A New Deep Learning Neural Network Architecture for Seafloor Characterisation

Yoann ARHANT[1], Xavier NEYT
Communications, Information, Systems and Sensors (CISS) Department,
Royal Military Academy, Brussels, BELGIUM

yoann.arhant@mil.be, xavier.neyt@mil.be

Aleksandra PIZURICA
[1](also) Telecommunications and Information Processing (TELIN) Departement,
Ghent University, Ghent, BELGIUM

yoann.arhant@ugent.be, aleksandra.pizurica@ugent.be

ABSTRACT

With the development of Autonomous Underwater Vehicles (AUVs) and high-resolution Synthetic Aperture Sonar (SAS) technology, Automatic Target Recognition (ATR) algorithms achieved detection accuracy above 95% in benign environment and defined new standards for Mine Countermeasures (MCM) at sea. Despite those achievements, their performances are highly dependent on the seafloor type and severely drop in natural environments with sand ripples and rocky terrain. Thus, seafloor characterisation is critical to MCM operations to assess the confidence of those algorithms above certain areas. Additionally, such mapping can be used for mission planning. Extensive research has been devoted to addressing this issue with classical methods. Unsurprisingly, standard deep learning models achieved promising results on such data, but overwhelmingly the scientific literature relied on transfer learning from models without tailoring them neither to the data nor to the task. Therefore, this work proposes a novel deep learning approach for seafloor characterisation based on a scaled down EfficientNet architecture. The complete model is tailored to address the training on limited high-resolution sonar data with data-specific augmentations, the search for the optimal input size and input patch resolution, and efficient post-processing. The resulting model achieves state-of-the-art results on a dataset comprised of challenging seafloors for ATR in real MCM operations.

1.0 INTRODUCTION

The seafloor refers to the underwater surface between the sea and the uppermost part of the oceanic crust. In shallow waters, near the shore, it is mostly composed of unconsolidated sediments that may arrange in patterns like sandwaves forged by local currents. The study of the seafloor is complexified by the absorption of light underwater. Sonar has long been used to overcome that issue and to remotely capture wide seafloor backscatters. With the development of Autonomous Underwater Vehicles (AUVs) and high-resolution Synthetic Aperture Sonar (SAS), Automatic Target Recognition (ATR) algorithms achieved detection accuracy above 95% in benign environment [1], [2] and defined new standards for Mine Countermeasures (MCM) in shallow waters. Despite those achievements, their performances greatly depend on the seafloor type and severely drop in natural environments with vegetation, sand ripples and rocky terrain. Indeed, from the point of view of a low altitude AUV, minimising the ground sample distance of reconstructed Single-Look Complex (SLC) backscatter images, from multiple aggregated pings, to improve the detection and recognition of targets, it also maximises acoustic shadows that might hide an object. Thus, seafloor characterisation is critical to MCM operations to assess the confidence of those algorithms above certain areas. Additionally, such segmentation maps are beneficial to MCM for mission planning or to improve the autonomy of AUVs, and especially if the algorithm could directly be embarked on the vehicle.

1.1 Related Work

Extensive research has been devoted to addressing seafloor characterisation with textural descriptors and regression models or clustering methods. The works of [3] clustered wide areas with statistical curve fitting of averaged side-looking strips of data called swaths. Additionally, studies in [4], [5] addressed binary image segmentation with Gray-Level Co-Occurrence Matrixes (GLCMs), possibly with active contours [4]. Similarly, Brandes and Ballard [6] investigated Multidistribution Dirichlet Clustering with circular Histograms of Oriented Gradients and dictionary learning. Finally, Gips [7] performed patch-based pixel-wise segmentation with gaussian-process classification. Aforementioned methods and others explored numerous algorithms to transform sonar data into maps of varying ground sample distances. Unsurprisingly, standard deep learning models achieved promising results on such data, but overwhelmingly authors relied on networks developed for daily object recognition tasks without tailoring them neither to the task nor to the data [8] – [12].

