Automatic recognition of underwater munitions from multi-view sonar surveys using semi supervised machine learning: a simulation study

Oscar Bryan[1] Roy Edgar Hansen[2] Tom S. F. Haines[1] Narada Warakagoda[2]
Alan Hunter[1]
1) University of Bath
2) Norwegian Defence Research Establishment

1. ABSTRACT

This paper presents a machine learning technique for using large unlabelled survey datasets to aid automatic classification. We have demonstrated the benefit of this technique on a simulated synthetic aperture sonar (SAS) dataset. We designed a machine learning model to encode a representation of SAS images from which new SAS views can be generated. This novel task requires the model to learn the physics and content of SAS images without the requirement for human labels. This is called self-supervised learning. The pre-trained model can then be fine-tuned to perform classification on a small amount of labelled examples. This is called semi-supervised learning. By using a simulated dataset we can step-by-step increase the realism to identify the sources of difficulty for applying this method to real SAS data, and have a performance bench mark from this more idealised dataset. We have quantified the improved accuracy for the re-view model (ours), against a traditional self-supervised approach (autoencoder), and no pre-training. We have also demonstrated generating novel views to qualitatively inspect the model's learned representation. These results demonstrate our re-view self-supervised task aids the downstream classification task and model interpretability on simulated data, with immediate potential for application to real-world UXO monitoring.

2. INTRODUCTION

Disposing of unwanted material at sea was common practice, most notably munitions were dumped in vast quantities following the Second World War.[1] To survey large areas underwater acoustic sensors (sonar) are required due to electromagnetic signal's (e.g., optical and radar) high attenuation in water. Autonomous underwater vehicles equipped with synthetic aperture sonar, a specific type of acoustic sensor, are capable of surveying large areas at cm resolution.[2] This large amount of data necessitates an automated approach to UXO detection and classification. Automatic approaches to detection and classification for sonar and more generally computer vision primarily use deep learning.[3,4]

Many automatic target recognition for sonar use machine learning (typically convolutional neural networks) in a supervised fashion: using pairs of images and labels to train the model's parameters.[5–8,8–10]

However, supervised learning requires many labelled examples.[11] Large amounts of data can be collected but labelling is time consuming and inaccurate.[12] This motivates training a model on an analogous task using the much larger unlabelled dataset: self supervised learning. This analogous tasks allows the model to learn something useful about the structure of the data, all or part of this model can then be reused for the desired down stream of classification : semi supervised learning. A common analogous task

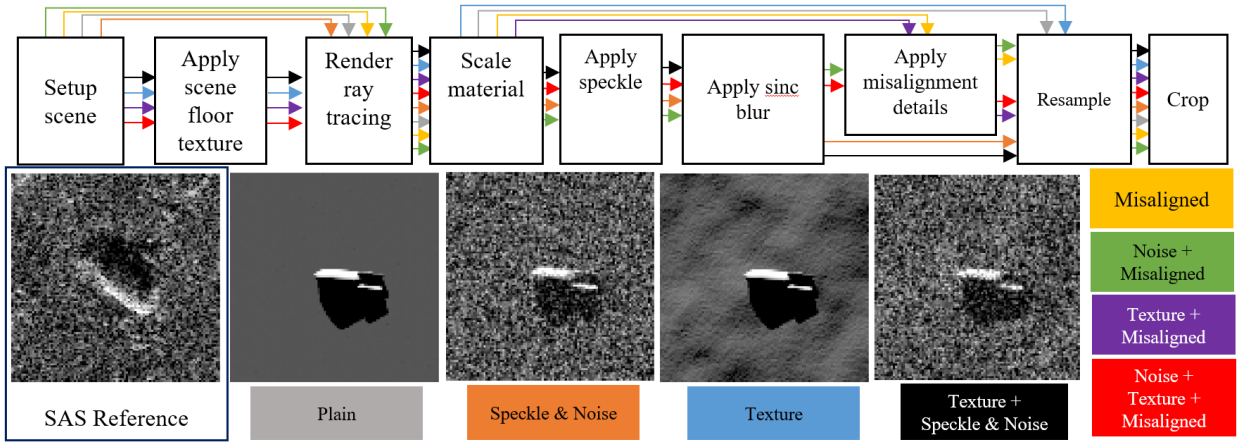

Figure 1: Overall flowchart for generating synthetic sonar data. Different levels of realism are configurable by selecting stages.

is encoding the image content to a compressed representation before recreating the image, this task is performed by two neural networks a encoder and generator together called an autoencoder.[13] Pre-training with an autoencoder has been demonstrated to improve classification of sea-floor types.[9]

The task we propose in this work is predicting the view from a different sonar location. This requires multiple registered sonar views, requiring multiple survey runs, but not human labels. Predicting new views requires the geometry of the scene and physics of the imaging technique to be learnt and can be done in a self supervised fashion. We aim to evaluate the potential of this approach by comparing it to a pre-trained autoencoder, and a randomly initialised model (no-pretraining). To isolate the impact of this new self supervised task the same training data, model architecture and model input was used to evaluate classification accuracy. We demonstrate application of the method on simulated SAS data.

