2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Improved Benthic Classification using Resolution Scaling and SymmNet Unsupervised Domain Adaptation

Heather Doig[1], Oscar Pizarro[1,2] and Stefan B. Williams[1]

***Abstract*— Autonomous Underwater Vehicles (AUVs) conduct regular visual surveys of marine environments to characterise and monitor the composition and diversity of the benthos. The use of machine learning classifiers for this task is limited by the low numbers of annotations available and the many fine-grained classes involved. In addition to these challenges, there are domain shifts between image sets acquired during different AUV surveys due to changes in camera systems, imaging altitude, illumination and water column properties leading to a drop in classification performance for images from a different survey where some or all these elements may have changed. This paper proposes a framework to improve the performance of a benthic morphospecies classifier when used to classify images from a different survey compared to the training data. We adapt the SymmNet state-of-the-art Unsupervised Domain Adaptation method with an efficient bilinear pooling layer and image scaling to normalise spatial resolution, and show improved classification accuracy. We test our approach on two datasets with images from AUV surveys with different imaging payloads and locations. The results show that generic domain adaptation can be enhanced to produce a significant increase in accuracy for images from an AUV survey that differs from the training images.**

I. Introduction

Autonomous Underwater Vehicles (AUVs) are used to conduct regular visual surveys of the marine environment to measure and monitor changes in the benthic environment due to stresses such as pollution, over-fishing and climate change [1], [2]. Typically, marine scientists label the images with point annotations to measure the presence and diversity of benthic species and physical features using a morphological hierarchy such as in [3]. Machine learning classifiers promise to make this process more efficient and increase the amount of information gained from a survey [4], [5] but there are typically only small amounts of labelled annotations available for training.

Classifiers of morphospecies do not transfer well between image datasets due to domain shift between different AUV surveys [3], [6], [5]. Domain shift arises from differences in cameras, imaging altitude, illumination and water column properties. In addition, classifying benthic species and physical features is a fine-grained classification problem as there are many classes with high intra-class variations and also similar inter-class features, increasing the difficulty of the classification task [7], [8].

[1]The authors are with the Australian Centre for Field Robotics, University of Sydney, NSW Australia, (h.doig, o.pizarro, stefanw)@acfr.usyd.edu.au

[2]O. Pizarro is also with the Marine Technology Department, Norwegian University of Science and Technology, Trondheim, Norway.

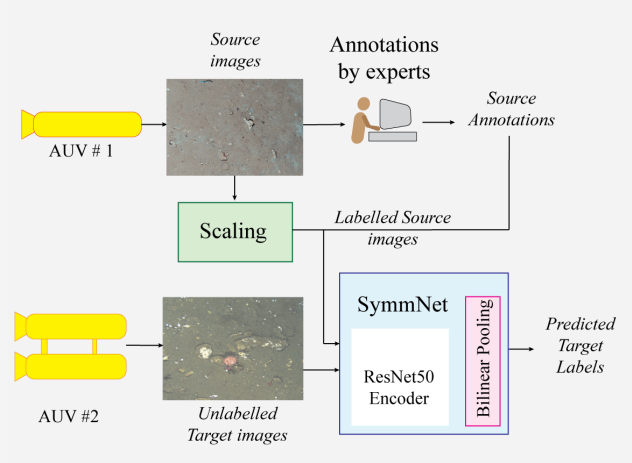

Fig. 1. **Overview** Framework to improve the transferability of a benthic morphospecies classifier when using labelled source images from one AUV survey for training and testing on unlabelled target images from another AUV survey. The framework includes a) scaling images to match spatial resolution between source and target b) using Unsupervised Domain Adaptation (SymmNet, [9]) on the labelled source images and unlabelled target images to train a classifier for source and target images and c) using an efficient bilinear pooling layer [10] for more discriminative features. The framework improves classification accuracy compared to training with the unscaled source images alone.

To address the challenges of making more efficient use of AUV imagery with scarce label annotations, this paper proposes a framework to explore and understand the use of Unsupervised Domain Adaptation (UDA) [11], [12] with the aim of increasing performance of classifiers trained with data from different AUV surveys. This process would allow labelling effort for images from one AUV survey to be migrated to a new survey with new operating parameters arising from changes to cameras, lighting or imaging altitude. Classification performance will also be improved by scaling images so that images have the same spatial resolution [13] and adding an efficient bilinear pooling layer to generate more discriminative features [10]. Figure 1 illustrates the proposed framework.

Performance of a machine learning model trained on a source domain, such as a classifier, can drop when applied to a target domain due to a shift in the distribution of the features. UDA reduces the difference between labelled source domain data and unlabelled target domain data. A successful range of methods update the feature representation for the target domain using either statistics of the distributions [11], [14] or adversarial training [12], [15].

