2014 IEEE International Conference on Robotics & Automation (ICRA)
Hong Kong Convention and Exhibition Center
May 31 - June 7, 2014. Hong Kong, China

Multimodal Learning for Autonomous Underwater Vehicles from Visual and Bathymetric Data

Dushyant Rao, Mark De Deuge, Navid Nourani-Vatani, Bertrand Douillard,
Stefan B. Williams and Oscar Pizarro

***Abstract*— Autonomous Underwater Vehicles (AUVs) gather large volumes of visual imagery, which can help monitor marine ecosystems and plan future surveys. One key task in marine ecology is benthic habitat mapping, the classification of large regions of the ocean floor into broad habitat categories. Since visual data only covers a small fraction of the ocean floor, traditional habitat mapping is performed using shipborne acoustic multi-beam data, with visual data as ground truth. However, given the high resolution and rich textural cues in visual data, an ideal approach should explicitly utilise visual features in the classification process. To this end, we propose a multimodal model which utilises visual data and shipborne multi-beam bathymetry to perform both classification and sampling tasks. Our algorithm learns the relationship between both modalities, but is also effective when visual data is missing. Our results suggest that by performing multimodal learning, classification performance is improved in scenarios where visual data is unavailable, such as the habitat mapping scenario. We also demonstrate empirically that the model is able to perform generative tasks, producing plausible samples from the underlying data-generating distribution.**

I. INTRODUCTION

Autonomous Underwater Vehicles (AUVs) are often deployed to obtain imagery of the sea floor through optical sensing [1][2]. These vehicles gather large volumes of image data to facilitate the monitoring of marine ecosystems. An important task in marine ecology is *benthic habitat mapping*, which involves classifying large regions of the sea floor into broad habitat categories (such as sand, reef, etc) [2].

These benthic habitats are primarily determined by the substrate (such as rock or sediment) and the organisms present (such as algae or coral) [3], making them relatively easy to distinguish using AUV image data [4]. However, AUVs can only traverse a small fraction of a larger area of interest, limiting the scale to which visual habitat classification can be performed. Conversely, bathymetric (ocean depth) data from shipborne multi-beam sonar is usually available a priori over an entire site, but has a low spatial resolution (1.6 m between adjacent points for the dataset used in this work).

*This work was supported by the Australian Research Council (ARC) through Discovery programme grant number DP110101986, the Australian Government through the SIEF programme, the Australian Centre for Field Robotics at the University of Sydney and the Integrated Marine Observing System (IMOS).

Bertrand Douillard is with the NASA Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA. Email: Bertrand.Douillard@jpl.nasa.gov. The other authors are with the Australian Centre for Field Robotics, The University of Sydney, NSW, Australia. Emails: f.lastname@acfr.usyd.edu.au.

Existing solutions to benthic habitat mapping tend to consider labelled visual data as ground truth for classification of acoustic data [5]. The use of visual data is purely discriminative, and so we propose learning a generative probabilistic relationship between the visual and bathymetric features. Such a model could aid planning of future surveys, by answering multimodal and unimodal queries such as *"Where are we likely to find algae?"* or *"What kinds of bathymetric features will we see in conjunction with this image?"*.

To this end, we propose a multimodal learning approach based on *denoising autoencoders* [6] that is able to find correlations between bathymetry and visual features. Our method can naturally deal with missing modalities, allowing it to use visual features in the learning process even when applied to large benthic regions where only bathymetry is available. Further, the method can handle both discriminative and generative tasks, allowing it to perform benthic classification as well as sample from the underlying data-generating distribution. In particular, performance is improved on a bathymetry-based classifier by performing multimodal learning before classification.

The rest of this paper is organised as follows: Section II assesses related work in benthic classification and learning; Sections III and IV provide some background and outline our proposed approach; Section V describes the datasets used in our experiments; and Sections VI and VII detail our results and conclusions from this research.

II. RELATED WORK

A. Benthic Classification

A number of existing methods address the problem of assigning benthic habitat labels to visual imagery, by performing supervised classification of coral reef survey images [7] [8], or clustering benthic imagery in an unsupervised fashion [9]. However, these approaches are constrained to previously imaged areas, precluding large-scale habitat mapping. As a result, many approaches to benthic habitat mapping are bathymetry-based, using in-situ visual imagery as "ground truth" [5]. One such technique clusters AUV-based benthic imagery, and uses the probabilistic output as training labels for classification of bathymetric features [10]. Another method extrapolates vision-based results to larger regions, using visual classification from a completed dive to determine the most informative future dive from a set of candidates [11]. Building on these techniques, our approach looks to incorporate both bathymetric and visual features into

the classification process, whilst maintaining the ability to classify either modality on its own.