3. METHOD

A. SYNTHETIC DATASET

A synthetic SAS dataset was created to compare semi supervised and supervised approaches. This allowed as much data as we needed to be produced with perfect ground truth. Further, elements of realism could be added to the dataset incrementally to help development of the machine learning architecture and assess the impact of these acoustic specific phenomena on the semi supervised approach.

Ray tracing was used to generate realistic image geometry and noise was then added to match the statistics of real SAS images. Real SAS images are HISAS 1030 images of the Skagerrak UXO dumpsite. The dataset included 8 UXO target classes and one 'nothing' class. The unlabelled dataset contained 18,000 images - 6,000 scenes each imaged "from" three different "SAS" positions. The labelled dataset contained up to 700 training images and 2,500 test images. The overall processing steps are summarised in Figure 1, greater detail on the generation and content of the dataset is given in the following sub-sections.

i. Scene - Ray Tracing

Objects were positioned in the centre of the scene, 3d models of targets 500 kg bomb, 250 kg bomb, 50 kg bomb, 150 mm shell and 105 mm shell or "nothing". The position of the object was selected at random as shown in figure 2 where $A, B \sim U(0\,\text{deg}, 360\,\text{deg})$, $C \sim U(-20\,\text{deg}, 20\,\text{deg})$ and $D \sim U(0\text{ m}, \frac{H}{3}\text{ m})$.

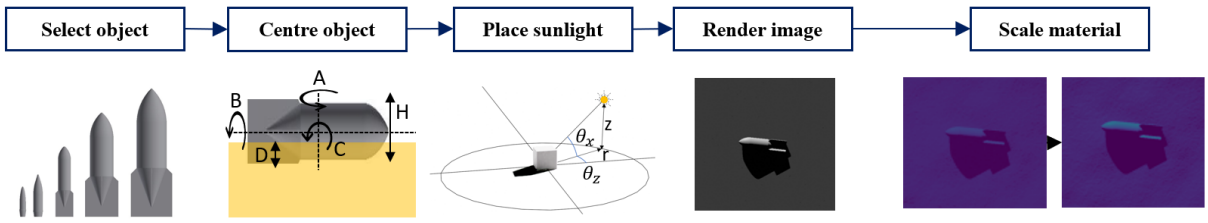

Figure 2: Synthetic data scene setup and rendering process.

A sun lamp is then used to mimic the sonar source, this zero sized light at equivalent infinate distance produces shadow and highlight geometry comparable to side scan or SAS sonar. The "sonar source" has a set height $Z = 25$ m, radius $r \sim U(30m, 170m)$, z rotation $\theta_z \sim U(0\,\text{deg}, 360\,\text{deg})$ and x rotation $\theta_x = \arctan(\frac{z}{r})$

Orthographic camera placed in a downward looking configuration rendered using blender cycles render with zero ray bounces . Rendered at (1 cm) equivalent resolution which is five times that of HISAS sonar system. This high resolution in addition to a wide field of view (10 m) allowed flexibility in the downstream processes (i.e. interpolation when rotating) before being down-sampled to resolution equivalent to the SAS system.

The reflectivity of materials used within the ray tracing software blender do not map to acoustic response of materials at the 100 kHz frequency being considered. We used a smoother material for the munitions resulting in a brighter render. We then normalised pixel intensity by dividing by the median,

$$z = \frac{p_i}{\text{med}(P)} \forall p_i \in P,$$

and made a gamma correction, $z' = z^k$. Shadow regions have zero intensity and the seafloor has unity intensity. Therefore this gamma correction changes the ratio between the seafloor and target. This parameter was set so that the ratio between shadow, seafloor, and target were matched to the real SAS data.

ii. Sea Floor Texture

Another realism element was the addition of a seafloor like texture to the synthetic data background. This was achieved by mixing (weighted average) of a small scale noise (sand) a large scale noise (spatial variation) and wave (ripples) texture. The output of this texture was used to scale the "sea-floor" mesh in the depth direction. We acknowledge this is a limited approximation of seafloor texture, there are comprehensive studies specifically to achieve this[14][15].[16] However, it fulfils the purpose of generating variable highlight and shadow independent to the target object. This allows us to investigate the impact of the background on the focus of this study: the proposed semi supervised model.

iii. Speckle and Noise

Real SAS data has multipath speckle due to the coherent nature of the acoustic signal. This interference manifests as cantered Gaussians in the complex and real domain of the returned signal, a Rayleigh distribution therefore describes the absolute value of this signal. Multiplicative noise was therefore added using statistics calculated from real SAS snippet of empty seafloor as shown in Figure 4 (derivation taken from[17]). The second source of noise is system noise this was modelled as additive Gaussian noise the statistics for the Gaussian were measured from real SAS data of a shadow region.

A sinc kernel was convolved over the blurred image to replicate a lobe spreading pattern caused by the coherent acoustic signal.[18] A sinc kernel with 4 lobes was applied to the complex components of the image.

