiple views because our atrous network is able to account for noise in a single image.

There has also been work analyzing camera imagery captured onboard underwater vehicles. For example, Kaeli, et al. identify a set of images that could be used to summarize the mission [11]. To identify images in this set they use Quantized Accumulated Histogram of Oriented Gradient (QuHOG) features, which rely on the gradients in the image. As shown in Fig. 1b, in sonar imagery the image gradients in noisy images can be deceptive features.

To address the noise and low-resolution of sonar imagery, we use the dilated filters in the atrous convolution architecture. In previous work we showed that the atrous architecture

provides superior transfer learning capabilities when tested on images of objects not in the training set [12]. Details on this approach are discussed in Section IV-A.

C. Underwater Acoustic Structure From Motion

There has been much previous work in the area of using an imaging sonar for underwater 3D environment reconstruction. One popular approach in solving the underwater SLAM problem involves feature correspondence between pairs of sonar images. Hover et al. extract features from sonar images by clustering points in the image with large gradients [1]. A Normal Distribution Transform (NDT) is used for image registration, and the vehicle trajectory is optimized using a pose-graph. However, due to the unknown elevation angle of sonar image features, a planar assumption is used and the points are projected into the same plane as the vehicle. While this may work well for environments with large objects (e.g., sea floor, ship hull), this assumption is broken in more complex and unstructured environments.

The Acoustic Structure From Motion (ASFM) method proposed by Huang and Kaess provides two main advantages over other approaches [4]. First, ASFM relaxes the planar assumption for projecting sonar features into 3D. Second, their approach is capable of utilizing more than single pairs of sonar images to reconstruct 3D points. They formulate the reconstruction as a factor graph optimization problem. With the assumption of Gaussian noise, this can be formulated as a nonlinear least-squares problem. The factor graph is initialized by setting the sonar feature's unknown elevation angle to 0° and performing optimization via iterative linearization.

Huang and Kaess further extend their work to automatically perform feature association between sonar images [13]. Their proposed algorithm searches over a tree of potential feature correspondences, similar to Joint Compatibility Branch and Bound [14]. We leverage this work here to perform automatic feature correspondence and 3D reconstructions from hand annotated images. We contribute a complementary framework for analyzing incoming imagery to allow for the human to select the features in real-time.

III. RECONSTRUCTION FORMULATION

We formulate the ASFM problem using a factor graph as proposed by Huang and Kaess [4]. An overview of the process is provided here, with a graphical representation shown in Fig. 4. For full details, please refer to [4], [15].

In formulating the reconstruction objective, we seek to find the maximal probability set of poses, $\Theta = \{b_i, l_j\}$, and observations, $Z = \{u_i, m_k\}$, where b_i is a robot pose, l_j is a landmark, u_i is a navigation measurement, and m_k is a landmark measurement. Assuming Gaussian noise, this can be formulated as a nonlinear least-squares problem:

$$\Theta^* = \underset{\Theta, Z}{\operatorname{argmin}}[\, \|b_0\|^2_\Lambda + \sum_{k=1}^{M} \|h(b_i, l_j) - m_k\|^2_{\Xi_k} + \sum_{i=1}^{N} \|g(b_i, b_{i-1}) - u_i\|^2_{\Lambda_i}\,], \tag{1}$$

where $\|b\|^2_\Sigma = b^T \Sigma^{-1} b$ is the Mahalanobis distance squared and M and N are the number of sensor and odometry measurements respectively. The sensor model is defined as $h(b_i, l_j) + \mathcal{N}(0, \Xi_k)$ and the odometry model is defined as $g(b_i, b_{i-1}) + \mathcal{N}(0, \Lambda_i)$, where Ξ_k is the variance of sensor measurement k and Λ_i is the variance in odometry measurement i.