Adversarial UDA uses a domain discriminator to distin-

guish between features from the source and target domain. An adversarial loss updates the network during training to align the features from the source and target domain. This paper uses state-of-the-art UDA, SymmNet [9], which has demonstrated success on a similarly challenging problem of classifying aerial habitats from different drone imagery [16]. SymmNet uses an adversarial loss applied at the class level producing alignment of the source and target distribution at both the domain and class level.

This paper makes the following contributions:

- We propose a framework using SymmNet UDA, and scaling to improve the classification of point annotations from images taken from different AUV platforms or locations.
- We explore the impact to classification performance when an efficient bilinear pooling layer that requires a small number of parameters with fast inference time is applied.
- We provide results and analysis from two datasets from benthic surveys both with images from two different AUV surveys.
- We introduce two curated datasets of benthic image patches to support further research into UDA between AUV surveys[1].

The remainder of the paper is structured as follows. Section II describes related work in Unsupervised Domain Adaptation, bilinear pooling and scaling spatial resolution. Section III describes the implementation of SymmNet and Two-Level Kronecker Product Factorization bilinear pooling as well as the two datasets used to demonstrate the framework. The results of the experiments are presented in Section IV followed by some concluding remarks and suggestions for future research in Section V. Our code is available at `https://github.com/hdoi5324/benthic-uda`.

II. Related Work

A. Unsupervised Domain Adaptation

UDA is a machine learning technique used to reduce the domain shift between labelled *source* data and unlabelled *target* data, such that a domain-adapted classifier originally trained with labelled source data can be used with the target data. The difference in the distribution between source and target data can been reduced by aligning the feature representations using higher order statistics [11], [14] or adversarial training [12], [15]. We focus on adversarial training techniques which have delivered state-of-the-art results for image classification. [12] and [15] align the feature representations using a discriminator network to distinguish samples between domains while using a confusion loss to adapt the target encoder. In this paper, we use SymmNet [9], [17] which replaces the domain discriminator with a multi-class discriminator loss giving more separable features at a class level.

[1]`https://data.mendeley.com/datasets/d2yn52n9c9`

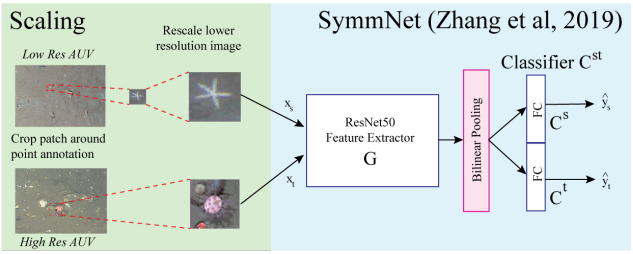

Fig. 2. Network diagram for full framework. Scaling of the lower resolution AUV patch is performed prior to training. The SymmNet network uses ResNet50 backbone for the encoder followed by the bilinear pooling layer then the symmetric source and target classifier. The bilinear pooling layer uses a Two-Level Product Factorization to calculate a projection of the bilinear product of the feature map from the encoder.

B. Bilinear Pooling

Bilinear pooling has been shown to provide more discriminative features from a deep learning Convolutional Neural Network (CNN) model [18], [19]. Applying this pooling layer between CNN modules or after the final CNN module has increased the accuracy of fine-grained and texture based classification tasks [20]. For example, [21] improved the retrieval of remote sensing images by adding a bilinear pooling layer after the CNN backbone. The resulting features provided more detail about the scene despite variations in atmospheric conditions, illumination and viewing angles which is similar to the variations that arise in benthic images. This work uses Two-Level Kronecker Product Factorization bilinear pooling [10] which aims to use fewer parameters than other implementations [18], [21]. Fewer parameters reduces computational resources and inference time which may allow the method to be used onboard an AUV providing real-time results.

C. Scaling

Adjusting images to a common spatial resolution can increase the performance of a CNN when a small number of annotations are available for training [22], [13]. Spatial resolution may differ between AUV surveys due to changes to the camera (e.g. imager size and/or resolution, lens field of view, etc) and altitude used during the mission. [13] performed scale adjustment on images from different AUVs using altitude and spatial resolution per pixel, improving the generalization error in an image segmentation task.

III. Method

The proposed framework is comprised of a classifier network based on SymmNet [9]. The classifier has a ResNet50 backbone [23] with the pooling layer after Layer 4 CNN module replaced with a bilinear pooling layer described below. Figure 2 provides an overview of the three main components of the framework being scaling, the SymmNet classifier and bilinear pooling which will now be described in more detail.

