

# Selection of pre-training datasets for sonar image classification

Yannik Steiniger\*, Jose Luis Quinones Gonzalez\*, Dieter Kraus<sup>†</sup> and Benjamin Lehmann<sup>†</sup>

\*German Aerospace Center, Institute for the Protection of Maritime Infrastructures, Germany

<sup>†</sup>City University of Applied Sciences Bremen, Germany

**Abstract**—Deep learning based computer vision models like convolutional neural networks (CNN) are more and more applied for the automatic analysis of sonar images. Since sonar image datasets typically have a limited number of samples, transfer-learning is used to train these models. However, commonly used pre-training datasets, like ImageNet, have a large domain gap to sonar images, i.e., images in these two datasets are fundamentally different. The selection of the pre-training dataset and the related domain gap have shown to have an impact on the final performance of the model. In this work, different datasets are analysed for applying transfer-learning to a CNN for the classification of sidescan sonar images. We quantify the domain gap using a variational autoencoder (VAE) and the t-distributed stochastic neighbor embedding (t-SNE) and link these values to the classification performance of the CNN after fine-tuning.

**Index Terms**—Deep learning, Sidescan sonar, Transfer-learning, Computer vision, Sonar image classification

## I. INTRODUCTION

Sonar image data is typically captured using autonomous underwater vehicles (AUVs) or ships equipped with sidescan or synthetic aperture sonars. These vessels survey the seafloor following a pre-defined path to search for sunken objects of interest. However, collecting data is cumbersome and costly as it requires specialized personnel and equipment. Although significant progress has been made in the field of sonar image classification, especially by developing deep learning models [1]–[3], the outcome of deep learning models remains limited due to the need for more data.

In applications in which the available amount of data is limited, transfer-learning can be applied to improve the overall training of the models [4]. With this concept, a model is first pre-trained on a large source dataset, which does not necessarily represent the final task. In a second step the model is fine-tuned on the real target dataset. For classification, one of the most common pre-training dataset is ImageNet [5], containing over one million optical RGB images and 1,000 object classes. However, compared to sidescan sonar images, the samples from ImageNet are fundamentally different in terms of the imaging sensor, image content and resolution, resulting in a large domain gap between ImageNet and a sonar image dataset (see Table I). Researchers have shown that selecting pre-training datasets with a small domain gap to the target dataset can improve the performance of the model [4].

This paper presents a study on classification datasets, which, compared to ImageNet, are expected to have a smaller domain gap to a sidescan sonar image dataset. For measuring the

domain gap we calculate the Euclidean distance in the latent space of a variational autoencoder (VAE) as well as the Euclidean distance in the t-distributed stochastic neighbor embedding (t-SNE) representation. We train a convolutional neural network (CNN) used in previous work [6] on the most promising datasets as well as ImageNet. The classification performance after fine-tuning on our sonar image dataset is compared to the one of a network trained from scratch.

The remaining of the paper is organized as follows: Section II introduces the datasets which are investigated in this work. Afterwards, Section III covers the measurement of the domain gap. The designed CNN and training parameters are briefly explained in Section IV. In Section V the results of our experiments are presented. Finally, Section VI closes the paper with a summary of the main findings and outlook to future work.

## II. SELECTED DATASETS

Sidescan sonar data was collected over multiple sea trials in the time span from 2019 to 2023 using a Edgetech 2205 sidescan sonar mounted on a SeaCat AUV. The sonar image dataset build from these trials was already described in [6]. It contains objects from the four classes *Tire*, *Rock*, *Cylinder* and *Wreck* as well as an additional *Background* class. The number of samples in the training and test set are given in Table I. Note that the number of test samples is larger than the number of training samples due to the fact that the number of training images from the class *Rock* was limited to keep the training set balanced (see [6] for more details).