In this formulation, there are six unknowns $(x, y, z, \alpha, \beta, \gamma)$ for every pose and three unknowns (x, y, z) for every landmark where α, β, and γ are the roll, pitch, and yaw respectively. With all landmarks visible from every pose, fixing the first frame yields a fully constrained system iff: $6(n-1) + 3m \leq 2mn$, where n is the number of robot poses and m is the number of landmarks in the factor graph. Rearranging this, we derive n, the number of frames necessary for reconstruction, to be

$$n \geq \frac{3m - 6}{2m - 6}\,. \tag{2}$$

While this formulation is well-suited to the reconstruction process, it does not directly support performing these reconstructions in real time. In developing such a solution, we introduce an image proposal system which gives a human operator a small set of frames containing information useful for reconstruction.

IV. METHOD

A. Proposing Informative Frames

To select informative frames automatically we leverage the atrous convolution architecture. We motivate and briefly summarize our approach here, which first appeared in our prior workshop paper [12].

Due to the lack of strong local features in sonar images, the CNN method of analyzing sonar imagery can be greatly improved by using dilated filters (shown in Fig. 2b). Dilated filters use the same number of pixels as a standard filter; however because they are distributed, they can alleviate the effects of noise and low-resolution in sonar images.

The ability to ignore and not overfit to local features also improves the transfer learning capabilities in identifying informative sonar images. That is, given an image of an object not seen in training data, the atrous CNN is able to classify informative images with a higher average precision than a standard CNN or frequency-component based method. This ability is particularly attractive when using an underwater

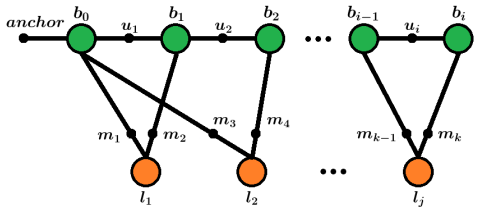


Fig. 4: Factor graph representation. l, m, b, and u are the landmark positions, landmark measurements, robot pose, and navigation measurements respectively. The landmarks are hand-labeled features in sonar images (green dots in Fig. 1a).

sonar, where there is a lack of training data as compared with terrestrial environments, and various sources of noise make similar looking objects appear different in sonar imagery.

B. Architecture Choice

To determine the appropriate architecture for configuring these dilated filters for use in analyzing sonar imagery, we test multiple architecture configurations and parameter choices. Full details of our selection process can be found in our prior work; a summary is below [12].

For the general architecture choice we evaluated three approaches. The first is similar to a normal CNN except the convolutional layers beyond the first have dilated filters instead of standard filters. This is inspired by Yu and Koltun, who use this configuration of dilated filters for dense prediction in image segmentation tasks [5]. The second pretrains two networks: one a standard CNN, the other an atrous network. The dense layers are then popped off of each and combined into a single dense layer for training again. This is meant to capture the expressibility of CNNs with the generalization capabilities of atrous networks. Finally, we test an architecture that uses filters of different dilation rates in the same layer. This is inspired by Chen, et al. who use this configuration for semantic image segmentation [6]. We complete a full parameter sweep on each architecture, including number of layers, number of filters, kernel size, dilation rate, and dense layer size. Average precision was used as the metric for comparing each architecture and parameter configuration. To account for variability in initialization, the results for each were averaged over 20 runs. We found the architecture in Fig. 6 to achieve the highest average precision. Due to the large dilation rate, this network starts to extract non-local features as early as the second convolutional layer. Therefore, it does not overfit to local patterns in the training set and has better generalization capabilities.

C. Frame Proposal Process

We treat the real-time frame proposal problem as a single-shot information-greedy selection. That is, every time a frame is proposed to the human operator, the most informative frame is selected from the set of candidate frames, $max(n_{info}) \in N$. The informativeness, n_{info}, of a frame is determined directly from the sigmoid output of our atrous convolution based deep network [12].

The set of candidate frames at any given time, N, is determined by two factors: n_{info} and the time since the last frame proposal, t_{motion}. Frame n_i is added to the set of candidate frames, $N \cup \{n_i\}$ if:

$$\begin{aligned} n_{info} &\geq T_{info} \\ t_{motion} &\geq T_{motion}. \end{aligned} \tag{3}$$

The threshold T_{info} allows for the tradeoff between frame information quality and frame quantity. The threshold T_{motion} is tuned to allow for view diversity while considering potential vehicle motion and sonar field-of-view.