B. Feature Learning and Multimodal Learning

Feature learning refers to a family of learning techniques that attempt to determine a set of basis vectors or features to describe a dataset, often with a sparse representation. Different algorithms can perform feature learning in practice, including sparse autoencoders (SAEs), k-means clustering, Gaussian mixture models and restricted Boltzmann machines (RBMs) [12]. These methods all tend to learn similar dictionaries of localised filters [12], such as Gabor-like edge filters for natural images, or handwriting "strokes" for the MNIST digits dataset. While RBMs are generative models that can sample from the data-generating distribution [13], SAEs are trained to optimise their reconstruction of the input data.

Deep learning approaches stack multiple feature learning layers into a single model, and can learn entire hierarchies of filters. Each layer of a deep network usually learns a set of features at a different scale or complexity, such as an edge, object-part, or whole object [14].

Deep networks have previously been applied to multimodal learning tasks, since multiple layers can capture higher order correlations between two data modalities. Ngiam et al. [15] use a deep learning approach to perform classification of phonemes from audio and video features. However, their network is tuned as a sparse autoencoder, which has been shown to perform poorly in generative tasks [6]. Another technique learns correlations between a large set of images and a set of keywords [16] using a deep RBM. By maintaining the generative properties of the RBM, they are able to sample keywords given an input image and vice versa. Another approach uses a Bayesian co-clustering algorithm to learn a relationship between a visual dictionary and textual words, in order to perform classification and keyword-based image retrieval tasks [17].

C. Our contributions

Building on these methods, we advocate the use of *denoising autoencoders* (DAEs) for multimodal learning tasks. DAEs are trained to reconstruct the clean input after a corruption is applied, motivated by the fact that learned features should be invariant to minor changes in the input data [6]. By applying masking noise stochastically to a fraction of the inputs, the training process naturally mimics the scenario where some modalities are missing, making them well equipped to handle multimodal learning tasks. Further, our application demands strong performance in *both* discriminative and generative tasks (classification and sampling, respectively). DAEs, unlike SAEs, can be used as generative models and can sample from the data-generating distribution [6].

We validate our method with both classification and sampling results, demonstrating its efficacy as a multimodal inference engine for visual and bathymetric data. In particular, we demonstrate that classification of bathymetric features is improved by learning multimodal correlations with visual data, and also demonstrate the ability of the model to sample from the underlying data distribution.

III. BACKGROUND

A. Autoencoder

An autoencoder is a single layer neural network in which the hidden layer learns to reconstruct the input. The input $\mathbf{x} \in [0,1]^d$ is *encoded* to a hidden layer representation $\mathbf{y} \in [0,1]^m$ according to $\mathbf{y} = \text{sigm}\left(\mathbf{W}\mathbf{x} + \mathbf{b}\right)$, and the output reconstruction $\mathbf{z} \in [0,1]^d$ is *decoded* with $\mathbf{z} = \text{sigm}\left(\mathbf{W}'\mathbf{y} + \mathbf{b}'\right)$. Here, d and m are the dimensions of the input and hidden representation, $\text{sigm}(x) = \frac{1}{1+e^{-x}}$ is the element-wise logistic sigmoid function, $\mathbf{W}$ and $\mathbf{W}'$ are weight matrices, and $\mathbf{b}$ and $\mathbf{b}'$ are bias vectors. In the case of real-valued data $\mathbf{x} \in \mathbb{R}^d$, a linear decoder $\mathbf{z} = \mathbf{W}'\mathbf{y} + \mathbf{b}'$ is usually used for the reconstruction. The model parameters are often further constrained by using *tied weights*, $\mathbf{W}' = \mathbf{W}^\top$ [12]. We adopt this constraint for our approach, since it acts as a regulariser and affords additional flexibility in the model (e.g. the option to fine tune the model as a RBM).

Given a training set of n input data vectors, each training vector $\mathbf{x}^{(i)}$ can be mapped to a hidden representation $\mathbf{y}^{(i)}$, followed by reconstruction $\mathbf{z}^{(i)}$. The model parameters $\theta = \{\mathbf{W}, \mathbf{b}, \mathbf{b}'\}$ are then tuned to minimise a loss function, often the mean squared reconstruction error over the training set. To prevent the weights from increasing unboundedly, we add an L_2 weight decay term, using the square of the Frobenius norm of weight matrix $\mathbf{W}$. Further, past work has empirically shown that hidden units that are selectively activated are more useful in discriminative tasks [12]. Thus, we also incorporate a sparsity cost, based on the cross entropy between the sparsity (average activation) of each unit, $\hat{\rho}_j = \frac{1}{n}\sum_{i=1}^{n} y_j^{(i)}$, and a user-defined sparsity ρ.

The entire objective function including weight decay and sparsity cost, is given by:

$$\begin{aligned} J\left(\theta\right) &= \frac{1}{n}\sum_{i=1}^{n}\left\|\mathbf{x}^{(i)} - \mathbf{z}^{(i)}\right\|_2{}^2 + \lambda\|\mathbf{W}\|_F{}^2 \\ &\quad + \beta\sum_{j=1}^{m}\left[\rho\log\frac{\rho}{\hat{\rho}_j} + (1-\rho)\log\frac{(1-\rho)}{(1-\hat{\rho}_j)}\right] \\ \theta^* &= \underset{\theta}{\text{argmin}}\ J\left(\theta\right) \end{aligned} \tag{1}$$

Here, λ and β are hyperparameters to tune the effects of weight decay and sparsity cost, respectively.