The most common dataset to pre-train deep learning models for the classification task is ImageNet. However, ImageNet contains optical RGB images with a higher resolution more details than sonar images and thus the domain gap to sidescan sonar dataset is expected to be large. During pre-training with ImageNet the network learns features based on color which are meaningless for the target task of classifying sonar images. Thus, one criteria for selecting the datasets in this work was that the images should be grayscale to ensure a small domain gap. Typical sensors whose images fulfill this requirement are ultrasonic transducers, synthetic aperture radar (SAR) or X-ray. Another aspect for the selection was the number of training samples, since a dataset used for pre-training should contain more samples than the target dataset. We used Kaggle and Roboflow to search for open source datasets and selected the following ones: Malo [7], Fetal [8], Ship&Boats [9],

SAR [10] and Hand X-ray [11]. Originally, the SAR dataset contains ships to be detected. To use it for a classification task, we extract the ships and additional background snippets at random positions. This results in a binary classification dataset. In addition to these datasets, we manually modified the images from ImageNet to be grayscale. Table I gives an overview about the selected datasets regarding the number of classes and number of samples.





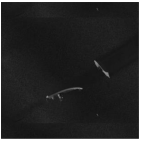




We also setup a simulation using a CAD software where we randomly placed models of ships, tires and rocks in a scene to generate a dataset of synthetic sonar images. Within the generated CAD environment, the essential features and components are: floor, light, and objects. The floor is an extruded planar surface which extends in a rectangular form. It includes different surface elevations along its area and in some other scenarios it also contained a ripple pattern using a simple  $\sin()$  function. For better resemblance with real sonar images, an additional sand texture was set to the planar surface. In a second step, for enhancing the environment to a more realistic scene, two light sources were placed on the 3D assembly. By alternating its location and angle a wide range of possibilities in image diversity was achieved. Finally, the objects of interest, e.g., ships, boats and tires, were downloaded from a free CAD source. Only the rock class was manually created implementing a free-form option from the same CAD software.

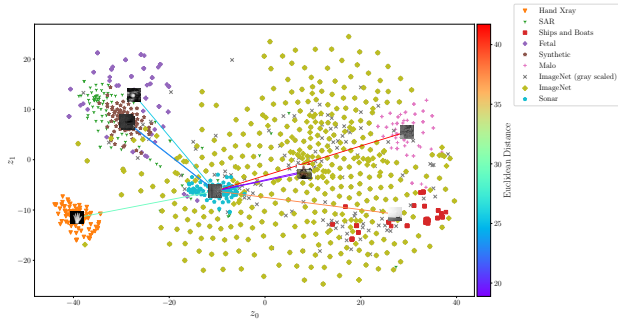
### III. MEASURED DOMAIN GAP

To quantify the domain gap between the pre-training datasets and the sidescan sonar dataset we first map the images to a lower dimensional space. Afterwards, the Euclidean distance between the center points of the datasets in this space is calculated. Mensink et al. propose to use a backbone CNN which was pre-trained on ImageNet to extract features of images from the individual dataset and calculate the distance between these feature vectors [4]. However, since ImageNet is one dataset to be investigated we disregard this approach. In this work, t-SNE as well as a VAE are used for the purpose of dimensionality reduction. The t-SNE method maps the images to a two dimensional space where images that are similar to each other should lie close to each other in the embedding. The encoder of the VAE maps the input images to a latent vector and the decoder learns to reconstruct the image from this. We design the VAE to have a latent vector of size two to obtain a two-dimensional representation of the images.

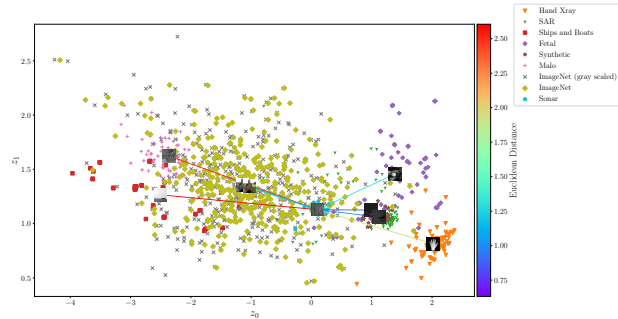
Because the images from ImageNet have three color channels and the input to the network has to be the same for all datasets, we repeat the grayscale images from the other datasets to match this requirement. Figure 1a and 1b show the distribution of the investigated datasets in the t-SNE and VAE embedding space. The calculated distances are also listed in Table II. In the VAE embedding space the SAR dataset has the closest distance to the sonar image dataset. Surprisingly, although sonar images and optical images are fundamentally different the distance for ImageNet measured in the t-SNE the smallest compared to the other datasets. This is partly

Table I: Datasets investigated for pre-training.