Frames are proposed to the human operator sequentially as soon as the labeling of the previous frame is complete.

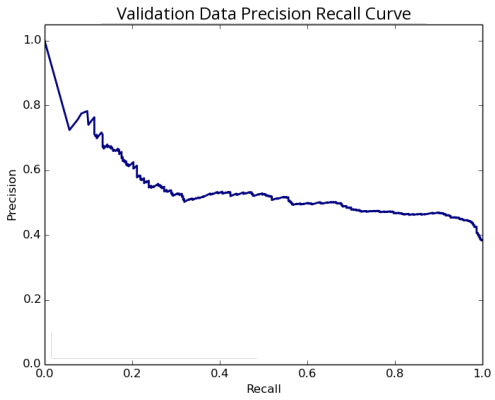


Fig. 5: The precision recall curve generated from validation data used to choose a threshold for informative vs. non-informative frames, denoted as T_{info}.

When a frame is proposed, the set of candidate frames, N, is cleared. Frame proposal stops when either an appropriate number of frames, N_{thresh}, have been labeled or the data stream ends (e.g., the deployment concludes).

Once a frame is proposed, the human operator selects the features to be reconstructed by simply clicking on the image. We found this usually involved a small number of features (3-6), which allowed for fast feature selection and reconstruction. An example of selected features for the "X" shaped object are shown as green dots in Fig. 1a. An interesting avenue for future work is the investigation of the effects of the experience or training of the human labeler. We note that given the simplicity of the task, we do not expect large variations across annotators.

D. Threshold Tuning

We set the threshold for proposal, T_{info} at 0.99. As demonstrated in Eq. 2, with objects having on the order of five features, often less than ten frames are needed to complete reconstructions. In these cases it is more beneficial to be selective in choosing frames (high threshold) rather than finding many informative frames. Experimentally this threshold provided more than enough frames to complete the reconstruction. For completeness, the precision recall curve from our model on validation data is shown in Fig. 5.

To determine N_{thresh}, we evaluated the reconstruction error for annotated set sizes of n (the minimum number of frames for reconstruction) to $2n + 1$. To obtain the average and variance on the sum squared reconstruction error (SSE) we use a set of data containing 7 proposed frames. From this set we test all of the combinations of frames for each annotated set size. These combinations are given by $\binom{2n+1}{k}$ where $2n + 1 = 7$ and $k = \{3, 4, 5, 6\}$. The same validation data used in completing the parameter sweep for the atrous network was used. As seen in Fig. 9, we find $N_{thresh} = 5$ to provide enough frames for a low reconstruction error/variance, with minor benefits increasing the set size to 6.

For the experiments we set $T_{motion} = 2$ sec. This threshold provided view diversity while allowing the objects

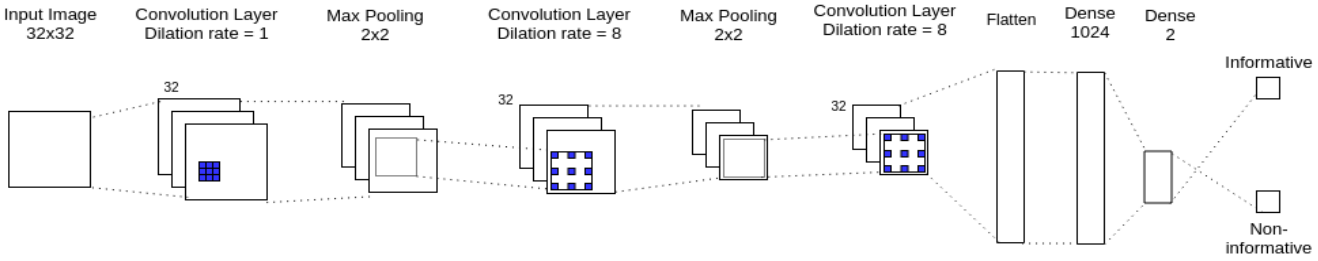


Fig. 6: The atrous convolutional network used. The parameters and filter configuration were tuned by testing discrete options.

of interest to remain in the sonar field-of-view. It also allowed the human operator to annotate each frame in real-time.