Typically, the parameters are learned using stochastic gradient descent or another gradient-based optimisation procedure. As a result, the autoencoder learns a hidden layer representation to minimise the mean squared error between the input and the model-based reconstruction.

Multiple layers of autoencoders can be stacked on top of one another in order to learn higher order relationships between features, forming a deep network. This type of network is learned using greedy layerwise training: training each layer, encoding the input features according to the learned parameters, and using these as input for the next layer.

B. Denoising Autoencoder

In a denoising autoencoder (DAE), a stochastic corruption is applied to each data vector $\mathbf{x}^{(i)}$ during training, and the resulting vector $\tilde{\mathbf{x}}^{(i)}$ is used as the training input. By comparing the model reconstruction with the *clean* input, the DAE learns to reconstruct input data with robustness to corruption / noise.

Typical options for the stochastic corruption include additive isotropic Gaussian noise, salt-and-pepper noise and masking noise. Since our application demands robustness to missing modalities / inputs, we apply masking noise to the input vectors, where a fraction η of the input dimensions are set to zero. The corruption process is stochastic, so different dimensions are masked for each training vector and for each iteration of learning. However, after training the model, the hidden representation is obtained using *clean* inputs, so that future tasks with the encoded features are not probabilistic.

IV. MULTIMODAL LEARNING

A. Notation

Our method utilises square patches of gridded bathymetry and AUV-based visual images. A bathymetry patch $\mathcal{B}$ can be considered as the sum of a mean ocean depth $\mathcal{B}_0 = \text{mean}(\mathcal{B})$, and a zero-mean patch capturing the local bathymetric variation (or "shape"), $\mathcal{B}_l = \mathcal{B} - \mathcal{B}_0$. The local variation is important in determining the habitat; for example, sandy regions are likely to exhibit smoother bathymetry gradients than reef habitats. Similarly, the depth is also significant (for example, kelp species prefer shallower water). However, since the depth has a much larger magnitude than the local variation, it is likely to dominate the feature representation if $\mathcal{B}$ is used directly. We solve this problem by separating the bathymetry data into these two variables. For the remainder of this paper, we refer to the mean ocean depth as $\mathcal{B}_0$, the zero-meaned local bathymetry patch as $\mathcal{B}_l$, and the visual input as $\mathcal{V}$.

B. Multimodal Architecture

Our proposed architecture learns multimodal correlations using a multi-layer hierarchy. As shown in Fig. 1, the features for $\mathcal{B}_0$, $\mathcal{B}_l$, and $\mathcal{V}$ are concatenated in the *mid-layer*. We then learn a DAE *multimodal layer* using the mid-layer as input, which learns correlations between the modalities and learns to reconstruct the whole mid-layer input even when large amounts ($\sim 50\%$ of the input dimensions) are missing.

The mid-layer feature representation of $\mathcal{B}_l$ is obtained by learning a DAE on the local bathymetry patches with 25% random masking noise and a prescribed sparsity of 0.05. Conversely, the raw depth $\mathcal{B}_0$ is fed directly into the mid-layer, and image-level features are used for the visual data. We can justify this architecture by considering the kinds of correlations that are likely to occur between these modalities. The ocean depth is unlikely to correlate with $\mathcal{B}_l$ patch pixels directly, but may be related to the first layer $\mathcal{B}_l$ features (local edge and gradient filters). For example, in the datasets used for this work, deeper areas often have smoother bathymetry gradients corresponding to sand habitats, while shallower reef regions exhibit localised 'blob-like' bathymetry. Similarly, the local bathymetry is more likely to correlate with image-level features than individual image pixels or sub-image patches.

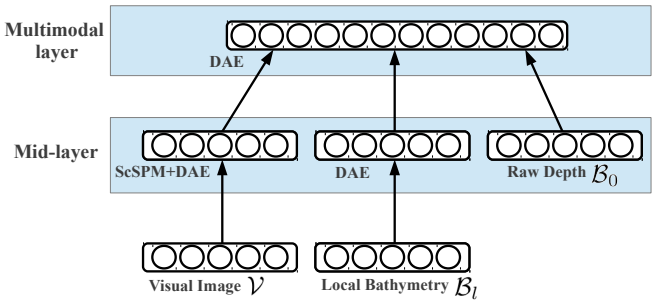


Fig. 1. The proposed model for multimodal learning. The ScSPM visual features, DAE bathymetric features, and encoded depth features are concatenated at the mid-layer, and the multimodal DAE is learned on top.