1.2 Problem Statement

Even though seafloor characterisation and mapping over a single survey appear as a simple task with its limited number of classes, seafloor patterns and dynamics highly vary due to different sensors, geographical location, survey altitude, seafloor topography, strong acoustic backscatterers, sensor calibration, etc. In addition, multiple sources of noise and bias concur in SAS images, such as reconstruction artefacts due to position and orientation errors in the aperture synthesis, speckle noise owing to coherent imaging, pass-scatterers such as organic life or bubbles, etc. All those effects account for the difficulty of the automatic segmentation of the seafloor. Besides, the important costs of seafloor ground thruthing with divers and in-field analyses, hinders the supervised learning process that fall in low-label regime. Standard Deep Learning (DL) methods and in particular typical Convolutional Neural Networks (CNNs) would address such regime with a possible combination of data augmentation [10], [12], AutoEncoder (AE) pretraining of the feature extractor [9], transfer learning [10], [11], [13], unsupervised clustering [11] or self-supervised training such as the geographic location contrastive learning method of [12].

Two main deep learning approaches derived from computer vision have been considered in the literature on seafloor characterisation and mapping: patch classification [10], [12] and pixel-wise segmentation [11], [13]. They automatically craft discriminative features from the distribution of the input data and assign labels respectively to an entire patch of the image or to each pixel. Despite the encouraging results of [10], an important limitation of these existing approaches lies in their incompatibility to simpler tasks and their overfitting with the low-label regime. Recent works [14], [15] alleviate this limitation but still involve a heavy Neural Architecture Search (NAS). In this work, we build on the block and architecture of Tan and Le [16] and scale down the resulting model by limiting the depth and the per sequence of layers of same number of filter maps.

1.3 Paper Contributions

Therefore, this work proposes a novel deep learning approach for seafloor characterisation based on a scaled down EfficientNet architecture derived from computer vision for daily object recognition tasks. The complete model is tailored to address the training on limited high-resolution sonar data with data-specific augmentations. Additionally, this work embodies the search for optimal input patch characteristics and efficient post-processing steps to achieve reliable and fine-scale segmentations with CNN classifiers trained in low-label regime. The proposed CNNs are validated through a small ablation study and the best one achieves state-of-the-art results on such data.

1.4 Paper Organisation

Section 1 introduces seafloor characterisation challenges for ATR and MCM. The proposed approach is presented in Section 2 where we start from the EfficientNet model and build a related model with significantly less parameters and specifically tailored to low-label regime with a particularly designed data augmentation procedure. Section 3 describes the pre-processing pipeline and manual annotation that were employed to create the training and evaluation datasets. Section 4 reports the experimental results including an ablation study and Section 5 concludes the paper.

2.0 METHOD

2.1 Neural Network Architecture

Tan and Le [16] proved the EfficientNet family of models a close to optimal trade-off between computational budget and prediction accuracy on ImageNet [17], outperforming existing models at its publication time. Those blocks were derived from the state-of-the-art for embedded systems MobileNetv2 models [18] by employing its inverted residuals and linear bottleneck block. They also added squeeze and excitation layers after the inverted residuals to accelerate the training by learning a multiplicative bias to be applied over the feature maps and to help select the most discriminative ones for a limited computation, latency, and memory increase. To further accelerate the training with residual blocks, they employed the stochastic depth wrapper for residual blocks which randomly discards the block path and its backpropagation during training, to keep only the identity operation of the residual path.

Henceforward, this work will refer to resolution as the image resolution. The EfficientNet family [16] constitutes a good search space exploration for network engineering meta-architecture parameters such as depth, width, and input resolution instead of relying on computationally expensive search-space exploration algorithms to craft a set of models [17], [18], [19]. Tan and Le used a neural search architecture for the smallest model named *EfficientNet-B0* trained on ImageNet and then solved the discretised optimisation problem of (1) via grid search, with reasonable hypotheses on computation growth constraints, such as (2). Their method is called compound scaling.

$$\begin{aligned} &\alpha\,\beta^2\,\gamma^2 = 2 \\ &s.t.\ \alpha > 1,\ \beta > 1,\ \gamma > 1 \end{aligned} \tag{1}$$

$$\begin{aligned} d &= \alpha^{\phi} \\ w &= \beta^{\phi} \\ r &= \gamma^{\phi} \end{aligned} \tag{2}$$

ϕ is the computational growth parameter, doubling the necessary Floating-point Operations (FLOPs) of the inference for each increase of ϕ by one. α, β and γ are the grid-search parameters for the optimal unit growth from the baseline model with compound scaling. Finally, d, w, and r are respectively the depth, width, and resolution multipliers to be applied on the architecture. In particular, the depth multiplier is applied on the number of block repetitions between two decimation operations and is directly linked to the depth of model.