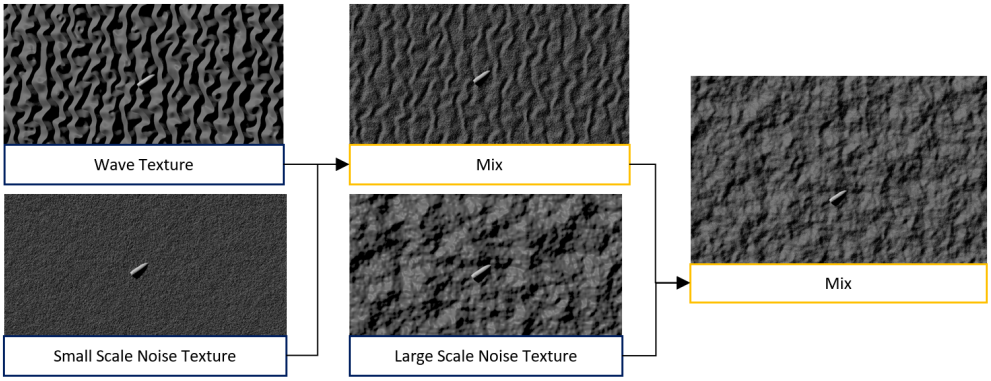

Figure 3: Main elements for procedual sea-floor texture generation.

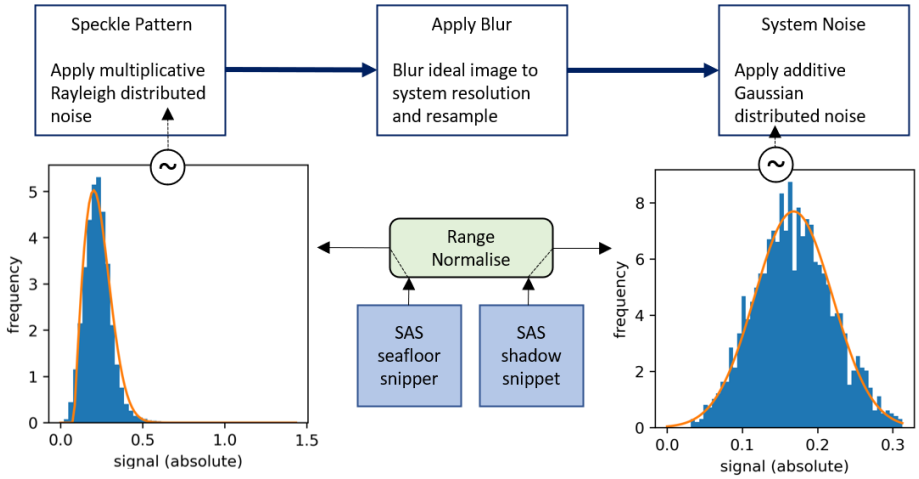

Figure 4: Process for approximating speckle and noise for synthetic data.

Finally the image was resampled at 5cm resolution which is the equivalent resolution of the HISAS sonar system used for UXO surveying at Skagerrak.[2]

B. RE-VIEW MACHINE LEARNING MODEL

This section details the model architecture and training method used to perform the self supervised task of generating novel sonar views (re-viewing). This "re-viewing" model is then re-purposed to perform the desired downstream supervised task of classification. The model is a neural network using both convolutional and fully connected layers implemented using the PyTorch library,[?] see[19] for a broad overview of neural network models.

The re-view model takes as input two views of the seafloor and a desired viewing location. The goal of the model is to output the view of the seafloor from this new viewing location. The model architecture shown in Figure 5 is a feedforward network with information moving from left to right. Note the information passes through a "bottleneck" this is a common feature of self supervised model including autoencoders.[13] Features useful for the self supervised task are extracted from the input in order to fit through this information bottleneck. The downstream classification task is therefore aided by using this pre-trained feature extractor. The motivation for performing the proposed re-viewing task is that it requires extraction of information about the 3d geometry of the scene and the physics of "sonar" imaging in order to generate novel views.