Since the mid-layer $\mathcal{B}_l$ features are the output of a DAE encoding (a sparse code in the interval $[0, 1]$), we use a modified 1-of-k encoding for $\mathcal{B}_0$. This is necessary because it is difficult for the multimodal layer to learn correlations between two modalities with different input ranges. In particular, we divide the observed depth range of $19 - 100\,\text{m}$ into 82 equally spaced bins with an increment of 1m, and encode each $\mathcal{B}_0$ as a value of 1 in the corresponding bin. To ensure that adjacent bins are correlated (i.e. to explicitly encode the continuous nature of the depth data) the value of nearby bins are set based on a Gaussian-like falloff. The width of this falloff is selected so that the sparsity of the encoded $\mathcal{B}_0$ features is approximately the same as that of the mid-layer $\mathcal{B}_l$ features. Outside this width, all other values are zero.

The mid-layer features for $\mathcal{V}$ are obtained using the method outlined in [4], based on the Sparse coding Spatial Pyramid Matching (ScSPM) algorithm [18]. A grid of overlapping square patches are extracted from an image and each encoded with a SIFT descriptor. These are converted to a sparse code (using a prelearned dictionary), pooled over the entire image using a spatial pyramid, and then compressed using Random Projections [4]. The output of the Random Projections step results in a 3000-dimensional dense feature vector, and we need to ensure the mid-layer features have a similar sparsity across all modalities. Thus, we train another DAE on these features, and use the encoded representation as the mid-layer visual features. This ensures that the visual features do not dominate the representation in the multimodal layer.

C. Implementation Details

We perform contrast normalisation on the zero-mean bathymetric patches $\mathcal{B}_l$ when training the first layer, to ensure that our learned bases cover a larger region of the input space [12]. We also perform ZCA whitening [12] on the bathymetric patches to remove redundancy and pixel-wise correlations, reducing the patches to 104 dimensions to preserve 95% of the variance.

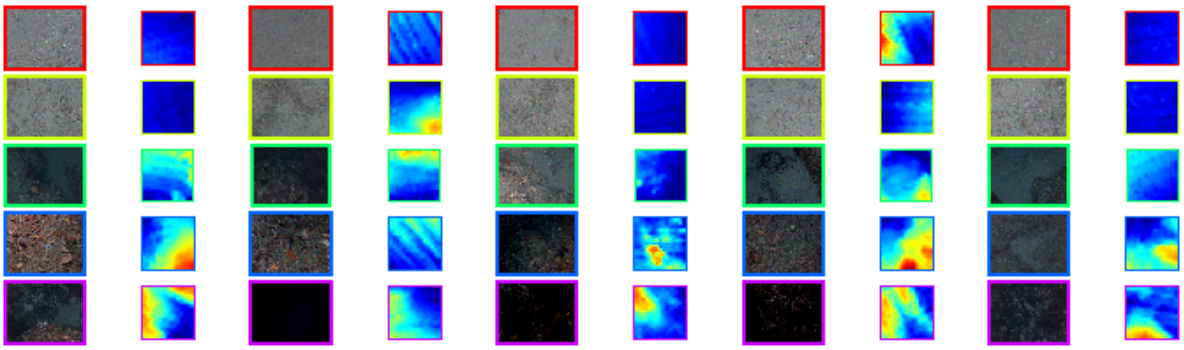


Fig. 2. Examples of the habitat classes and raw data used. Each coloured row represents one of the habitat classes, which are (from top to bottom) sand, screw shells / rubble, reef / sand interface, reef and Ecklonia Radiata (kelp). A number of data samples are shown in each row, comprising an image obtained from the AUV (left) with the corresponding patch from gridded bathymetry data (right). The images are 1360×1024 pixels and are taken $1.5 - 3.5\,\mathrm{m}$ above the ocean floor, while the 15×15 bathymetric patches cover a $22.4 \times 22.4\,\mathrm{m}^2$ area. In the bathymetric patches, red regions represent shallow waters relative to blue regions.

The DAEs for $\mathcal{B}_l$ and $\mathcal{V}$ contain 1000 hidden units, while the multimodal DAE contains 2000. Experiments were performed with 100, 200, 500, 1000, 2000 and 4000 hidden units, but the selected numbers exhibited the best performance in unsupervised learning and discriminative training.

D. Classification and Habitat Mapping

To perform benthic classification with $\mathcal{B}_0$, $\mathcal{B}_l$, and $\mathcal{V}$ all available, we can encode the mid-layer representation using the multimodal layer, and pass these "multimodal encoded" features into a logistic regression classifier. When modalities are missing (such as for benthic habitat mapping), we simply set the missing input dimensions to zero in the mid-layer before performing the multimodal encoding. Given that the DAE learns features robust to masking noise, we expect that this procedure will yield better results than if we were to perform classification on the mid-layer features directly. This is confirmed in our results in Section VI.