E. Reconstruction Process

Inspired by Huang and Kaess, initial landmark locations in the factor graph are calculated by setting the elevation angle ϕ to 0°and projecting the sonar feature points (r, θ, ϕ) into Euclidean coordinates (X, Y, Z) [4]. The nonlinear least-squares problem is then optimized via iterative linearization using the Levenberg-Marquardt (LM) algorithm.

V. EXPERIMENTS AND RESULTS

To demonstrate the capability of our framework to complete accurate reconstructions in real-time, we ran three experiments on real sonar data played back offline. The first shows the transfer learning capabilities of our atrous network when tested on imagery of objects not seen in training. In the second experiment, we show that we select an appropriate number of frames for our reconstruction. That is, we choose enough to fully constrain the optimization but do not propose many superfluous frames. Finally, the third experiment shows that our proposal method chooses a subset of frames that enable real-time reconstructions with errors lower than baseline methods.

Throughout each test we maintain a real-time property by ensuring that at the end of a given experiment, our algorithm has output the reconstructed shape. Experiments 1 and 2 are used only for analyzing components of our process and are thus not timed. Experiment 3 mimics a deployment and thus in Section V-C.3 we provide runtime from start to finish of our process. For completeness we provide average numbers for each component of our system. Our CNN processes images in 18.1 ± 0.8 ms (55.2 ± 2.41 Hz). We found that a human operator annotated images of the X object in 2.63 ± 0.231 sec. Finally, the factor graph bundle adjustment process took on average 77.1 ± 9.4 ms.

A. System Overview

For our experiments we use a Tritech Gemini 720i multi-beam sonar onboard a tethered Seabotix vLBV300 Remotely Operated Vehicle (shown in Fig. 7).

The Gemini 720i is a standard multi-beam imaging sonar with similar parameters to the SoundMetrics DIDSON, BlueView M900-90, and the Aris Explorer 3000, all of which have been used in previous work [4], [16], [7]. It operates at 720kHz, images a 120° swath horizontally, has a maximum

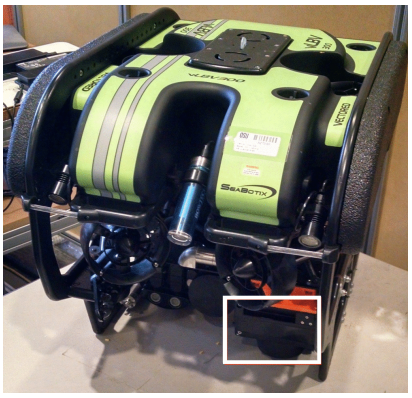


Fig. 7: The Seabotix vLBV300 and Gemini 720i imaging sonar (designated by the white rectangle in the lower right).

range of 120m, and produces images at 20Hz. This data rate, combined with large amounts of image noise, creates a situation in which the human operator is presented with a large number of frames, many of which will not be useful.

Our vehicle is also equipped with a Doppler Velocity Log and Inertial Navigation System which we use for pose estimates in the factor graph mapping scheme. The vehicle's tether provides continual power as well as data transmission back to an offboard computer which can be used to analyze sonar imagery in real-time. The use of standard underwater sensors demonstrates our approach can be deployed on a variety of platforms for real-time underwater reconstructions.

B. Dataset Overview

In this work we use four datasets captured onboard our vehicle while it operated in a variety of environments. Each dataset was collected in a passive manner, meaning each dataset is a single contiguous video stream captured while our vehicle moved through underwater environments. The content and number of informative frames varied across datasets and is summarized in Table I.