Dataset	Classes	Training snippets	Test snippets	Example
Sonar	5	129	1486	
ImageNet	500	252772	75000	
ImageNet (grayscale)	500	252772	75000	
Malo	3	776	98	
Synthetic	3	408	103	
Fetal	6	7129	5271	
Ship&Boats	2	812	49	
SAR	2	62981	20000	
Hand X-ray	6	40637	10160	



(a)



(b)

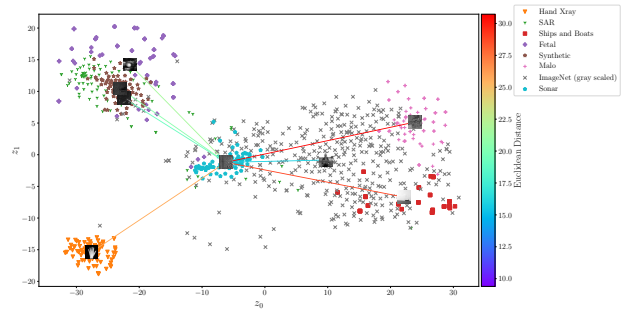
Figure 1: Comparison of the domain gap. (a) Based on t-SNE. (b) Based on VAE.

Table II: Measured distances between to sonar and pre-training dataset.

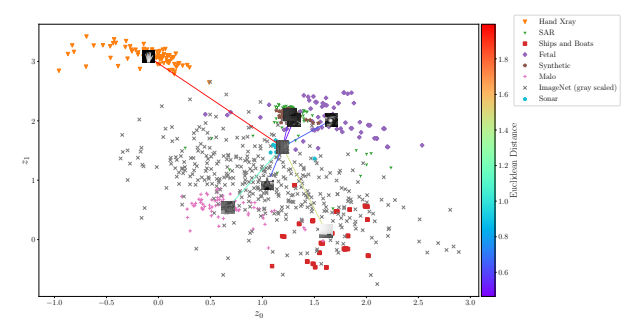
Dataset	Distance using t-SNE	Distance using VAE
ImageNet	18.880	1.150
ImageNet (grayscale)	19.029	1.266
Malo	41.713	2.516
Synthetic	23.321	1.020
Fetal	25.542	1.324
Ship&Boats	37.788	2.605
SAR	22.647	0.888
Hand X-ray	29.334	1.946

due to the large spread of the ImageNet samples in the embedding space. Comparing the distances for the original and the grayscaled version of ImageNet only a slight change is visible with the original one even having a smaller distance in the VAE representation. Intuitively, the grayscaled images are more similar to sonar images and should lead a smaller domain gap. This shows that both methods might not be optimal to measure the domain gap and further research needs to be done to quantify the domain gap between datasets.

Additionally, Figure 2a and 2b show the distance plots for the grayscale images when the input to the network only consists of one channel. Thus, the original ImageNet is excluded. It can be seen that this modification has only minor influence on the distribution in the t-SNE representation. For



(a)



(b)

Figure 2: Comparison of the domain gap with one channel input. (a) Based on t-SNE. (b) Based on VAE.

the VAE the effect is slightly stronger. However, the general distribution of the datasets stays the same, e.g. in both cases (three channel and one channel) the datasets SAR, Synthetic and Fetal are grouped in the same area.

#### IV. NETWORK ARCHITECTURE AND TRAINING

For the experiments we utilize a network architecture described in our previous work, where it trained from scratch to classify sidescan sonar images [6]. This CNN consists of three convolutional layers with 8, 16 and 32 kernel of size  $3 \times 3$ . The output of a convolutional layer is passed through a ReLU activation function, batch normalization and  $2 \times 2$  max pooling. Features from the last convolutional layer are compressed using a fully connected layer with 100 neurons prior to the final output layer. Depending on the pre-training dataset the number of neurons in the output fully connected layer matches the number of classes for each dataset, e.g., the CNN trained on the Malo dataset has three output neurons. Dropout is added before both of the fully connected layers. All input images are scaled to match the input size of the network, which is  $64 \times 64$  pixel.

In the following experiment a CNN is first pre-trained on one of the source datasets and afterwards fine-tuned on the sidescan sonar image dataset. In all cases, pre-training is done for a maximum number of 100 epochs using the Adam optimizer. Early-stopping is used to avoid over-fitting so that the training might be stopped before the last epoch is reached.