E. Sampling

Unlike the standard sparse autoencoder, the denoising autoencoder can produce plausible samples from the data-generating distribution [6]. However, contrary to RBMs, DAEs lack a model of the marginal distribution of the hidden layer [19] and cannot generate samples from an arbitrary hidden layer representation. Vincent [19] proposes that this marginal distribution be modelled as an empirical distribution, comprised of the set of hidden codes obtained by encoding the training vectors. Using the example of a single layer DAE, a sample can be obtained as follows. First, we encode a training sample to obtain the hidden layer representation. Then, we perform Bernoulli sampling, where the value of each unit acts as the probability of activation of the unit (as with an RBM), resulting in a binary code. Finally, deterministic decoding yields a new input sample.

For our network, we adopt the same procedure for the DAE layers, comprised of a deterministic bottom-up pass, followed by alternating Bernoulli sampling and deterministic decoding. Missing modalities are again handled by masking the corresponding dimensions in the mid-layer. Note that since the ScSPM approach cannot perform top-down decoding (due to the pooling layers), we cannot sample visual features below the image level.

One specific task of interest is to sample depth / bathymetric features conditioned on an input image; i.e. the query *"Given this image, what kinds of bathymetric features might be present?"*. To handle this task, we use the same sampling procedure but instead clamp the visual features at the observed values (i.e. the evidence) on each iteration.

V. DATASETS

The bathymetric patches were obtained from large-scale gridded data from Geoscience Australia [20]. The visual ScSPM features were computed from 1360×1024 pixel images taken during 14 dives by our AUV *Sirius* in 2008 off the Eastern coast of Tasmania, Australia [2].

The DAE layer for $\mathcal{B}_l$ was trained using a set of $500,000$ 15×15 patches extracted randomly from the gridded data. With a separation of $1.6\,\mathrm{m}$ between grid points, they each represent an area of $22.4 \times 22.4\,\mathrm{m}^2$. The selected size of this region was based on two considerations: it had to be sufficiently large to capture enough texture in the bathymetry, and sufficiently small to avoid covering many different habitat classes. We note that the approach outlined in [10] uses multi-scale features up to a $50 \times 50\,\mathrm{m}^2$ area.

Matched multimodal data was obtained by extracting a bathymetry patch centred at the AUV position corresponding to each image. The AUV navigation accuracy is comparable to the bathymetric grid spacing, and the habitats of interest typically vary at much larger scales. We therefore assume that any potential misregistration between the images and bathymetry (from localisation errors) will have a minimal effect on the relationship between the two modalities.

The expert labels for the AUV image data were consolidated into 5 habitat classes, characterised by keywords "sand", "screw shells / rubble", "reef / sand interface", "reef", and "kelp" (*Ecklonia Radiata*). Examples of the images and bathymetric patches corresponding to each class are shown in Fig. 2. The labelled dataset contained $68,700$ image / bathymetric pairs, split equally into a training and

test set. For the habitat mapping and sampling results, we use a subset of this data, corresponding to an area of interest known as O'Hara Bluff.

VI. RESULTS

The learned bathymetric features are shown in Fig. 3. Interestingly, the DAE learns edge and gradient filters similar to those obtained from natural image patches [12][14].

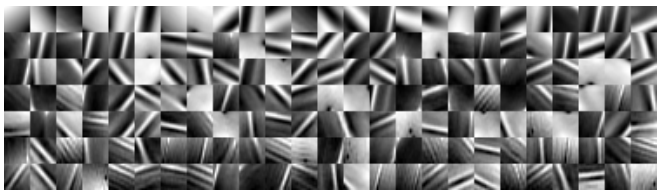


Fig. 3. A subset of the 1000 learned bases for 15×15 bathymetry patches, representing a $22.4 \times 22.4\,\mathrm{m}^2$ area.

A. Classification Performance

For the discriminative training, a regularised logistic regression classifier was used, with a "One-vs-All" scheme for multi-class classification. For each classification scenario, a 3-fold cross validation was performed on the training set for various values of the regularisation hyperparameter. The value that maximised the validation accuracy was then used to retrain the classifier on the entire training set.

To quantify the effect of multimodal learning, two classification scenarios are presented. In the standard encoding scenario, classification is performed directly on the mid-layer features with the available modalities. In the multimodal encoding scenario, the missing modalities are filled in with zeros at the mid-layer, and the entire feature vector is encoded by the multimodal layer and passed into the classifier. The multimodal DAE was trained 10 times, and we report the mean and standard deviations of the subsequent classification results. The performance of the algorithm is shown in Table I, with the classification accuracy and associated F1-score.

As expected, results are significantly better when visual data is present. When comparing the multimodal scenario with the standard scenario, we observe an improvement in performance for the scenarios in which visual data is unavailable. Interestingly, with visual data available, the performance with multimodal encoding is slightly lower than with the standard encoding. This may indicate that the visual features are already fairly precise for this task, and gain little information by encoding their relationship with bathymetric features. The most important result is the benthic habitat mapping case ($\mathcal{B}_0$ and $\mathcal{B}_l$), where the multimodal approach yields a 10% improvement in accuracy. This suggests that the discriminative power of bathymetric data is improved by transformation to a feature space which encodes correlations with visual imagery.