A. Scaling

The input to the network is an image patch cropped around a human-generated point annotation and scaled to the same spatial resolution and pixel size (224x224) for the source and target images. Table III provides the spatial resolutions for the original images and the crop size used.

B. SymmNet Classifier

The Domain-Symmetric Network or SymmNet by [9] is an UDA network that uses adversarial training to adapt a target classifier using labelled source data and unlabelled target data. The network uses a symmetric design with a combination of loss terms using domain and class level confusion loss to improve the adaptation of the target classifier to the target domain distribution as well as the class distribution, improving on previous methods such as [15].

The network comprises a shared feature extractor G using a ResNet50 backbone and pooling layer [23] followed by a fully connected classifier C^{st} that provides classifications for both target and source images. Its symmetric design combines a source classifier C^s and the target classifier C^t each with K outputs where K is the number of classes. When using bilinear pooling, the average pooling layer after Layer 4 of ResNet is replaced with the bilinear pooling described in the next section.

The following detail provides a brief overview of the losses that are used to train the feature extractor and classifier while also adapting the target classifier to the unlabelled target data. x denotes the input data being the image and y is the label for the input image. The source data is defined as $D_s = (\mathbf{x}_i^s, y_i^s)_{i=1}^{n_s}$ with n_s labelled samples and the unlabelled target data is $D_t = (\mathbf{x}_j^t)_{j=1}^{n_t}$ with n_t unlabelled samples.

Firstly, the classifiers C^s and C^t are each trained using the labelled source data with a cross-entropy loss as is common for a classifier:

$$\min_{C^{dom}} E_{cls}^{dom}(G, C^{dom}) = -\frac{1}{n_s}\sum_{i=1}^{n_s} log(p_{y_i^s}^{dom}(\mathbf{x}_i^s)) \quad (1)$$

where $p^{dom} = C^{dom}(G(\mathbf{x}))$ and dom is the source s or target t domain. Domain discrimination in the next loss terms adapt the target classifier C^t so that it differs from the source classifier C^s.

Instead of using a separate domain discriminator network as in [15], the discrimination training uses the sum of probabilities for source samples through C^s and target samples through C^t to train the C^{st} classifier using the following two-way cross-entropy loss:

$$\min_{C^{st}} E_{dom}^{st}(G, C^{st}) = -\frac{1}{n_t}\sum_{j=1}^{n_t} log(\sum_{k=1}^{K} p_{k+K}^{st}(\mathbf{x}_j^t)) - \frac{1}{n_s}\sum_{i=1}^{n_s} log(\sum_{k=1}^{K} p_k^{st}(\mathbf{x}_i^s)) \quad (2)$$

where $p^{st} = C^{st}(G(\mathbf{x}))$ and K is the number of classes. The source and target classifiers now discriminate between source and target samples.

The following confusion loss term discriminates at the class level between the source and target domain leading to better adaptation of the target classifier at a class level.

$$\min_G F_{class}^{st}(G, C^{st}) = -\frac{1}{2n_s}\sum_{i=1}^{n_s} log(p_{y_i^s+K}^{st}(\mathbf{x}_i^s)) - \frac{1}{2n_s}\sum_{i=1}^{n_s} log(p_{y_i^s}^{st}(\mathbf{x}_i^s)) \quad (3)$$

A domain level confusion loss using the unlabelled target data compares the sum of probabilities from source and target classifiers of the combined C^{st} classifier using the target data:

$$\min_G F_{dom}^{st}(G, C^{st}) = -\frac{1}{2n_t}\sum_{j=1}^{n_t} log(\sum_{k=1}^{K} p_{k+K}^{st}(\mathbf{x}_j^t)) - \frac{1}{2n_t}\sum_{j=1}^{n_t} log(\sum_{k=1}^{K} p_k^{st}(\mathbf{x}_j^t)) \quad (4)$$

The final loss uses an entropy minimization objective to increase the classification task performance by summing the probabilities for each matching class in the combined C^{st} classifier:

$$\min_G M^{st}(G, C^{st}) = -\frac{1}{n_t}\sum_{j=1}^{n_t}\sum_{k=1}^{K} q_k^{st}(\mathbf{x}_j^t) log(q_k^{st}(\mathbf{x}_j^t)) \quad (5)$$

where $q_k^{st}(\mathbf{x}_j^t) = p_k^{st}(\mathbf{x}_j^t) + p_{K+k}^{st}(\mathbf{x}_j^t), k \in \{1, \cdots, K\}$.