First, we investigate the use of (1) and (2) but for a smaller input size r – which is called input resolution for ImageNet as the images are interpolated to a smaller input resolution. On the contrary, our mapping approach tiles SAS images with patches of maximum resolution. As we aim for similar global average pooling height and width reduction than the baseline model, we reduce the number of decimation operations – in the baseline model there are five strides of two, which decimates by a factor two the resolution of the

feature maps. Our approach is to perform an average pooling over at least a grid of four-by-four feature maps and to increase the number of decimations by one just after reaching the next power of two. Therefore, the layer floors depth l is:

$$l = \left\lceil \log_2 \left(\left\lceil \frac{r}{8} \right\rceil \right) \right\rceil \tag{3}$$

Additionally, and contrary to the actual meaning of input resolution, for which images are interpolated, reducing the size altogether with the complexity of patterns, the use of maximum input resolution does not reduce the complexity of patterns. Hence, Networks still requires to be deep enough to be discriminative to them. Then, we study the effect of resolution by interpolating input images, reducing their size by a factor 2 or 4. Finally, we reduce the growth of the number of feature maps after a decimation operation reverting to the design of [18], which is illustrated in Figure 1. As a result, this work considers multiple architectures to understand how models can be robust to overfitting with such limited data, while addressing the mapping of the seafloor from noisy and real-world data.

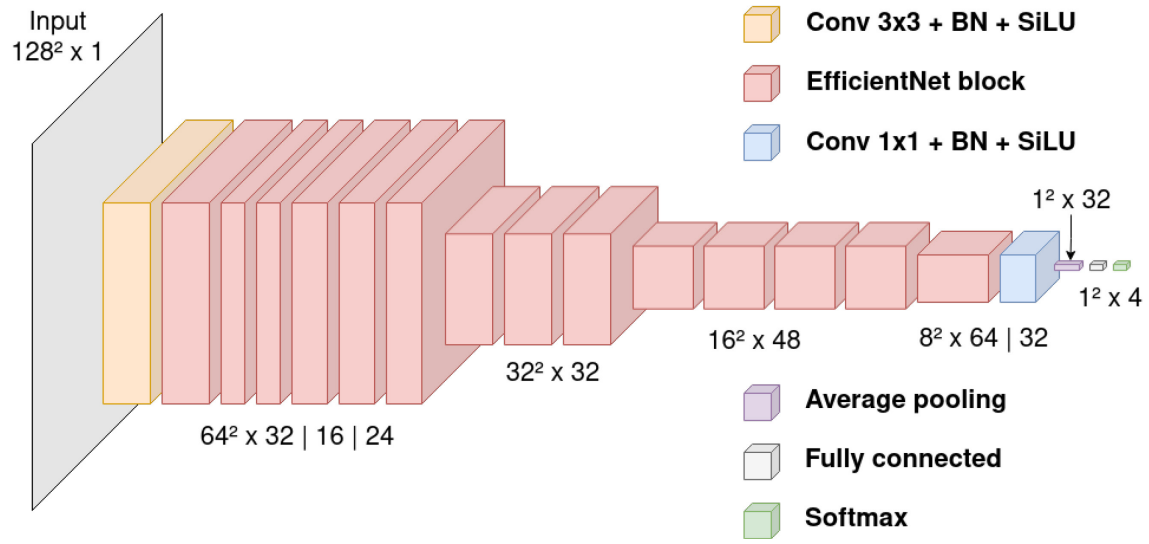

Figure 1: Diagram illustrating the architecture of our proposed and scaled down EfficientNet classifier represented by its feature maps output size after every layer and operation. We reverted to the filter maps growth of [18] for which the model's number of filter maps is doubling every two decimation operations.

2.2 Data Augmentation

The data augmentation is beneficial to the training of CNN models within the low-label regime by reducing even more overfitting and improving the generalisation ability of the model. The pipeline is derived from efficient standard computer vision operations and is comprised of the sequential and random application of flips, rotation, scaling, translation, and contrast and intensity jittering. Thus, it can compensate biases introduced by some input distributions such as the partial representation of the possible orientations of sand ripples.

We also employ a modified sliding window strategy tiling the entire sonar image to consider only input information from the natural distribution fed to models. Given an input patch size and a stride, each sonar image is decomposed as a set of possibly overlapping windows identified by pixel coordinates. Then, depending on its position whether being too close to the edges to be augmented without the introduction of padded pixels or not, the pipeline keeps the unaugmented patch or randomly applies the transforms.