With this improvement, we are able to more accurately perform large-scale habitat classification off the coast of Tasmania, in an area of interest known as O'Hara Bluff (Fig. 4). The classifier used to produce the map was trained on a subset of the matched data: rather than using the image labels from the entire Tasmanian dataset, only the training points and labels within the O'Hara Bluff region were used. White regions represent areas over which bathymetry data was unavailable, and the five classes, from red through to purple, represent sand, screw shells / rubble, reef / sand interface, reef, and kelp. The AUV dive trajectory is overlaid on the habitat map, with the colour at each location representing the ground truth label for the image obtained at that location. A bathymetry map is also shown for comparison, with depth contours plotted.

TABLE I

CLASSIFICATION PERFORMANCE FOR VARIOUS INPUT MODALITIES, REPORTED AS % ACCURACY, WITH F1-SCORE IN PARENTHESES. THE HIGHLIGHTED CASE IS THE BENTHIC HABITAT MAPPING SCENARIO.

Modalities used	Standard Encoding	Multimodal Encoding
$\mathcal{B}_0$ only	59.79% (0.523)	59.79±0.01% (0.523±0.0001)
$\mathcal{B}_l$ only	63.49% (0.582)	70.53±0.08% (0.670±0.0009)
$\mathcal{B}_0$ and $\mathcal{B}_l$	**73.36% (0.703)**	**83.27±0.05% (0.824±0.0005)**
$\mathcal{V}$ only	80.32% (0.791)	79.77±0.04% (0.783±0.0005)
$\mathcal{B}_0$ and $\mathcal{V}$	83.32% (0.825)	83.22±0.05% (0.823±0.0007)
$\mathcal{B}_l$ and $\mathcal{V}$	83.52% (0.827)	80.45±0.07% (0.784±0.0008)
$\mathcal{B}_0$, $\mathcal{B}_l$, and $\mathcal{V}$	86.97% (0.865)	84.71±0.06% (0.835±0.0007)

In general, the habitat map is qualitatively similar to that produced by Bender et al. [10]. Crucially, however, our approach also encodes visual feature information into the classification process. We can also analyse the individual class probabilities to gauge the performance of the model (Fig. 5). As expected, the distribution of kelp correlates strongly with depth, and is most likely to be observed in the shallower waters towards the bottom-left (South-west) corner of the map. Similarly, screw shells / rubble and sand are more likely to be observed in deeper waters towards the East, though they are distributed over a larger region.

B. Generative Sampling Performance

The model can also sample from the underlying data-generating distribution, and we present the results of sampling $\mathcal{B}_0$ and $\mathcal{B}_l$ given an input image $\mathcal{V}$. Some generated samples are shown in Fig. 6, with the model using each input image (top row) to generate samples of encoded depths and bathmetric patchs, respectively (following rows). Recall that the depth features are encoded as a 1-of-k with Gaussian falloff, with 82 bins between the range of $19 - 100$m. Therefore, the depth "signal" should be interpreted as an activation function, where a high activation value suggests a higher likelihood of observing that depth.

To quantitatively analyse the results, the model was used to generate 1000 such samples of $\mathcal{B}_0$ and $\mathcal{B}_l$ for every image in the O'Hara Bluff region. The images were then grouped by class label, and the depth samples averaged over each class to show the distribution over the entire depth range (Fig. 7). Additionally, the rugosity, a measure of roughness or terrain complexity [21], was computed for each generated bathymetry patch, and the mean and standard deviation over each class are shown in Table II.

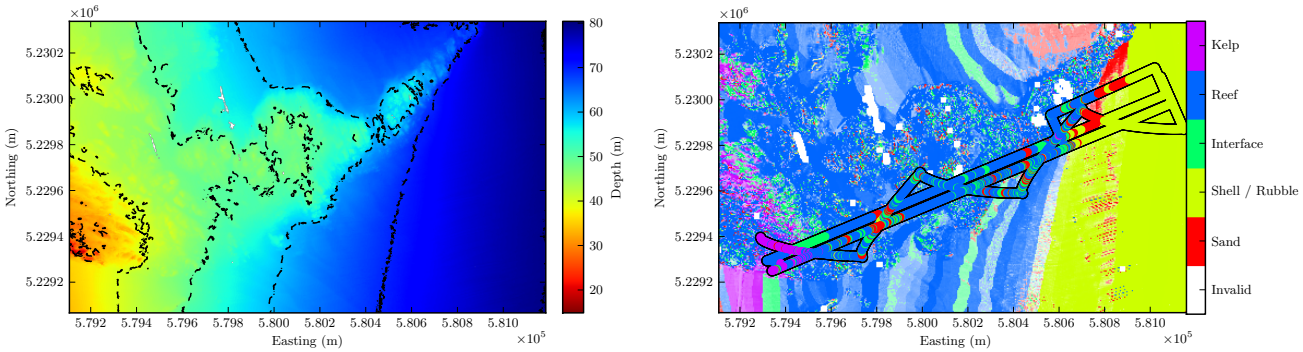
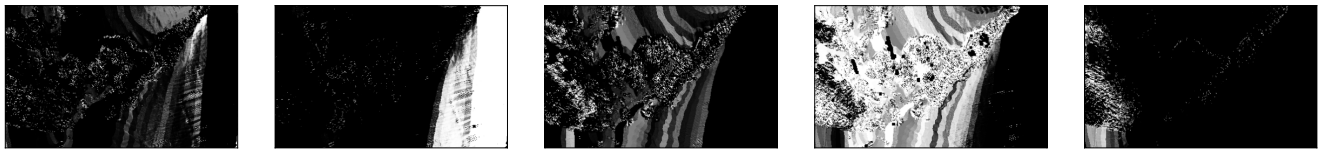