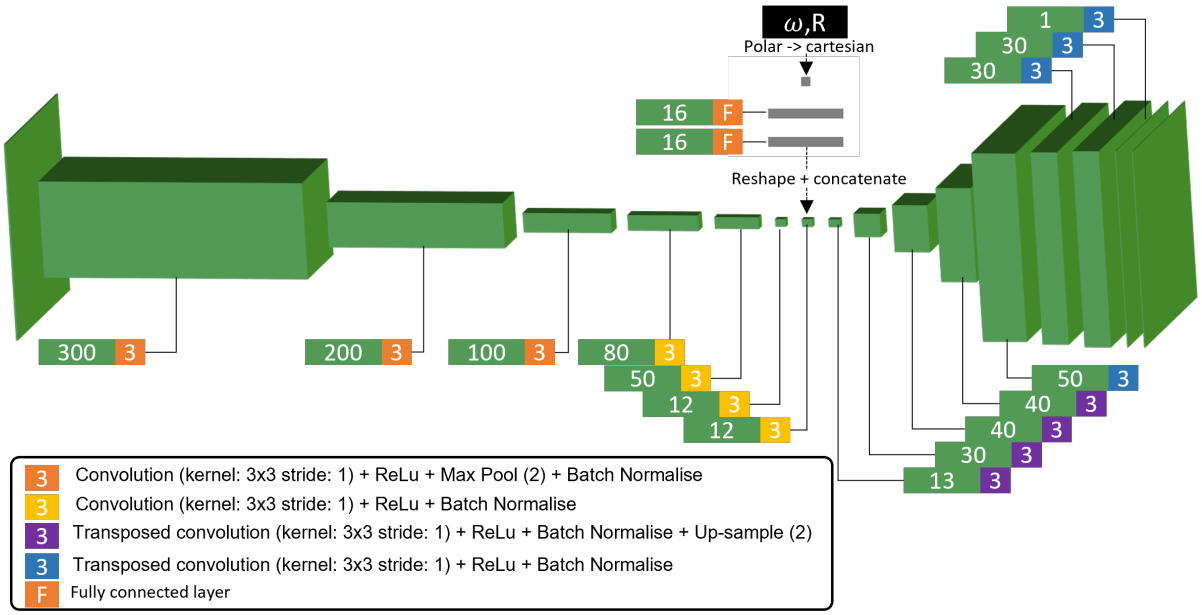

Figure 5: Re-view model architecture - self supervised pre-training

This information is then be helpful for classifying objects contained in the scene.

Model architecture shown in Figure 5 are inspired by similar tasks for re-lighting optical images.[20] Two different sonar views of size 100×100 pixels are input. The model input is a vector of length 20,000 which is compressed by the neural network to a representation vector of length 192. A new viewing position is input and concatenated to the compressed representation. The following section of the model generates the new view from the desired esonifying position. Model parameters are randomly initialised and are then trained using mean square error loss between the true and generated views.

The downstream task of classification is performed by a small network of fully connected layers as shown in Figure 6. This classification network takes the compressed representation as input and outputs a classification "probability" for each class. The difference between classifier output and the true label is calculated by cross entropy loss which is used to train the classifier.

For this study two additional regimes for classification are considered for comparison to the review model, the three variations are:

1. No pre-training
2. Autoencoder pre-trained
3. Re-view pre-trained

All three regimes use the same model architecture to allow for a fair comparison. Model one takes the form of Figure 6 but has its weights randomly initialised before being trained on labelled data. Model two is a pre-trained version of Figure 5 but with zeros input to new sonar position (w and R). Further, a single image duplicated as the input and mean squared error loss calculated between the reconstructed and original images. All models use the same labelled data, and both self-supervised approaches use the same unlabelled data. All models were trained for 50 epochs total, initially the classification network (grey on Figure 6) was trained for 25 epochs with the rest of the network frozen. For practical computer memory considerations a batch size of four was used with the popular Adam gradient descent optimizer.[21]