These losses combine to create two overall losses to minimise:

$$\min_{C^s, C^t, C^{st}} E_{cls}^s(G, C^s) + E_{cls}^t(G, C^t) + E_{dom}^{st}(G, C^{st}),$$
$$\min_G F_{class}^{st}(G, C^{st}) + \lambda(F_{dom}^{st}(G, C^{st}) + M^{st}(G, C^{st})) \quad (6)$$

where $\lambda \in [0, 1]$ is a trade-off parameter that is increased during training.

C. Bilinear Pooling

Bilinear pooling calculates a more discriminative feature $\mathbf{b}$ from the outer product of an input feature map ($\mathbf{X}\mathbf{X}^\mathbf{T}$) but this approach is highly memory intensive. The bilinear pooling layer in this paper uses the low parameter count Two-Level Kronecker Product Factorization (TKPF) method [10] to approximate a projection of the bilinear product. The configuration used in this work adds only 4K parameters to the ResNet50 backbone which has 23M parameters. Two smaller scale matrices $\mathbf{A}$ and $\mathbf{B}$ are learned to create a projection of the bilinear product of the input feature map $\mathbf{X}$. The network is repeated q times and the results are averaged to increase the representative capability. The parameters a, b, r and q are set at training. This process is visualized in Figure 3 and a high-level description is provided below. The complete factorization method can be found in [10].

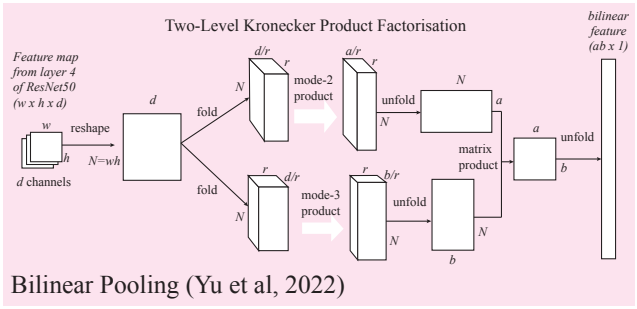

Fig. 3. Network diagram for the bilinear pooling layer with Two-level Kronecker Product Factorization.

A projection matrix $\mathbf{P}$ can be used to generate a lower dimension bilinear feature:

$$\mathbf{b} = \mathbf{P}vec(\mathbf{X}\mathbf{X}^T) \tag{7}$$

where $\mathbf{X}$ is the feature map from the last layer of ResNet. $\mathbf{P}$ is constrained to be a Kronecker product as this reduces a very larger matrix to two smaller matrices:

$$\mathbf{P} = \mathbf{A} \otimes \mathbf{B} \tag{8}$$

By substituting in the Kronecker product into Equation 7, there can be a further factorization using the 'vec' trick described in [10]:

$$(\mathbf{A} \otimes \mathbf{B})vec(\mathbf{X}\mathbf{X}^T) = vec(\mathbf{S}\mathbf{T}^T) \tag{9}$$

where $\mathbf{S} = \mathbf{B}\mathbf{X}$ and $\mathbf{T} = \mathbf{A}\mathbf{X}$. A second level of Kronecker product can be used to factorize further:

$$\mathbf{A} = \mathbf{I_r} \otimes \hat{\mathbf{A}}, \mathbf{B} = \hat{\mathbf{B}} \otimes \mathbf{I_r} \tag{10}$$

allowing $\mathbf{S}$ and $\mathbf{T}$ to be calculated efficiently using modal folding and tensor modal products. $\mathbf{X}$ is folded into $\mathcal{X}_a \in \mathbb{R}^{N \times \frac{d}{r} \times r}$ and $\mathcal{X}_b \in \mathbb{R}^{N \times r \times \frac{d}{r}}$. After folding, the tensors are multiplied using the tensor modal product:

$$\mathbf{T} = \mathcal{X}_a \times_2 \hat{\mathbf{A}} \tag{11}$$

$$\mathbf{S} = \mathcal{X}_b \times_3 \hat{\mathbf{B}} \tag{12}$$

The matrix product gives the projected bilinear feature:

$$\hat{\mathbf{b}} = vec(\mathbf{S}\mathbf{T}^T) \tag{13}$$

with $\hat{\mathbf{b}} \in \mathbb{R}^{ab}$.

Element-wise signed square root and L2 normalization is applied to $\hat{\mathbf{b}}$. The features from the q duplicated networks are averaged giving the bilinear feature for the classifier. A dropout layer with probability of 0.7 is used before the classifier to add regularization. Dropout of 0.7 is a common starting amount and was not tuned further.

D. Datasets

The proposed framework was used on two datasets with images taken from two different AUV surveys [24].