Fig. 4. Left: The bathmetry map over the O'Hara Bluff region. Right: The habitat map of the O'Hara Bluff region produced by classifying $\mathcal{B}_0$ and $\mathcal{B}_l$ features after multimodal encoding. The colour represents the assigned habitat class label, while the intensity of the colour is proportional to the class probability (i.e. white represents equal probabilities over all classes). The class labels from the training / test dataset are also plotted as coloured dots, indicating the coverage of the visual images.



(a) Sand (b) Shells / Rubble (c) Reef / Sand Interface (d) Reef (e) Kelp

Fig. 5. The probability maps for each of the classes, with white areas indicating high probability.

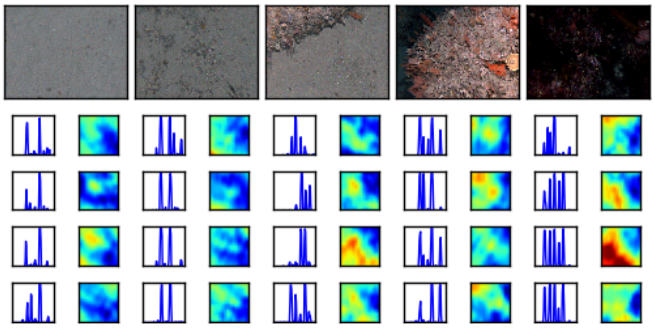


Fig. 6. Bathymetry samples obtained from the learned data-generating distribution, conditioned on the input image. For each input image representing a single class (top row), the subsequent rows display model-generated samples of $\mathcal{B}_0$ in encoded form (left) and $\mathcal{B}_l$ patches (right). Shallower regions are represented as red in the patches, and the $\mathcal{B}_0$ signal should be interpreted as an 'activation function' over the depth range $19 - 100$m).

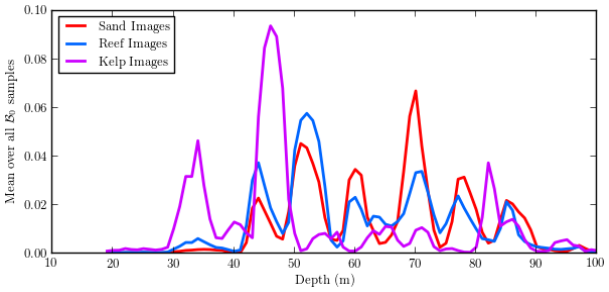


Fig. 7. Average of encoded $\mathcal{B}_0$ (depth) samples conditioned on every image in O'Hara Bluff (1000 samples per image). Samples are grouped by the class of the image used to generate them, and a few key classes are shown here.

TABLE II

RUGOSITY OF $\mathcal{B}_l$ (BATHYMETRY PATCH) SAMPLES, CONDITIONED ON EVERY IMAGE IN O'HARA BLUFF (1000 SAMPLES PER IMAGE), GROUPED BY CLASS.

	Sand	Shell / Rubble	Interface	Reef	Kelp
Mean	1.0773	1.0804	1.0989	1.1218	1.1846
Std	0.0184	0.0307	0.0365	0.0478	0.0580

The results suggest that the model is learning the underlying data distribution. The sampled bathymetric patches are, on average, smoother for the sand classes and more rugose for the reef and kelp images (Table II). Similarly, while the kelp image activates depth features at the lower (shallow) end of the range, higher depths are activated for sand and reef (Fig. 7). We also note that the variation within each class is indicative of the spatial distribution of the class.

VII. CONCLUSIONS

In this paper, we have presented a multimodal approach to perform learning and inference from AUV-based image data and coarse shipborne bathymetry. Our model learns the relationship between both modalities, and is able to naturally deal with missing modalities. As such, the algorithm can capture the visual features of a region, but can still be applied to large-scale bathymetry where visual data is unavailable.