In dataset X_1 we collect a set of 5000 frames of the X shaped object (Fig. 8a) in the Oregon State University pool. In dataset X_2 we collect 1107 frames of the same X object in the same environment on a different date. In dataset $Multi$ we collect a set of 5000 frames containing insonified imagery of the X, Square, Triangle, and T-shaped objects (Figs. 8a-8d) in the pool. This dataset contains images with just a single object in addition to images containing multiple objects. Finally, in dataset $Cinder$ we capture 1470 frames of a cinder block target (Fig. 8e) in Yaquina Bay, Newport,

TABLE I: Summary of the datasets

Dataset	Total number of frames	Percent informative frames
X_1	5000	36%
X_2	1107	56%
$Multi$	5000	39%
$Cinder$	1470	42%

Note: An informative frame is one in which a human operator can clearly identify an object and its features in the sonar image.

TABLE II: Average precision for baselines and our method

Method	Avg. Precision		
	Seeding X present	Seeding No X	No Seeding No X
DCT [17]	0.57±0.01	0.55±0.02	0.48±0.02
Two-layer CNN [9]	0.69±0.04	0.58±0.05	0.42±0.04
Three-layer CNN	0.70±0.05	0.64±0.04	0.42±0.03
Atrous Convolution (ours)	**0.72±0.02**	**0.68±0.03**	**0.54±0.04**

OR. While in this work we use representative shapes for the underwater domain (resembling objects such as underwater moorings) an interesting avenue for future work involves the use of our architecture on more complex shapes.

The binary ground truth label (informative or non-informative) is obtained by having an expert evaluate if each image contains enough well-defined features to clearly identify the object. For example, simply seeing two lines intersect does not identify one of the five targets; however observing three well defined lines that each meet in an acute fashion clearly defines the triangle object seen in Fig. 8d.

Throughout the three experiments we train primarily using datasets X_1 and X_2 (both only containing data of the X target in Fig. 8a). 1000 frames of the $Multi$ dataset are used as validation data for the tuning of parameters.

C. Experiments

We conduct three experiments to demonstrate that our framework selects informative subsets of images thus enabling real-time reconstructions. In the first experiment we test using 4000 frames of the *Multi* dataset (the remaining 1000 frames were used as validation data). In Experiment 2 we use these same frames as well as the *Cinder* dataset. In Experiment 3, we only use data from the *Cinder* dataset. Unless stated otherwise, test data for each experiment contains imagery of objects not trained on, thus demonstrating the flexibility or lack thereof of each approach examined. This also serves to demonstrate the performance of our approach when trained prior to deployments.

1) Validating the transfer learning capabilities of our atrous network: In the first experiment we show that our model is able to identify objects not in the training set with a higher average precision than baseline methods. We are able to do this while also classifying objects seen in the training set with precision comparable to baseline methods. This ability is a result of our atrous network, which balances the power of standard convolutional architectures with the more global context afforded by dilated filters.

We compare against three baselines. The first is the Discrete Cosine Transform (DCT) method of global analysis. Defined for pixels m, n of an image A of size $M \times N$ as:

$$B_{pq} = \alpha_p \alpha_q \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} A_{mn} cos \frac{\pi(2m+1)p}{2M} cos \frac{\pi(2n+1)q}{2N}, \tag{4}$$

where

$$\alpha_p = \begin{cases} \frac{1}{\sqrt{M}}, p = 0 \\ \sqrt{\frac{2}{M}}, 1 \leq p \leq M-1 \end{cases} \tag{5}$$

$$\alpha_q = \begin{cases} \frac{1}{\sqrt{N}}, q = 0 \\ \sqrt{\frac{2}{N}}, 1 \leq q \leq N-1 \,, \end{cases} \tag{6}$$

and B_{pq} is the coefficient on the basis cosine function defined by the constant p and q [17]. The resulting coefficients represent the frequency components contained in the image, which can be used to separate noisy images from those with more complex geometries. We complete this classification with the use of a Support Vector Machine using the frequency components as the input data. This method is natural for sonar imagery because it takes a global image analysis approach to imagery with poorly-defined features. The second baseline is the approach taken by Kim, et al. (called "two-layer CNN") [9]. The third and final baseline is a network we developed inspired by Kim, et. al that contains an additional convolutional layer (called "three-layer CNN").