Contrary to standard patch-based classification approaches derived from computer vision that would first cut the image in windows, assigning a label to each of them and then performing transforms with the mirroring or padding of the edges, our method works on pixel coordinates and augments both the backscatter and semantic maps to generate a patch and its label by pixel majority voting over the semantic map. Such a design implies that different random augmentations of the same pixel coordinate may result in different class tags. Besides, as acoustic shadows are only cast in increasing range, the random rotation is limited to 15° to avoid introducing non-plausible data in the input distribution. The data augmentation pipeline is illustrated in Figure 2 and an ablation study will further examine the performance improvement of models. In the end, the images are fed to the data augmentation stage and then to the DL classifiers for training contrary to the evaluation where images remain untransformed to ensure the test stage is representative of the model's ability to create segmentation maps from its predictions.

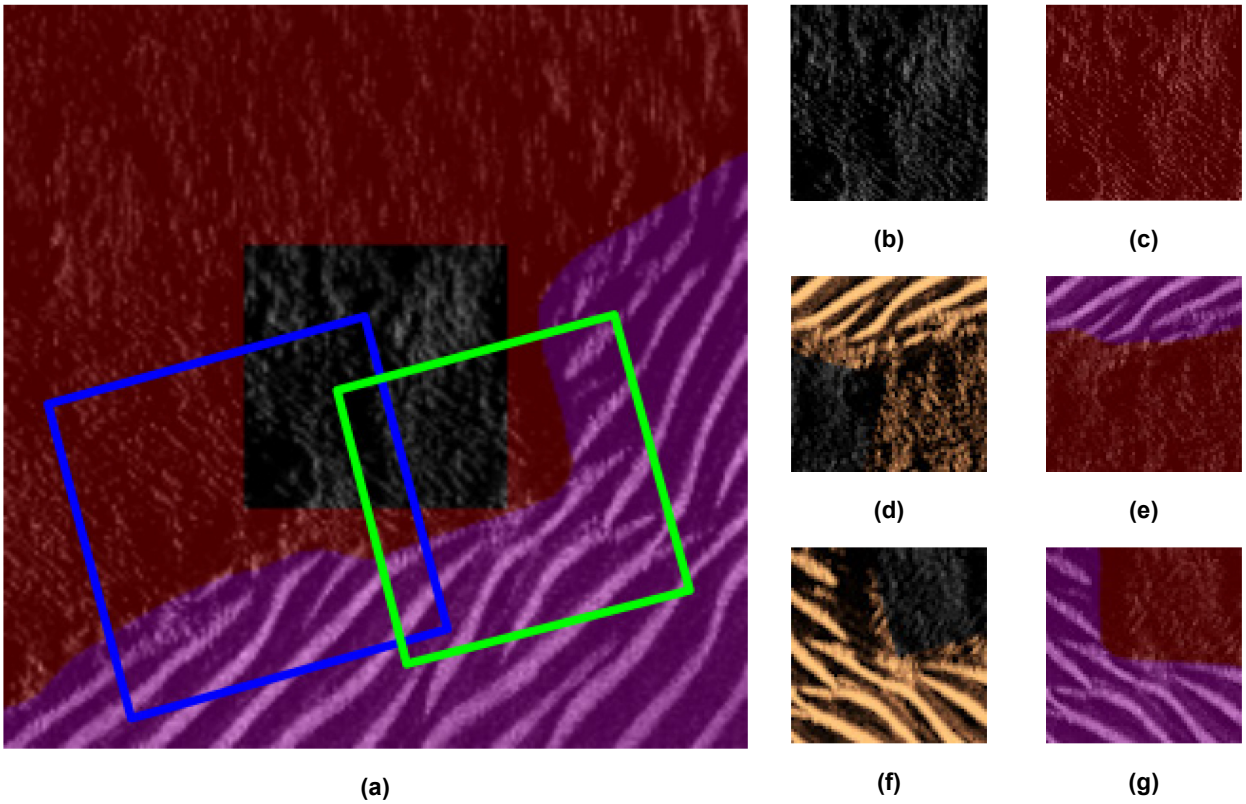

Figure 2: A pixel coordinate boundary between small sand ripples in red and large sand ripples in magenta and its corresponding ground truth (a) displayed through different views of data augmentation (b), (d), (f) for which the semantic maps are also augmented (c), (e), (g). The centre crop corresponding to the unaugmented patch (b) is displayed in grey colour map also in (a), (d), (f) while our augmentation pipeline, keeping filled edges with natural input information, is embodied by the copper colour map. (d) – (e) and (f) – (g) represent two possible outcomes of randomly augmented patches with flips, rotation, scaling, and translation for which the pixel majority voting assigning the class tag would result in small sand ripples for (e) and large sand ripples (g). (d) and (f) are represented back on the bigger ground truth map (a) respectively by a blue and green square. Such a design reduces the class tag bias otherwise introduced by fixed labels and improves models' generalisation ability.