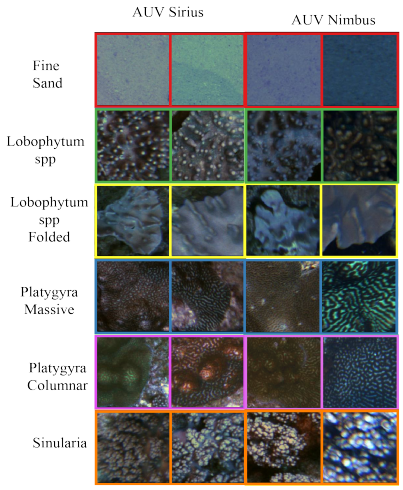

Fig. 4. Samples of the six classes from the EMR dataset.

TABLE I
NUMBER OF SAMPLES FOR EMR DATASET

Survey Name	NG06	SS07	SS09
AUV	*Nimbus*	*Sirius*	*Sirius*
Fine Sand	135	159	66
Lobophytum spp	153	199	60
Lobophytum spp Folded	152	194	70
Platygyra Massive	225	125	62
Platygyra Columnar	196	163	70
Sinularia	258	278	93
Total Samples	1119	1118	421

1) Elizabeth and Middleton Reef: The first dataset was collected at Elizabeth and Middleton Reef (EMR) during a marine environmental cruise completed in January 2020 using the AUV *Sirius* and AUV *Nimbus* [25]. Manual point annotations of the five most common coral species as well as sand were used to create the dataset. The images are from three separate surveys by the AUVs. Figure 4 shows examples of the six classes and Table I shows the count of point annotations by class and survey.

2) South Hydrate Ridge: The second dataset was from the South Hydrate Ridge (SHR) collected on the Adaptive Robotics cruise in the Eastern Pacific Ocean in 2018 [26]. This dataset includes three physical features (Bacterial Mat, Sand/Mud and Rock) and five benthic species (Soft Coral, Sea Star, Crab, Rockfish and Sole) collected by the AUV *AE2000f* and the AUV *Tuna-sand* at depths of around 800 m. The AUV *AE2000f* has a very large image footprint operating at around 6m altitude while the AUV *Tuna-sand* has high-resolution cameras working at 2 m altitude. Figure 5 provides samples of each class taken by the two AUVs as well as the scaled patch for the AUV *AE2000f* and Table II provides the number of samples in each class by AUV.

E. Experiments

For each dataset, each source and target pair was trained with all combinations of with or without scaling, bilinear pooling and SymmNet. This resulted in eight possible combinations for the six source-target pairs. When SymmNet was

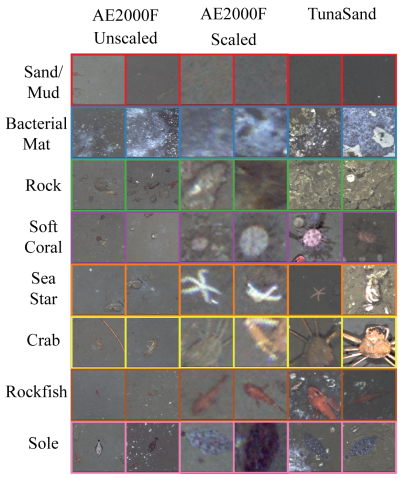

Fig. 5. Samples from the 8 classes in the SHR dataset. Each row shows two unscaled patches from AUV *AE2000f*, two patches from AUV *AE2000f* scaled to the same spatial resolution as AUV *Tuna-sand*, followed by two patches from AUV *Tuna-sand*. This shows how the same spatial scale displays species at similar size as seen in the Soft Coral and Rockfish samples.

TABLE II
NUMBER OF SAMPLES FOR SHR DATASET

AUV	*AE2000f*	*Tuna-sand*
Sand	65	57
Bacterial Mat	60	60
Rock	62	54
Soft Coral	59	49
Sea Star	67	58
Crab	64	52
Rockfish	71	66
Sole	64	41
Total Samples	512	437

not used, only the losses in Equation 1 and 3 were used for training with the labelled source data.

No scaling used a crop of 224x224 pixels for both domain datasets. When scaling was used, it was applied to the domain with the lower resolution by cropping to the scaled crop size in Table III and resizing the patch to 224x224. The spatial resolutions for the EMR dataset were calculated using the average mission altitude, the calibrated focal length in pixels and the camera sensor pixel size. The resolutions for the SHR dataset are from [27]. Note that for the SS09 and SS07 domain pair there was no scaling required as they were from the same AUV platform operating at the same average altitude so have the same spatial resolution already.

TABLE III
AUV RESOLUTION

Dataset	AUV	Resolution (mm/pixel)	Scaled Crop (pixels)
EMR	Nimbus	0.65 @ 2m	148
	Sirius	0.43 @ 2m	224
SHR	AE2000f	6 @ 6m	32
	Tuna-sand	0.8 @ 2m	224

F. Training

The training protocols from previous UDA studies [9], [14], [17] were used. All labelled source data and unlabelled target data were used for training SymmNet. The data was augmented by adding a horizontal flip as used in [9]. The same model and training parameters were used for all experiments on both datasets.