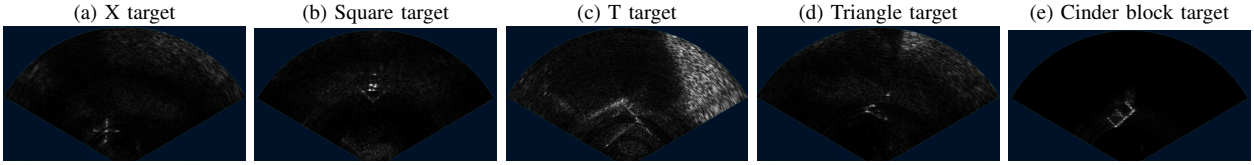
We compare the four approaches in three scenarios, the results of which are shown in Table II. The first scenario allows each classifier to train with 600 frames of multi-object data (called "seeding"). The test data contains imagery of each of the four objects in Figs. 8a-8d (including the X). These results show that the DCT method of global analysis is not expressive enough to capture the features present in both the training and test data. The two and three-layer CNNs, as well as our method, all perform similarly.

The second scenario still seeds the training of the models, but the test data no longer contains imagery of the X object. We find that our model outperforms the others, which is expected because dilated filters capture a larger neighborhood of influence than localized filters.

Finally, in the third scenario, true transfer learning is evaluated. The models are not given imagery of any object but the X to train on and the test set only contains imagery of the Triangle, Square, and T shaped objects. In this scenario we find that our model significantly outperforms other methods, which is intuitive as dilated filters give a nice balance between the power of deep learning based methods and the larger context of global evaluation methods. As demonstrated in the first scenario, this does not come at the cost of performing worse than baselines in the case where the test data is similar to the training data.

2) Choosing the appropriate number of frames to propose: In the second experiment, we show that our method proposes an appropriate number of frames. That is, our method achieves lower reconstruction error than using the minimum number of frames to fully constrain the reconstruction (n) as well as achieving comparable performance to reconstructions that use a large number of frames ($2n$). Given the lack of

(a) X target (b) Square target (c) T target (d) Triangle target (e) Cinder block target



(f) X in sonar (g) Square in sonar (h) T in sonar (i) Triangle in sonar (j) Cinder block in sonar

Fig. 8: Each target and its representation in sonar image space. (a)-(d) are images taken from the vehicle underwater. As a note, in Fig. 8j the second long side of the cinder block has been occluded.

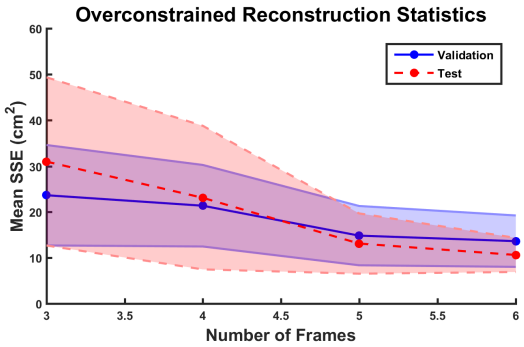


Fig. 9: Selection of the appropriate number of frames to use for reconstructions. For both validation data (blue line) as well as test data (red line) five frames is at a "knee" point as increasing to six frames gives only a marginal benefit.

global ground truth coordinates, known object dimensions are used to compute the reconstruction error.

The results of this experiment can be seen in Fig. 9. As expected, increasing the number of frames used in the reconstruction decreases the reconstruction error. When only using the minimum number of frames required, n, we can see that on average the reconstruction has both the highest sum squared error (SSE) as well as the highest variance. At five frames in both plots we observe a "knee" point, where increasing the size to six gives only marginal benefit.