2.3 From Patch Coordinates to Seafloor Characterisation Maps

To evaluate such an ability, the easiest strategy would be to generate non overlapping patches tiling all images and to accumulate the predictions into maps. However, big patches and the stride of the sliding window prevent those maps of the seafloor to accurately represent the fine scale boundaries contrary to pixel-wise segmentation. To circumvent this issue, an overlapping patches strategy is used to reduce de discretisation, but for which a post-processing fusion strategy is required. This work studies either the centre crop scheme discarding the external borders of overlapping predictions – for which the extreme case with a stride of one pixel corresponds to patch-based pixel prediction as in [5], [7] – or a complicated majority voting process reverting to the former procedure in case of a tie. With such a design, the patch-based classification approach produces finer-scale maps and becomes more comparable to the pixel-wise segmentation one.

3.0 DESCRIPTION OF THE DATASET

The data comes from a MCM exercises campaign at sea with the MUSCLE AUV equipped with a SAS sensor, where targets and confusers were laid on the seafloor. It is comprised of a variety of seafloor patterns ranging from flat sandy bottoms to rock outcrops and through sand ripples. The sizes of the reconstructed SLC images are 7333 pixels acrosstrack and 2000 pixels alongtrack.

3.1 Data Pre-Processing

First, those complex images are reduced to their modulus. Then, they are curated from inconsistent values, median-normalised, clipped and log-converted following the procedure described in [22] to standardize and compress the high acoustic dynamic of the seafloor. Finally, the resulting images are bilinearly interpolated to reach the same alongtrack and acrosstrack resolution, in other words to reach a square ground pixel. That pre-processing pipeline was found effective for seafloor characterisation with CNN models in [7], [11].

3.2 Manual Annotations

The seafloor being homogeneous within vast areas and hardly inhomogeneous for MCM related seafloor characterisation, the polygon annotation strategy was adopted to carry out the manual post-survey ground thruthing. Seventeen images, one from each acquisition segment, were randomly selected to establish a supervised training set. The polygons are characterised as Flat Bottom, Rocks, and Small and Large Sand Ripples. They represent the generic seafloor type classification scheme for ATR in MCM operations [23] to identify difficult terrain where shadows might be cast over targets. Seeking for swiftness for early experiments, annotators preferred to avoid diving too deep into details of the image and drop pieces of seafloor characterisation ground truth smaller than a target, such as small boulders isolated in a sandy area. Likewise, the hard annotation strategy enforces annotators to bias one class over the others when dealing with mixed composition areas. Contrary to the weak annotations of [11] and [13], for which an unknown class is assigned where no labels were given by annotators, this work endorses that the small label bias would not be detrimental to resulting models. According to the ability of CNN architecture to be robust to imperfect annotations, this work assumes instead the added amount of labelled data to be beneficial to the training. Ultimately, the most challenging image is put aside to constitutes the evaluation set. Unseen seafloor configurations and boundaries, for which the ground truth, which is displayed in Figure 3(a), is characterised by a hundred of polygons, will account for the models' generalisation ability. Since the other images are spanning all textures of interest, they are sufficient to address the training of seafloor characterisation CNNs with a representative distribution of the seafloor and represents the train dataset which is like the one of [9], but with fewer images.

4.0 EXPERIMENTS AND RESULTS

4.1 Training Details

The training of classifiers is performed within a Distributed Data Parallel (DDP) strategy over 4 GPUs with batch size of at 256 and patches of size ranging from 64 to 256. The train dataset is split at 90% into the training set and the remaining 10% into the validation set. To mitigate the effect of imbalance in the training distribution, the cross-entropy loss is weighted by coefficients inversely proportional to the number of pixel occurrences of the class in the ground truth. Additionally, to ensure the best version of models are reached, an early stopping strategy with patience is employed to keep training models until no improvement over the validation accuracy can be noticed. Both designs enforce the training procedure to be more efficient and results to be more replicable. The starting learning rate is initiated at 0.001 and is employed within an Adam Optimiser strategy coupled with a scheduler reducing all individual learning rates for each layer by a small factor every 100 epochs, to avoid having to tailor them manually. Additionally, a small weight decay is introduced as a regulariser to prevent the explosion of weights. Trainings, which ranged from a couple of minutes to few hours depending on model size, are fast enough to perform an ablation study.