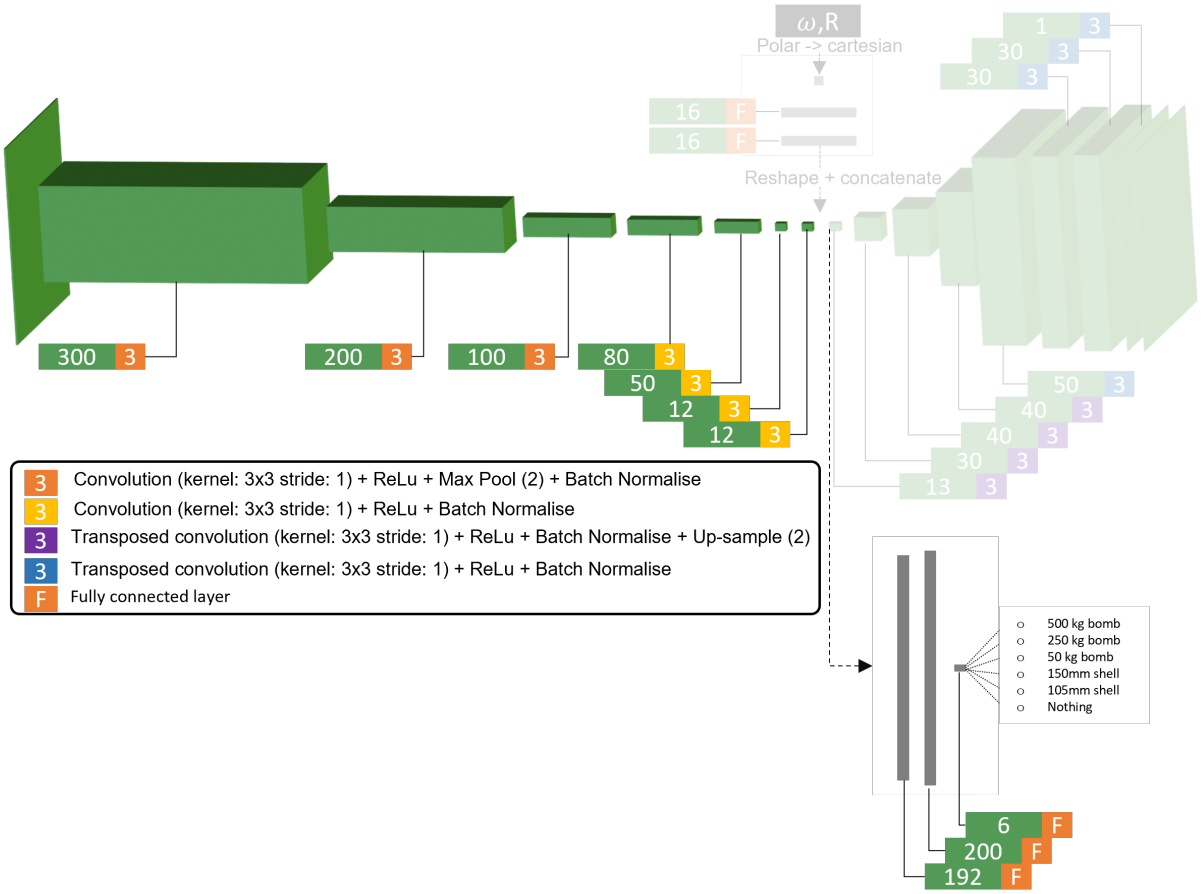

Figure 6: Re-view model architecture - classification

4. RESULTS

We can qualitatively inspect the performance of the self supervised task of re-viewing by generating novel views. We present in Figure 7 the result from two input images of rotating the sonar source around. A large target (250 kg bomb - 1.8m in length) was used with increasingly realistic simulated variations from top to bottom.

The results demonstrate the model generates realistic highlight and shadow progression as the modelled sonar source is moved. The size, shape and orientation of the object is also captured correctly. There is a loss of detailed features, for example the tail fins. Both size and shape reduce in accuracy as more realism is added to the synthetic data. The addition of misalignment between registered images results in significant detriment to the generated re-viewing images.

Another qualitative result is presented in Figure 8. This time the different object classes are shown (top to bottom) All configurations of the synthetic data's realism elements are also shown for comparison (left to right).

From top to bottom the targets get smaller, and predictably poorer realisations of object orientation, realistic highlight and shadow can be observed. Surprisingly texture appears to improve performance especially for the smaller objects: compare column two (speckle only) to column six (speckle and texture). A more informative representation of shadow and highlight may be learnt from the undulations of the seafloor in addition to the target. By way of explanation, with both texture and an object the model has many more

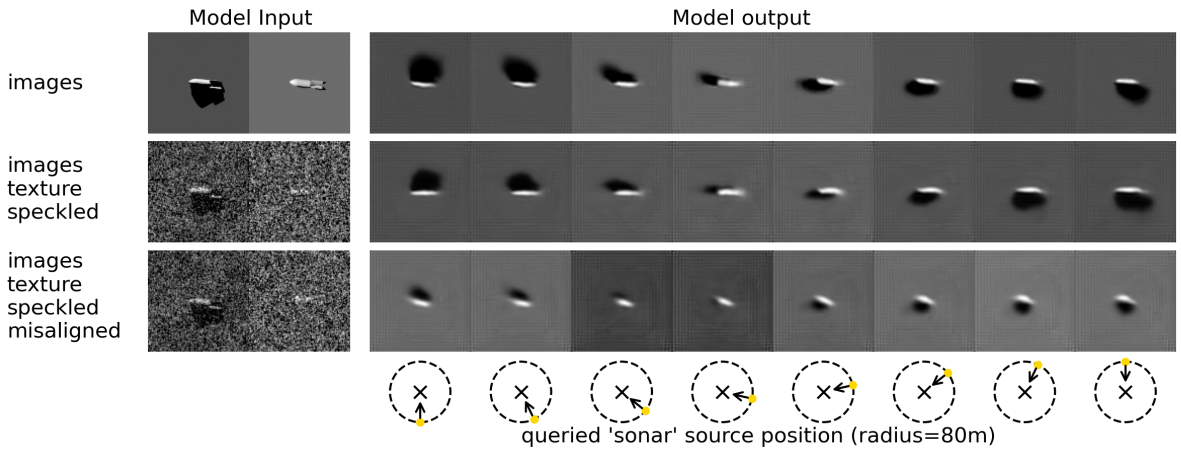

Figure 7: Generated images with a rotating the sonar source.

instances of highlight and shadow per image to learn from compared with a non-textured image. This observation is backed up by following quantitative results (Figure 10).

Following the self-supervised task of re-viewing we have a pre-trained model capable of producing a succinct representation encoding the geometry of the scene. we use this pre-trained model to aid the downstream task of classification, the results of which can be quantified as classification accuracy. We compare re-view pre-training with autoencoder pre-training using the same number of unlabelled images and comparable architectures (see Section 3.2) We also use the same architecture with no pretraining (randomly initialised weights) as a baseline. The classification accuracy on a separate test dataset is shown in Figure 9 for increasing amounts of labelled training data. Labelled training images are sampled randomly from a larger balanced dataset 10 times and the medium and 68th/32nd percentiles plotted. These results use synthetic data with texture and speckled (quantitative comparison with all synthetic data versions is presented later in Figure 10).