Table III: Classification performance for different pre-training datasets. Best value marked in bold.

Dataset	Macro F1-score	
	Source task	Target task
Sonar	0.430	0.429
ImageNet	0.132	0.439
ImageNet (grayscale)	0.113	0.379
Malo	0.701	0.199
Synthetic	0.808	0.306
Fetal	0.469	<b>0.491</b>
Ship&Boats	0.824	0.297
SAR	0.980	0.367
Hand X-ray	0.846	0.302

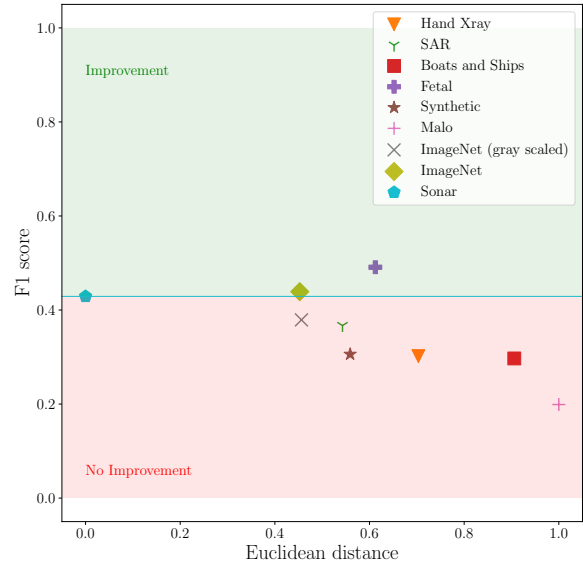
The learning rate in the pre-training step is set to 0.001 for all source datasets. As shown in [12] an optimal performance on the source dataset is not necessary to achieve a good transfer learning result. Thus, we did not focus on optimizing the performance of the CNN after pre-training. For fine-tuning, the output layer is adapted to match the sonar image classification task with five classes, i.e., the output fully connected layer now has five neurons. We noticed that improvement in training required higher patience values in early-stopping. Furthermore, the number of epochs is increased to a maximum of 400. We experimented with different learning rates and achieved the best results with a value of 0.001.

## V. CLASSIFICATION RESULTS

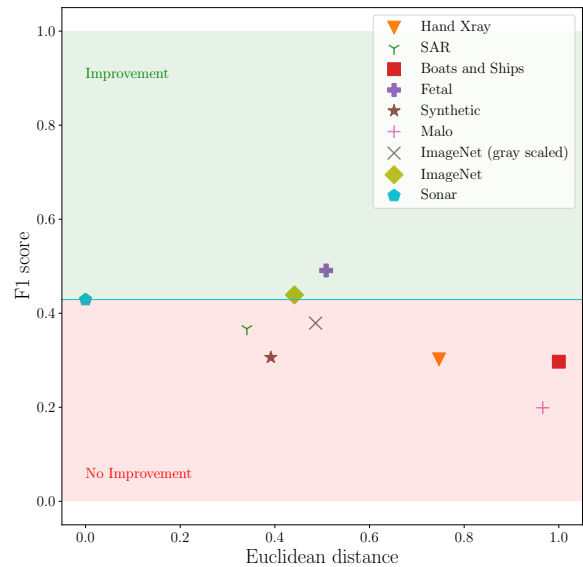
The classification performance of all models is assessed using the macro F1-score to account for the unbalanced test dataset. Table III shows the performance of the CNNs pre-trained on the aforementioned datasets and fine-tuned on the sonar image dataset. In addition, the performance on the source task after pre-training is given. It is worth noting that pre-training on the whole ImageNet dataset was not sufficiently possible due to the limited capacity of the CNN. Thus, for the number of classes was reduced to 500. Our results show that pre-training on the Fetal dataset improves the classification performance compared to pre-training on ImageNet and training from scratch.

To link the classification performance and the measured domain gap, Figure 3a and 3b plots the macro F1-score against the Euclidean distance in the t-SNE and VAE representation, respectively. Both figures show that the performance slightly drops with increasing distance between the source and the sonar dataset. Note however, that the distance not necessarily reflect the intuitive domain gap, since ImageNet shows the smallest distance in the t-SNE representation.