The initial learning rate was 0.02 for the classifier and pooling layer while the pre-trained ResNet50 layers had a learning rate one-tenth lower. The learning rate followed an annealing strategy as in [9] which reduced the learning rate according to $\eta_p = \frac{\eta_0}{(1+\alpha p)^\beta}$ where p is the ratio of current epoch to total epochs, η_0 = 0.02, α = 10 and β = 1.5.

Following the two-phase training approach of [10] when training the bilinear pooling layer, the learning rate was fixed at 0.1 for the first 3 epochs, followed by a learning rate of 0.02 using the annealing strategy described earlier. The parameters for the bilinear pooling layer were a and b = 64, r = 16 and q = 4.

The average accuracy from three training runs for classifying all the target data was calculated using the model at the final epoch. Batch size was 32 and training was for 50 epochs using a Stochastic Gradient Descent optimiser with momentum = 0.9.

IV. RESULTS AND DISCUSSION

Table IV shows the results of the experiments on each source and target pair. The average accuracy for the unlabelled target data from the last epoch of training is shown with the best accuracy in red and the second best in blue. Applying either scaling or SymmNet on their own always produced an increase in accuracy. Using scaling and SymmNet increased accuracy by up to 28% compared to using the classifier trained without scaling or domain adaptation. For the majority of source-target domain pairs, the top two results used both scaling and SymmNet while bilinear pooling did not provide a consistent improvement.

The most significant improvement in accuracy occurred with the SHR dataset where there is a larger difference in spatial resolution between the AUV surveys. This result demonstrates the ability to use manual annotations from platforms with lower resolution to classify images from higher resolution payloads by reducing the domain shift. This is relevant to real-world, long-term use of survey platforms that are upgraded after extensive annotation of older data sets.

Replacing the ResNet average pooling layer with the bilinear pooling layer only improved accuracy in five of the eight scaled results and none of the unscaled results. The bilinear pooling layer does not appear to learn a feature that is more discriminative than the average pooling layer that is used by ResNet. This may be due to the hyper-parameters and losses being used to train the network. A specific training approach and losses for this layer may improve the result.

Finding the optimal parameters for UDA through model tuning and selection is not possible as there is no labelled

TABLE IV

CLASSIFICATION ACCURACY OF TARGET DOMAIN FOR ALL SOURCE→TARGET PAIRS WITH COMBINATIONS OF SCALING, BILINEAR POOLING AND SYMMNET UDA. THE BEST RESULT IS SHOWN IN RED AND THE SECOND BEST RESULT IS SHOWN IN BLUE.

Scaling	Bilinear Pool	SymmNet	EMR dataset						SHR dataset	
			NG06→SS07	SS07→NG06	NG06→SS09	SS09→NG06	SS07→SS09	SS09→SS07	AE2000f→Tuna-sand	Tuna-sand→AE2000f
			71.86	70.41	65.41	64.30			63.92	54.52
	✓		70.29	70.32	64.46	63.71			60.79	50.88
		✓	77.55	79.47	73.13	74.44			84.44	57.44
	✓	✓	78.88	77.22	72.97	74.77			77.12	63.29
✓			75.51	75.48	69.50	72.42	71.47	69.16	74.14	73.50
✓	✓		74.50	77.12	71.63	72.68	72.42	69.36	67.81	66.60
✓		✓	79.43	80.19	76.67	74.26	75.65	72.64	93.59	83.27
✓	✓	✓	80.48	79.27	75.26	76.41	73.13	74.50	91.91	78.58

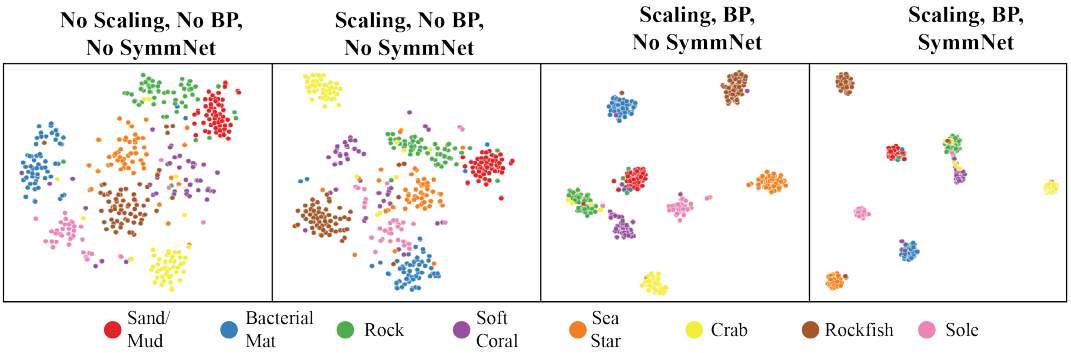

Fig. 6. A t-SNE visualization of the features for AUV *Tuna-sand* data for AE2000f→Tuna-sand source-target pair. The plots show the effect of adding scaling, bilinear pooling and SymmNet corresponding to lines 1, 5, 6 and 8 in Table IV.