Figure 9 shows that both self supervised pre-training methods (orange and blue) outperform the randomly initialised model (green). Further doing the novel re-viewing self supervised task results in better classification accuracy than the traditional self supervised autoencoder. since they use the exact same unlabelled images this demonstrates the re-viewing task forces a better (more informative for classification) representation of the model to be learnt. A levelling off of accuracy at around 80% can be seen even as more labelled examples are added. We observed that small objects can be hard to observe for a human operator given the noise, speckle and texture and the model predictable struggles to classify these objects (not shown). We also observed some convergence of the three lines and we would expect as the number of labelled training tend to infinity the model would converge to a common solution. However, at 700 training examples we still have a clear benefit from autoencoder pre-training and again from re-view pre-training.

Taking a slice from Figure 9 at 100 labelled training images we quantify in Figure 10 the impact of synthetic data realism on classification accuracy. We again split the results to compare the two self-supervised pre-training method and no pretraining.

Looking first at each realism element independently we observed speckle and noise have the largest negative impact on classification accuracy. It is also clear added realism extends the advantage of the re-viewing pre-training method over both autoencoder and no pre-training. This is especially true for adding texture to the sea-floor backing up qualitative observations in Figure 8, the reasons are discussed previously. Classification accuracy is relatively low with the realism elements with speckle and noise being the most significant detractor. However, the new re-view pre-training regime is superior to the other classification methods in all cases. Further, identifying specific object types may be unnecessary for UXO monitoring and

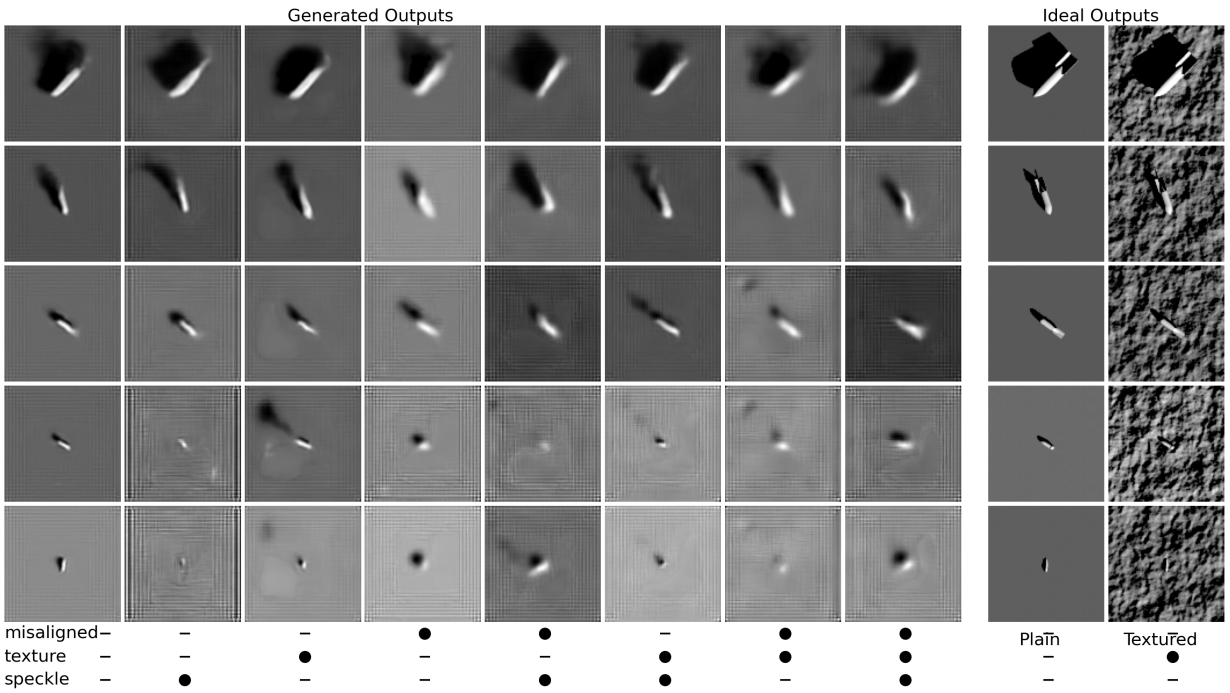

Figure 8: Generated images with a fixed queried sonar source at 80m range. Different object types and levels of realism are used as model input.

remediation and so accuracy may be improved by grouping similarly sized classes. This result also provides a performance benchmark from which the application to real data can be judged.

5. CONCLUSION

In this paper, we proposed and tested a novel self-supervised task which makes use of large unlabelled datasets to pre-train a network for improving classification of UXOs. The self-supervised task required a model to predict the view from a new sonar source (re-view) using simulate synthetic aperture sonar data. Inspecting the new generated views demonstrated the model correctly learnt the scene geometry and physics correctly displaying the progression of highlight and shadow as the sonar source was moved. The ability to inspect the learnt representation by qualitatively assessing re-viewing results allows improved trust in the model. This was supported by improved classification accuracy achieved by using the re-view model compared to an autoencoder and no pre-training. The impact of different levels of synthetic data realism was assessed qualitatively and quantitatively with speckle and noise having the most negative effect. This work demonstrates the potential of re-viewing as a semi supervised approach to classifying UXOs and provides a benchmark to assess an application to real data.