4.2 Ablation Study

We conduct ablation experiments to explore some meta-architecture hyperparameters for EfficientNet models derived from the procedure elaborated in Section 2.1 and study their effects in term of accuracies and computation growth, which are reported in Table 1. Early experiments with the depth and width scaled down architectures, owing to lower input size, performs poorly. Indeed, the reduced number of filters prevents models from learning discriminative features for all possible orientations and scaling factors given by data augmentation. Consequently, instead of drastically reduce the number of filters at all stages we revert to a filter maps growth similar to the one of [18]. Alongside and as models were addressing patches of similar input size, we adopt the largest depth multiplier to slightly compensate the significant decrease of training parameters. Experiments lead to a scaled down architecture, which is compared to the standard *EfficientNet-B0* model in Section 4.4 and a pixel-wise segmentation model from our previous work in Section 4.5. The models are also compared on the number of training parameters and the Floating-point Operations (FLOPs) necessary to perform an inference, which account for the simplicity of the model and its computing requirements. Equally important, results highlight the discretisation effect over the labelled ground truth. The larger input size, the better models perform patch-based classification, which is a simpler task with increasing input size, to the detriment of the pixel-wise accuracy, which is a more complex task.

Table 1: Ablation study of our scaled down EfficientNet family with input size, input resolution, model architecture and reduced number of feature maps in terms of patch-based and pixel-wise accuracy (ACC).

Input patch size	Down-sampling factor	Reduced feature maps	*w*	*d*	Patch-based ACC	Pixel-wise ACC	Number of Training parameters	Inference GFLOPs	Map creation GFLOPs
224	1	No	1.0	1.0	92.8 %	83.6 %	3.6 M	2.09	1386
224	1	Yes	1.0	1.0	90.6 %	84.4 %	1.1 M	1.12	743
128	1	No	0.68	0.48	88.6 %	84.2 %	148 k	0.11	232
64	1	No	0.43	0.20	83.2 %	82.8 %	8 k	0.01	85

Input patch size	Down-sampling factor	Reduced feature maps	*w*	*d*	Patch-based ACC	Pixel-wise ACC	Number of Training parameters	Inference GFLOPs	Map creation GFLOPs
256	1	No	1.0	1.0	96.5 %	81.3 %	3.6 M	2.73	1529
256	1	Yes	1.0	1.0	**97.2 %**	83.9 %	1.1 M	1.46	818
256	1	No	1.0	1.1	96.5 %	81.3 %	4.6 M	3.76	2106
256	1	Yes	1.0	1.1	86.5 %	84.0 %	1.4 M	2.09	1170
128	2	No	1.0	1.1	95.1 %	83.4 %	580 k	0.56	1180
128	2	Yes	1.0	1.1	93.1 %	81.7 %	276 k	0.41	209
64	2	Yes	1.0	1.1	91.4 %	84.3 %	46 k	0.06	126
64	4	Yes	1.0	1.1	95.8 %	83.3 %	46 k	0.06	**31**
128	1	No	1.0	1.1	85.6 %	85.6 %	580 k	0.56	1180
128	1	Yes	1.0	1.1	91.8 %	**86.4 %**	276 k	0.41	864
64	1	Yes	1.0	1.1	85.9 %	85.5 %	**46 k**	0.06	510

4.3 Prediction Mapping Analysis

Even though resulting maps are imperfect, they are promising enough from imperfect annotation, in low-label regime, and over unseen seafloor configurations Figure 3(b) and Figure 3(d). Indeed, models created with different input sizes could capture boundaries of different sizes. In addition, models could recover from annotation bias of rocks over the other classes. Besides, small boulders in between sand ripples are also recovered even though they were not labelled in the Train set. Finally, inconsistencies between the imperfect ground truth and produced maps with the simplistic post-processing design of Section 2.3, are often visually impossible to share to one another. Therefore, annotations will have to be refined to conclude over comparable performances of models to generate seafloor characterisation maps.