target data. More research on suitable metrics or validation process without labelled target data such as [12] or [28] may allow the network to be tuned to provide more discriminative features.

Figure 6 shows a t-SNE visualization of the features for the AUV *Tuna-sand* by class for four of the results with the *AE2000f→Tuna-sand* source-target pair. As can be seen, the clusters for each class become more compact showing features that are more discriminative at a class level as scaling, bilinear pooling and SymmNet are added. This is particularly evident with the introduction of SymmNet, which is able to use the unlabelled target domain information to find a feature space that is discriminative at a class level despite some of the classes being visually similar. Some clusters have co-mingled classes showing that the images are not always labelled correctly. A more accurate source model trained with a larger amount of labelled source data may reduce the number of incorrect labels in these clusters.

V. CONCLUSION

This paper investigated whether performance could be improved for a classifier trained on labelled source data and unlabelled target data from AUV surveys with differing payloads, location and altitudes by applying a framework of resolution scaling, bilinear pooling and SymmNet Unsupervised Domain Adaptation. Using scaling and SymmNet UDA to classify images from two benthic datasets from different AUV surveys consistently improved accuracy and the separation of classes in the feature space. The framework can increase the value and longevity of a classifier trained on manual point annotations from previous AUV surveys and applied to higher resolution images from upgraded AUV payloads. While accuracy was improved with both scaling and SymmNet, by reducing the domain shift between source and target data, tuning of model and training parameters particularly for the bilinear pooling layer may increase these gains. Other combinations of state-of-the-art UDA such as Source Hypothesis Transfer (SHoT) [29] and different implementations of bilinear pooling could be investigated to provide further improvements to the framework.

ACKNOWLEDGMENT

The images from the EMR dataset were part of Australia's Integrated Marine Observing System (IMOS), enabled by the National Collaborative Research Infrastructure Strategy (NCRIS). The EMR annotations were based on work undertaken for the Marine Biodiversity Hub, a collaborative partnership supported through funding from the Australian Government's National Environmental Science Program (NESP). The images for the SHR dataset were collected during the Schmidt Ocean Institute's FK180731 #Adaptive Robotics campaign, with support from the Japanese Government's Zipangu in the Ocean Strategic Innovation Program.

REFERENCES

[1] J. Monk, N. S. Barrett, D. Peel, E. Lawrence, N. A. Hill, V. Lucieer, and K. R. Hayes, "An evaluation of the error and uncertainty in epibenthos cover estimates from AUV images collected with an efficient, spatially-balanced design," *PLoS ONE*, vol. 13, no. 9, 2018.

[2] N. Perkins, J. Monk, and N. Barrett, "Analysis of a time-series of benthic imagery from the South-east Marine Parks Network," Institute of Marine and Antarctic Studies, Tech. Rep., 2021.

[3] G. Pavoni, M. Corsini, N. Pedersen, V. Petrovic, and P. Cignoni, "Challenges in the deep learning-based semantic segmentation of benthic communities from ortho-images," *Applied Geomatics*, vol. 13, no. 1, pp. 131–146, 2021.

[4] M. González-Rivero, O. Beijbom, A. Rodriguez-Ramirez, D. E. P. Bryant, A. Ganase, Y. Gonzalez-Marrero, A. Herrera-Reveles, E. V. Kennedy, C. J. S. Kim, S. Lopez-Marcano, K. Markey, B. P. Neal, K. Osborne, C. Reyes-Nivia, E. M. Sampayo, K. Stolberg, A. Taylor, J. Vercelloni, M. Wyatt, and O. Hoegh-Guldberg, "Monitoring of coral reefs using artificial intelligence: A feasible and cost-effective approach," *Remote Sensing*, vol. 12, no. 3, 2020.

[5] I. D. Williams, C. Couch, O. Beijbom, T. Oliver, B. Vargas-Angel, B. Schumacher, and R. Brainard, "Leveraging automated image analysis tools to transform our capacity to assess status and trends on coral reefs," *Frontiers in Marine Science*, vol. 6, no. APR, 2019.