REFERENCES

[1] H Lindsey Arison III. *European disposal operations: the sea disposal of chemical weapons*. CreateSpace Independent Publishing Platform, Scotts Valley, California, US. December 2013.

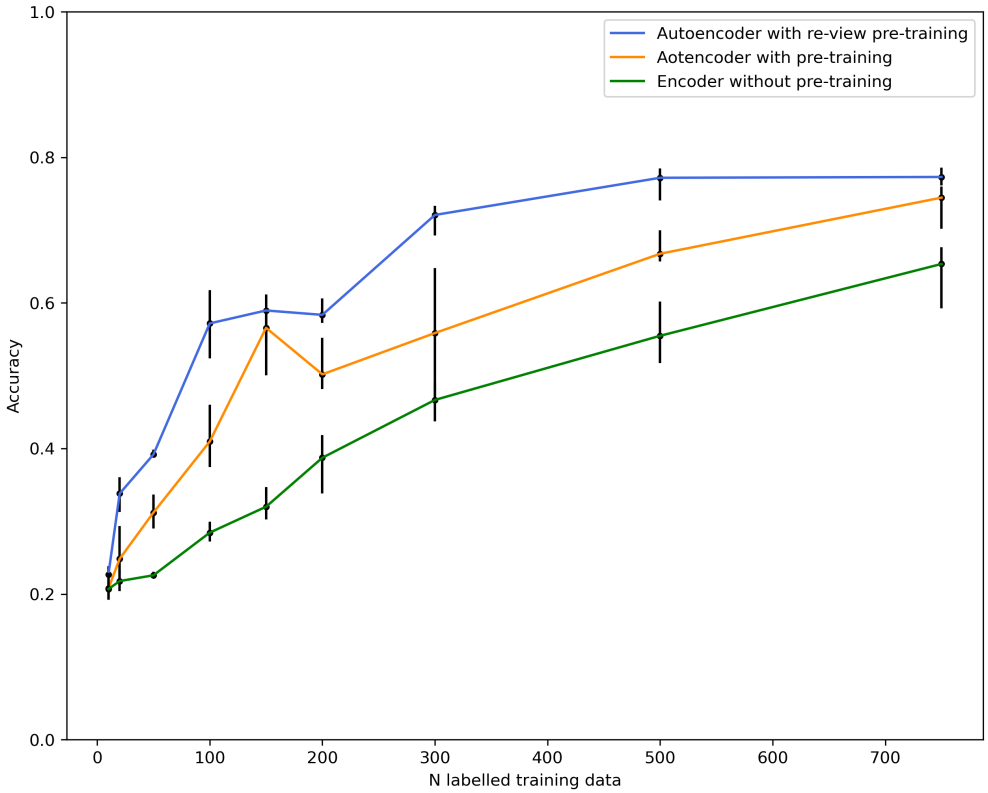

Figure 9: Classification accuracy from a single image with texture, speckle and noise after training on different amounts of labelled images (x-axis) . Blue and orange lines are pre-trained on 13000 unlabelled snippets, and green is not pre-trained.

[2] Roy Edgar Hansen, Hayden John Callow, Torstein Olsmo Sabo, and Stig Asle Vaksvik Synnes. Challenges in seafloor imaging and mapping with synthetic aperture sonar. *IEEE Transactions on geoscience and Remote Sensing*, 49(10):3677–3687, 2011.

[3] Dhiraj Neupane and Jongwon Seok. A review on deep learning-based approaches for automatic sonar target recognition. *Electronics*, 9(11):1972, 2020.

[4] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018, 2018.

[5] T Scott Brandes, Brett Ballard, Sowmya Ramakrishnan, Ethan Lockhart, Bradley Marchand, and Patrick Rabenold. Environmentally adaptive automated recognition of underwater mines with synthetic aperture sonar imagery. *The Journal of the Acoustical Society of America*, 150(2):851–863, 2021.

[6] Pingping Zhu, Jason Isaacs, Bo Fu, and Silvia Ferrari. Deep learning feature extraction for target recognition and classification in underwater sonar images. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 2724–2731. IEEE, 2017.

[7] John McKay, Isaac Gerg, Vishal Monga, and Raghu G Raj. What's mine is yours: Pretrained cnns for limited training sonar atr. In *OCEANS 2017-Anchorage*, pages 1–7. IEEE, 2017.