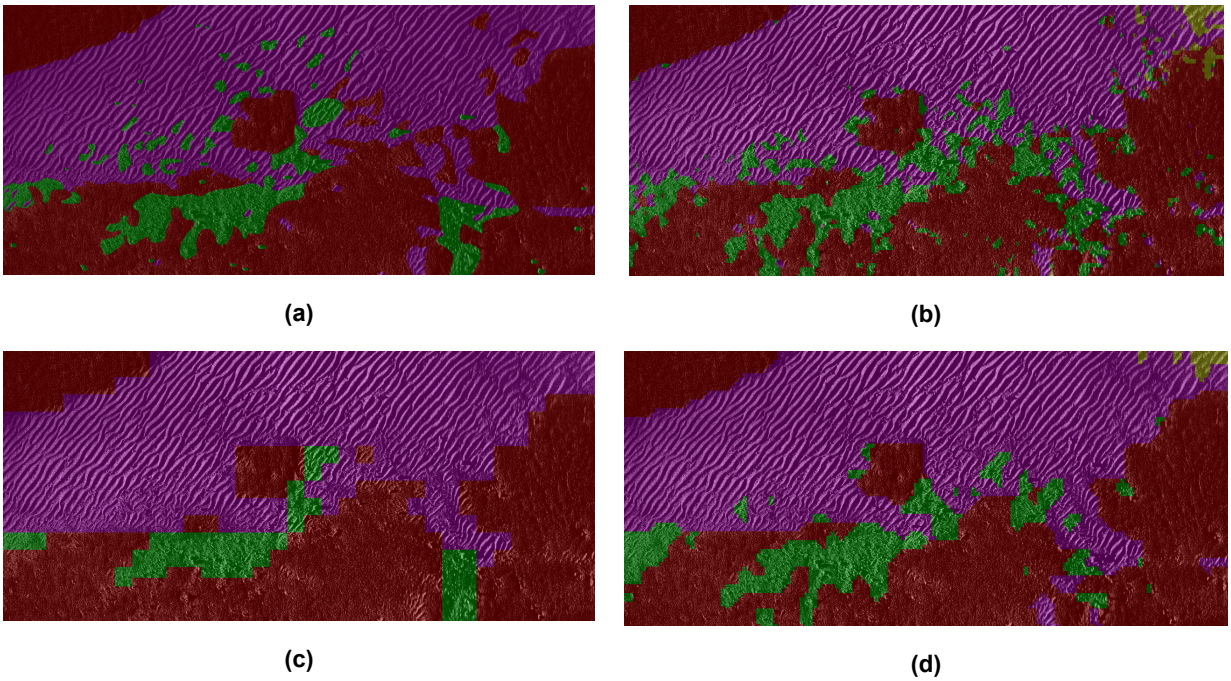

Figure 3: The semantic (a) and 256-patch-based classification (c) ground truth maps compared to the mappings respectively produced by our 64 (b) and 128 (d) pixels input size scaled down EfficientNet (b) with the half overlap tiling strategy. The Green, Yellow, Red, Magenta colours respectively corresponds to the Rocks, Flat Bottom, Small and Large Sand Ripples.

4.4 Comparison with Transfer Learning and the Baseline *EfficientNet-B0* Model

Seafloor characterisation datasets and algorithms are hardly shared. Therefore, comparing early developments of methods could be troublesome. In this work, we choose to compare to the *EfficientNet-B0* model from [10], pretrained on ImageNet and its own data augmentation schemes. To that extent, we considered different state-of-the-art transfer learning retraining strategies ranging from complete retraining to frozen layers and through model part specific learning rate. The best results were obtained with the latter. Additionally, models need to be modified to be consistent with the new task and number of classes. The fewer layers are replaced to maximise the number of transferred weights from the feature extractor. Hence, only the first convolution, now addressing a single channel instead of colour, and the final mapping convolution to a fixed number of channels before the global average pooling and the classifier part, are discarded and reinitialised. Table 2 proves our method right and also reveals the inability of the patch-based accuracy to correctly embody the performances of models due to different discretisation sizes.

Table 2: Comparison of the *EfficientNet-B0* model applied to seafloor characterisation with different configurations of data augmentation and transfer learning.