[6] D. Langenkämper, R. van Kevelaer, A. Purser, and T. W. Nattkemper, "Gear-induced concept drift in marine images and its effect on deep learning classification," *Frontiers in Marine Science*, vol. 7, 2020.

[7] N. Ani Brown Mary and D. Dharma, "Coral reef image classification employing Improved LDP for feature extraction," *Journal of Visual Communication and Image Representation*, vol. 49, pp. 225–242, 2017.

[8] H. Zhu, W. Ke, D. Li, J. Liu, L. Tian, and Y. Shan, "Dual cross-attention learning for fine-grained visual categorization and object re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4692–4702.

[9] Y. Zhang, H. Tang, K. Jia, and M. Tan, "Domain-symmetric networks for adversarial domain adaptation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, 2019, pp. 5026–5035.

[10] T. Yu, Y. Cai, and P. Li, "Efficient Compact Bilinear Pooling via Kronecker Product," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, pp. 3170–3178, 2022.

[11] B. Gong, K. Grauman, and F. Sha, "Learning kernels for unsupervised domain adaptation with applications to visual object recognition," *International Journal of Computer Vision*, vol. 109, no. 1-2, pp. 3–27, 2014.

[12] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, p. 2096–2030, 2016.

[13] J. Walker, T. Yamada, A. Prugel-Bennett, and B. Thornton, "The effect of physics-based corrections and data augmentation on transfer learning for segmentation of benthic imagery," in *2019 IEEE International Underwater Technology Symposium, UT 2019 - Proceedings*, 2019.

[14] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *34th International Conference on Machine Learning, ICML 2017*, vol. 5, 2017, pp. 3470–3479.

[15] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, 2017, pp. 2962–2971.

[16] N. Nagananda, A. M. N. Taufique, R. Madappa, C. S. Jahan, B. Minnehan, T. Rovito, and A. Savakis, "Benchmarking domain adaptation methods on aerial datasets," *Sensors*, vol. 21, no. 23, p. 8070, 2021.

[17] Y. Zhang, B. Deng, H. Tang, L. Zhang, and K. Jia, "Unsupervised multi-class domain adaptation: Theory, algorithms, and practice," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[18] T. Y. Lin, A. Roychowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2015 International Conference on Computer Vision, ICCV 2015, 2015, pp. 1449–1457.

[19] S. Kong and C. Fowlkes, "Low-rank bilinear pooling for fine-grained classification," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, 2017, pp. 7025–7034.

[20] L. Liu, J. Chen, P. Fieguth, G. Zhao, R. Chellappa, and M. Pietikäinen, "From BoW to CNN: Two Decades of Texture Representation for Texture Classification," *International Journal of Computer Vision*, vol. 127, no. 1, pp. 74–109, 2019.

[21] Y. Wang, S. Ji, M. Lu, and Y. Zhang, "Attention boosted bilinear pooling for remote sensing image retrieval," *International Journal of Remote Sensing*, vol. 41, no. 7, pp. 2704–2724, 2020.

[22] M. Zurowietz and T. W. Nattkemper, "Unsupervised knowledge transfer for object detection in marine environmental monitoring and exploration," *IEEE Access*, vol. 8, pp. 143 558–143 568, 2020.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, 2016, pp. 770–778.

[24] H. Doig, S. Williams, and O. Pizarro, "Dataset: Application of SymmNet Unsupervised Domain Adaptation and Resolution Scaling for Improved Benthic Classification," 2022, Mendeley Data, v3, http://dx.doi.org/10.17632/d2yn52n9c9.

[25] A. Carroll, J. Monk, N. Barrett, S. Nichol, S. Dalton, N. Dando, J. Siwabessy, A. Leplastrier, H. Evans, and Z. Huang, "Elizabeth and Middleton Reefs, Lord Howe Marine Park, Post Survey Report," Geoscience Australia, Report to the National Environmental Science Program, Marine Biodiversity Hub, 2021.

[26] T. Yamada, A. Prügel-Bennett, and B. Thornton, "Learning features from georeferenced seafloor imagery with location guided autoencoders," *Journal of Field Robotics*, vol. 38, no. 1, p. 52 – 67, 2021.

[27] J. Walker, A. P. Bennett, and B. Thornton, "Towards observation condition agnostic fauna detection and segmentation in seafloor imagery for biomass estimation," in *OCEANS 2021: San Diego – Porto*, 2021, pp. 1–8.

[28] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer joint matching for unsupervised domain adaptation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1410–1417.

[29] J. Liang, D. Hu, Y. Wang, R. He, and J. Feng, "Source Data-absent Unsupervised Domain Adaptation through Hypothesis Transfer and Labeling Transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.