3) Demonstrating the ability of our network to select informative subsets: In the third experiment, we demonstrate the ability for our method to select informative frames and operate more efficiently than standard baselines. For this experiment we use a subset of 400 contiguous frames. The results of this experiment can be seen in Table III. Statistical bounds have been provided for the comparison to [13] due to the random nature of frame proposals in this baseline. These bounds were found averaging over 5 runs.

The first baseline we compare against is naively presenting the user with every frame for annotation. We note that this method requires the human operator to annotate every frame and thus cannot be completed in real-time. The large amount of error is due to the fact that many poor quality frames are annotated by the user. Such large error demonstrates the low-quality of many of the images, which do not contain enough clear features to be annotated by the human.

The second baseline we test is proposing every informative frame to the user for annotation. The informativeness of a frame is determined by the sigmoid output of our atrous CNN. This baseline achieves a reconstruction error similar to that of our method; however we note that it cannot be run in real-time. The low reconstruction error is expected because the labeled images are all informative frames.

In the third baseline, we naively subsample all frames at an even interval to only present the human operator with the same number of frames as our method. This is equivalent to naively reducing the data rate to allow for real-time labeling. While this allows for real-time performance, we can see that the reconstruction error achieved is poor. This is due to the fact that without first assessing a frame's informativeness, poor quality frames are still presented to the user.

The fourth baseline allows us to compare against a method inspired by the state-of-the-art in underwater reconstruction [13]. In this previous work, Huang and Kaess perform automatic data association between annotated frames to determine whether or not all annotated feature points can be associated. If they cannot, the frame is rejected and not used for reconstruction purposes. To extend this approach and enable real-time reconstructions, we randomly select frames to propose to the human operator. If an association exists between the annotated feature points, the frame is used in the reconstruction. If not, it is discarded and another is proposed. The proposal process runs until the minimum number of frames required, n, have been successfully annotated. While this method is able to achieve high quality reconstructions, the reconstruction error is both higher and more variable than our method. This is due to the fact that only the minimum number of frames required are used in the reconstruction process. We also note that since frames are not evaluated before proposal, this baseline typically proposes a larger number of frames than our method.

The last row of the table presents the results of our method. We note that we achieve the lowest average reconstruction

TABLE III: Evaluation of our reconstruction framework

Method	Number of frames proposed	Reconstruction SSE (cm^2)
Every frame	400	8761.9
Every informative frame	167	16.37
Naive subsampling	**5**	71.239
Huang and Kaess [13]	6.93±1.91	18.42±6.47
Atrous proposal system (ours)	**5**	**13.12±3.72**

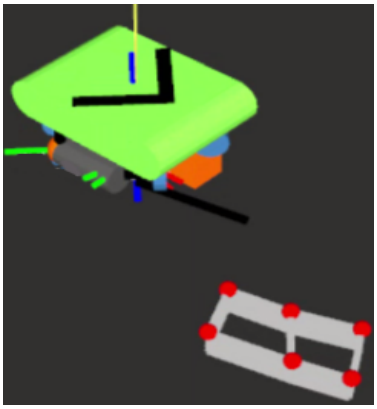


Fig. 10: The vehicle and reconstruction of the cinder block in the RViz simulation environment. Our method outputs the red dots (chosen as features by the operator) and the gray section was extrapolated.

error and variability, while simultaneously proposing the lowest number of sonar frames to the user. This low reconstruction error is achieved while using N_{thresh} = 5 frames. We also maintain real-time performance during the 20 second experiment. Our method, from start to finish, took on average 14.88 ± 0.49 seconds to complete.

VI. CONCLUSION

In this work we present a novel framework that enables the capability to complete accurate underwater 3D reconstructions in real-time and overcoming the large amounts of noise present in sonar imagery. We enable this functionality by identifying a small subset of informative frames for the human to annotate. Through experiments on real sonar images we show our atrous network outperformed other classifiers in identifying informative frames for proposal. We also experimentally validate the ability of our network to leverage non-local features to complete transfer learning.