[8] A Galusha, J Dale, JM Keller, and A Zare. Deep convolutional neural network target classification for underwater synthetic aperture sonar imagery. In *Detection and Sensing of Mines, Explosive Objects, and*

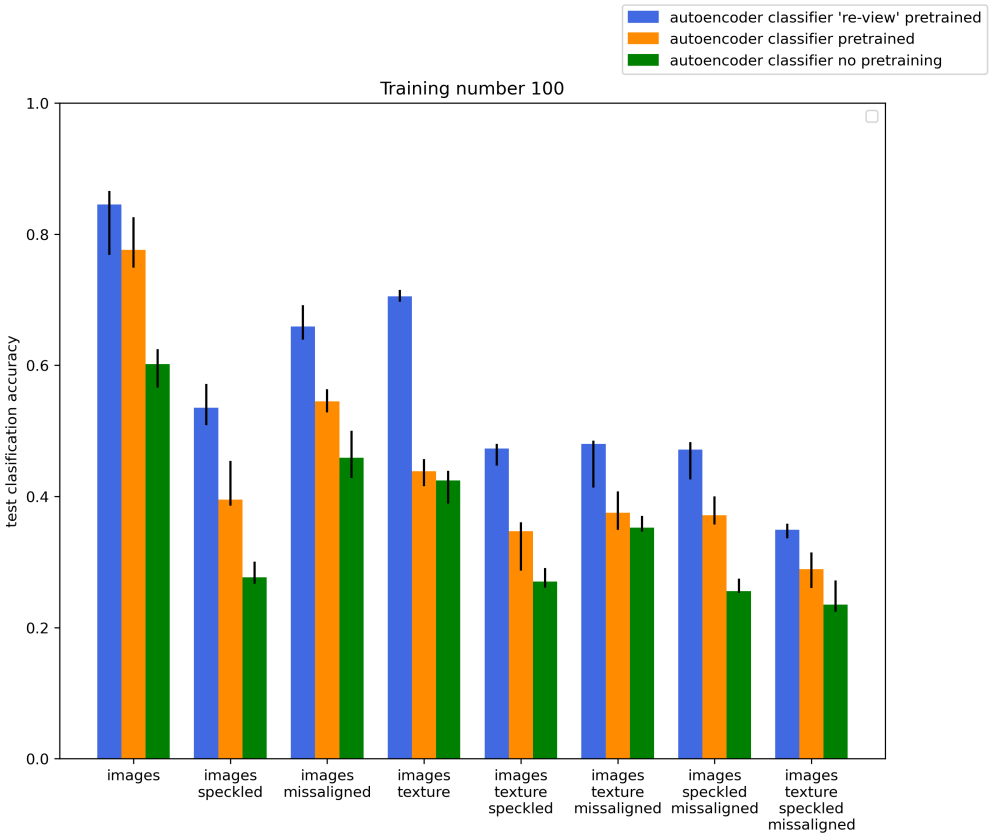

Figure 10: Classification accuracy after training on 100 labelled images. Median, 28th and 68th percentile shown for 10 repeats with randomly sampled training data and randomly initialised model weights.

Obscured Targets XXIV, volume 11012, page 1101205. International Society for Optics and Photonics, 2019.

[9] Johnny L Chen and Jason E Summers. Deep neural networks for learning classification features and generative models from synthetic aperture sonar big data. In *Proceedings of Meetings on Acoustics 172ASA*, volume 29, page 032001. Acoustical Society of America, 2016.

[10] David P Williams. Underwater target classification in synthetic aperture sonar imagery using deep convolutional neural networks. In *2016 23rd international conference on pattern recognition (ICPR)*, pages 2497–2502. IEEE, 2016.

[11] David P Williams. On the use of tiny convolutional neural networks for human-expert-level classification performance in sonar imagery. *IEEE Journal of Oceanic Engineering*, 46(1):236–260, 2020.

[12] Oscar Bryan, Roy Edgar Hansen, Tom SF Haines, Narada Warakagoda, and Alan Hunter. Challenges of labelling unknown seabed munition dumpsites from acoustic and optical surveys: A case study at skagerrak. *Remote Sensing*, 14(11):2619, 2022.

[13] Andrew Ng et al. Sparse autoencoder. *CS294A Lecture notes*, 72(2011):1–19, 2011.

[14] Dajun Tang, Frank S Henyey, Brian T Hefner, and Peter A Traykovski. Simulating realistic-looking sediment ripple fields. *IEEE Journal of Oceanic Engineering*, 34(4):444–450, 2009.

[15] Shawn F Johnson and Daniel C Brown. Sas simulations with procedural texture and the point-based sonar scattering model. In *OCEANS 2018 MTS/IEEE Charleston*, pages 1–7. IEEE, 2018.

[16] Samantha Dugelay and Vincent Myers. A correlated k-distributed model for seabed reverberation in synthetic aperture sonar imagery. *Proc. Inst. Acoust.*, 32(PART 4):163–168, 2010.

[17] MM Siddiqui. Statistical inference for rayleigh distributions. *Journal of Research of the National Bureau of Standards, Sec. D*, 68(9):1007, 1964.

[18] Alan Joseph Hunter. Underwater acoustic modelling for synthetic aperture sonar. PhD Thesis, University of Canterbury, June 2006.

[19] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[20] Zexiang Xu, Kalyan Sunkavalli, Sunil Hadap, and Ravi Ramamoorthi. Deep image-based relighting from optimal sparse samples. *ACM Transactions on Graphics (ToG)*, 37(4):1–13, 2018.

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.