Another important aspect when pre-training a deep learning model is the number of training samples. Comparing the classification performance with the size of the datasets given in Table I it can be seen that Malo and Ship&Boats, which perform worst, contain only a few hundred samples per class. This indicates that a small domain gap itself is not the only requirement for a good transfer-learning result. The source dataset also has to be sufficiently large. It should be noted



(a)



(b)

Figure 3: Domain gap and classification performance. (a) Based on t-SNE. (b) Based on VAE. Note that the distance has been normalized to the maximum value.

that the results hold for the case of a CNN with a one channel input. However in this case the improvement by pre-training on the Fetal dataset becomes smaller.

## VI. CONCLUSION

This work has presented a study on different pre-training datasets for the classification of sonar images. Since the number of sonar snippets is limited, pre-training can be a beneficial way to learn relevant features in a first training step. However, the domain gap to the sonar image dataset should be small. Our analysis shows that distance measured in the t-SNE and VAE embedding space does not match the intuitive domain gap. Additional work needs to be done in order to measure the domain gap more convincingly. Nevertheless, using the Fetal dataset to pre-train a CNN the performance when classifying sonar images after fine-tuning could be improved by more than 6 percentage points compared to the CNN trained from scratch.

The CNN used in this work has three convolutional layers and thus is small compared to other commonly used networks like ResNet or MobileNet. Deeper networks have a higher capacity and the influence of the pre-training might be larger. Extending this study to a broader pool of networks will be left for future work.

## REFERENCES

- [1] S. L. Phung, T. N. A. Nguyen, H. T. Le, P. B. Chapple, C. H. Ritz, A. Bouzerdoum, and L. C. Tran, "Mine-like object sensing in sonar imagery with a compact deep learning architecture for scarce data," in *2019 Digital Image Computing: Techniques and Applications (DICTA 2019)*. Piscataway, NJ: IEEE, 2019, pp. 1–7.
- [2] D. P. Williams, "On the use of tiny convolutional neural networks for human-expert-level classification performance in sonar imagery," *IEEE Journal of Oceanic Engineering*, vol. 46, no. 1, pp. 236–260, 2021.
- [3] Y. Steiniger, J. Stoppe, D. Kraus, and T. Meisen, "Investigating the training of convolutional neural networks with limited sidescan sonar image datasets," in *OCEANS 2022 MTS/IEEE Hampton Roads*. IEEE, 2022, pp. 1–6, in press.
- [4] T. Mensink, J. Uijlings, A. Kuznetsova, M. Gygli, and V. Ferrari, "Factors of influence for transfer learning across diverse appearance domains and task types," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 9298–9314, 2022.
- [5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
- [6] Y. Steiniger, A. Bueno, D. Kraus, and T. Meisen, "Tackling data scarcity in sonar image classification with hybrid scattering neural networks," in *OCEANS 2023 - Limerick*. IEEE, 2023, pp. 1–7.
- [7] asd, "malo dataset," 2023, visited on 2024-06-26. [Online]. Available: <https://universe.roboflow.com/asd-vp0be/malo-vhhkp>
- [8] X. P. Burgos-Artizzu, D. Coronado-Gutierrez, B. Valenzuela-Alcaraz, E. Bonet-Carne, E. Eixarch, F. Crispi, and E. Gratacós, "FETAL\_PLANES\_DB: Common maternal-fetal ultrasound images," 2020. [Online]. Available: <https://doi.org/10.5281/zenodo.3904280>
- [9] C. W. Ng, "ships detection test dataset," 2023, visited on 2024-06-26. [Online]. Available: <https://universe.roboflow.com/chee-wee-ng-efupv/ships-detection-test>
- [10] Y. Wang, C. Wang, H. Zhang, Y. Dong, and S. Wei, "A sar dataset of ship detection for deep learning under complex backgrounds," *Remote Sensing*, vol. 11, no. 7, 2019. [Online]. Available: <https://www.mdpi.com/2072-4292/11/7/765>
- [11] R. Projects, "X ray test dataset," 2023, visited on 2024-06-26. [Online]. Available: <https://universe.roboflow.com/rf-projects/x-ray-test>
- [12] S. Gutstein, B. Lance, and S. Shakkottai, "Does optimal source task performance imply optimal pre-training for a target task?" *CoRR*, vol. abs/2106.11174, 2021.