Our results show that by learning the multimodal correlations between visual and bathymetric data beforehand, we can significantly improve classification performance when visual data is unavailable. Results also show that the approach can act as a model for the underlying data-generating distribution, generating plausible samples of bathymetric

features given an input image.

Future work will focus on experimental validation across various other AUV dives, as well as application to other multimodal domains, such as incorporating multi-beam backscatter or hyperspectral imagery. Further, we aim to utilise multimodal inference to perform survey selection and mission planning.

ACKNOWLEDGMENTS

The authors thank Asher Bender, Ariell Friedman, and Daniel Steinberg, for access to some of the datasets and data processing scripts used for this research.

This work was supported by the Australian Research Council (ARC) and the New South Wales and Tasmanian State Governments, and the Integrated Marine Observing System (IMOS) through the DIISR National Collaborative Research Infrastructure Scheme. The authors would like to thank the Captain and crew of the R/V Challenger. Their sustained efforts were instrumental in facilitating successful deployment and recovery of the AUV. Thanks to Justin Hulls and Jan Seiler for help and support on-board the ship and for providing the supervised image labels. The ship-borne multibeam sonar data were collected, processed and gridded to produce DEMs by Geoscience Australia. We also acknowledge the help of all those who have contributed to the development and operation of the IMOS AUV Facility.

REFERENCES

[1] H. Singh, A. Can, R. Eustice, S. Lerner, N. McPhee, O. Pizarro, and C. Roman, "Seabed AUV offers new platform for high-resolution imaging," *Transactions American Geophysical Union*, vol. 85, no. 31, pp. 289–296, 2004.

[2] S. Williams, O. Pizarro, M. Jakuba, and N. Barrett, "AUV benthic habitat mapping in South Eastern Tasmania," in *Field and Service Robotics*, 2010, pp. 275–284.

[3] V. E. Kostylev, B. J. Todd, G. B. Fader, R. Courtney, G. D. Cameron, and R. A. Pickrill, "Benthic habitat mapping on the Scotian Shelf based on multibeam bathymetry, surficial geology and sea floor photographs," *Marine Ecology Progress Series*, vol. 219, pp. 121–137, 2001.

[4] D. Steinberg, "An unsupervised approach to modelling visual data," Ph.D. dissertation, University of Sydney, 2013.

[5] C. J. Brown, S. J. Smith, P. Lawton, and J. T. Anderson, "Benthic habitat mapping: A review of progress towards improved understanding of the spatial ecology of the seafloor using acoustic techniques," *Estuarine, Coastal and Shelf Science*, vol. 92, no. 3, pp. 502–520, 2011.

[6] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *The Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, 2010.

[7] M. Marcos, M. Soriano, and C. Saloma, "Classification of coral reef images from underwater video using neural networks," *Optics Express*, vol. 13, no. 22, pp. 8766–71, 2005.

[8] O. Beijbom, P. J. Edmunds, D. I. Klinez, B. G. Mitchellz, and D. Kriegman, "Automated Annotation of Coral Reef Survey Images," in *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2012.

[9] D. M. Steinberg, S. B. Williams, O. Pizarro, and M. V. Jakuba, "Towards autonomous habitat classification using Gaussian Mixture Models," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2010, pp. 4424–31.

[10] A. Bender, S. B. Williams, and O. Pizarro, "Classification with probabilistic targets," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2012, pp. 1780–86.

[11] P. Rigby, O. Pizarro, and S. B. Williams, "Toward adaptive benthic habitat mapping using gaussian process classification," *Journal of Field Robotics*, vol. 27, no. 6, pp. 741–758, 2010.

[12] A. Coates, H. Lee, and A. Y. Ng, "An analysis of single-layer networks in unsupervised feature learning," *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2010.

[13] G. Hinton, "A practical guide to training restricted boltzmann machines," Department of Computer Science, University of Toronto, Tech. Rep., 2010.

[14] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th Annual Int. Conf. on Machine Learning*, 2009, pp. 609–616.

[15] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th Annual Int. Conf. on Machine Learning*, 2011, pp. 689–696.

[16] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Advances in Neural Information Processing Systems*, vol. 25, 2012, pp. 2231–2239.

[17] G. Irie, D. Liu, Z. Li, and S.-F. Chang, "A bayesian approach to multimodal visual dictionary learning," in *IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 329 – 336.

[18] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2009, pp. 1794–1801.

[19] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural computation*, vol. 23, no. 7, pp. 1661–1674, 2011.

[20] M. Spinoccia, "Bathymetry grids of south east tasmania shelf," *Geosciences Australia*, 2011. [Online]. Available: http://www.ga.gov.au/marine/bathymetry.html

[21] A. Friedman, O. Pizarro, S. B. Williams, and M. Johnson-Roberson, "Multi-scale measures of rugosity, slope and aspect from benthic stereo image reconstructions," *PloS one*, vol. 7, no. 12, 2012.