One particularly interesting extension to this work is the formal treatment of incorporating image diversity into our framework. While in this work we enforce diversity with T_{motion}, the effects of using diverse images into the reconstruction process remains an interesting avenue to explore.

ACKNOWLEDGMENT

We would like to thank Dylan Jones, Nicholas Lawrance, Seth McCammon, Lauren Milliken, and Thane Somers who helped in the deployment of our vehicle. Nicholas Lawrence also assisted with the Seabotix vehicle RViz simulation.

REFERENCES

[1] F. S. Hover, R. M. Eustice, A. Kim, B. Englot, H. Johannsson, M. Kaess, and J. J. Leonard, "Advanced perception, navigation and planning for autonomous in-water ship hull inspection," *The International Journal of Robotics Research*, vol. 31, no. 12, pp. 1445–1464, 2012.

[2] M. Johnson-Roberson, O. Pizarro, S. B. Williams, and I. Mahon, "Generation and visualization of large-scale three-dimensional reconstructions from underwater robotic surveys," *Journal of Field Robotics*, vol. 27, no. 1, pp. 21–51, 2010.

[3] M. D. Aykin and S. S. Negahdaripour, "Modeling 2-d lens-based forward-scan sonar imagery for targets with diffuse reflectance," *IEEE Journal of Oceanic Engineering*, vol. 41, no. 3, pp. 569–582, 2016.

[4] T. A. Huang and M. Kaess, "Towards acoustic structure from motion for imaging sonar," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2015, pp. 758–765.

[5] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. of the International Conference on Learning Representations (ICLR), San Juan, Puerto Rico*, 2016.

[6] 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*, 2017.

[7] Y. Ji, S. Kwak, A. Yamashita, and H. Asama, "Acoustic camera-based 3D measurement of underwater objects through automated extraction and association of feature points," in *Proc. IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, 2016, pp. 224–230.

[8] H. Johannsson, M. Kaess, B. Englot, F. Hover, and J. Leonard, "Imaging sonar-aided navigation for autonomous underwater harbor surveillance," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010, pp. 4396–4403.

[9] J. Kim, H. Cho, J. Pyo, B. Kim, and S.-C. Yu, "The convolution neural network based agent vehicle detection using forward-looking sonar image," in *Proc. IEEE/MTS OCEANS, Monterey, CA*, 2016.

[10] D. P. Williams and S. Dugelay, "Multi-view sas image classification using deep learning," in *Proc. IEEE/MTS OCEANS, Monterey, CA*, 2016.

[11] J. W. Kaeli, J. J. Leonard, and H. Singh, "Visual summaries for low-bandwidth semantic mapping with autonomous underwater vehicles," in *Proc. IEEE/OES Conference on Autonomous Underwater Vehicles, Oxford, MS*, 2014.

[12] R. DeBortoli, A. Nicolai, F. Li, and G. A. Hollinger, "Assessing perception quality in sonar images using global context," in *Proc. IEEE Conference on Intelligent Robots and Systems Workshop on Introspective Methods for Reliable Autonomy, Vancouver, CA*, 2017.

[13] T. A. Huang and M. Kaess, "Incremental data association for acoustic structure from motion," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2016, pp. 1334–1341.

[14] J. Neira and J. D. Tardós, "Data association in stochastic mapping using the joint compatibility test," *IEEE Transactions on Robotics and Automation*, vol. 17, no. 6, pp. 890–897, 2001.

[15] M. Kaess, A. Ranganathan, and F. Dellaert, "isam: Incremental smoothing and mapping," *IEEE Transactions on Robotics*, vol. 24, no. 6, pp. 1365–1378, 2008.

[16] E.-h. Lee and S. Lee, "Development of underwater terrain's depth map representation method based on occupancy grids with 3D point cloud from polar sonar sensor system," in *Proc. IEEE International Conference on Ubiquitous Robots and Ambient Intelligence*, 2016, pp. 497–500.

[17] J. Makhoul, "A fast cosine transform in one and two dimensions," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 1, pp. 27–34, 1980.