Model	Our augmentations	Transfer Learning	Patch-based classification accuracy	Pixel-wise segmentation accuracy
EfficientNet-B0 [10]	No	Yes	85.6 %	79.4 %
EfficientNet-B0	Yes	Yes	90.6 %	82.2 %
EfficientNet-B0	Yes	No	92.8 %	83.6 %
Our scaled down *EfficientNet-B0*	Yes	No	90.6 %	**84.4 %**

4.5 Comparison with Pixel-Wise Semantic Segmentation

The best resulting model achieves state-of-the-art results on a dataset comprised of challenging seafloors of real-world MCM surveys. However, and despite models being hardly comparable, our scaled down model is slightly outperformed by our previous work on pixel-wise segmentation, as shown by Table 3. We evaluate classifiers with a half patch overlap and a majority voting strategy. It was experimentally found an acceptable trade-off between prediction accuracy and computation increase, roughly quadrupling the number of patches to be processed for the same image, since it reduces the mapping discretisation comparable or below the size of MCM targets of interest.

Table 3: Comparison of the pixel-wise segmentation model from our previous work with classifiers with the half overlap patch majority voting fusion strategy.

Model	Inference GFLOPs	Training Parameters	Pixel-wise classification accuracy	Complete Image mapping GFLOPs
EfficientNet-B0 [10]	2.15	3.6 M	79.4 %	1386
D4SC (Previous work)	4.34	960 k	**87.3 %**	**625**
Ours	0.41	276 k	86.4 %	864

5.0 CONCLUSION

This work investigated the effect of network engineering for a faster and more efficient patch-based classification mapping, to reach less memory requirements and necessary computing power, and for the deployment of seafloor characterisation algorithms for MCM directly inside AUVs. It failed at outperforming the pixel-wise semantic segmentation model over unseen seafloor configurations. Nevertheless, this work embodies an efficient exploration of network engineering meta-architecture search space of feature extractors that will be beneficial to the improvement of all segmentation models and the exploration of bigger ones addressing multiple geographic locations, sensors, etc. Additionally, comparable accuracies highlight the limits of the low amount of annotated and imperfect data. Thus, further research should extend the dataset to make segmentation models more robust and more capable of handling difficult sonar and seafloor configurations. Future work should also explore models' ability to inform about their being out-of-bounds or to adapt themselves to the new distribution of data. Careful designs presented in this work, might also be beneficial to other in low-label regime applications such as in medical images.

6.0 REFERENCE SECTION

[1] D. P. Williams, 'Fast Target Detection in Synthetic Aperture Sonar Imagery: A New Algorithm and Large-Scale Performance Analysis', *IEEE Journal of Oceanic Engineering*, vol. 40, no. 1, pp. 71–92, 2015, doi: 10.1109/JOE.2013.2294532.

[2] I. D. Gerg and V. Monga, 'Structural Prior Driven Regularized Deep Learning for Sonar Image Classification', *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, 2022, doi: 10.1109/TGRS.2020.3045649.

[3] L. J. Hamilton and I. Parnum, 'Acoustic seafloor segmentation from direct statistical clustering of entire multibeam sonar backscatter curves', *Continental Shelf Research*, vol. 31, no. 2, pp. 138–148, 2011, doi: https://doi.org/10.1016/j.csr.2010.12.002.

[4] M. Lianantonakis and Y. R. Pétillot, 'Sidescan Sonar Segmentation Using Texture Descriptors and Active Contours', *IEEE Journal of Oceanic Engineering*, vol. 32, pp. 744–752, 2007, doi: 10.1109/JOE.2007.893683.

[5] T. Celik and T. Tjahjadi, 'A Novel Method for Sidescan Sonar Image Segmentation', *IEEE Journal of Oceanic Engineering*, vol. 36, no. 2, pp. 186–194, 2011, doi: 10.1109/JOE.2011.2107250.

[6] T. S. Brandes and B. Ballard, 'Adaptive Seafloor Characterization With Hierarchical Bayesian Modeling of SAS Imagery', *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 3, pp. 1278–1290, 2019, doi: 10.1109/TGRS.2018.2865606.

[7] B. Gips, 'Texture-Based Seafloor Characterization Using Gaussian Process Classification', *IEEE Journal of Oceanic Engineering*, pp. 1–11, 2022, doi: 10.1109/JOE.2022.3162023.

[8] J. Chen and J. Summers, 'Deep neural networks for learning classification features and generative models from synthetic aperture sonar big data', *Journal of the Acoustical Society of America*, vol. 140, p. 3423, Oct. 2016, doi: 10.1121/1.4971014.

[9] J. Chen and J. E. Summers, 'Deep convolutional neural networks for semi-supervised learning from synthetic aperture sonar (SAS) images', in *Proceedings of 30th Meeting on Acoustics*, 2017. doi: 10.1121/2.0001018.

[10] G. Chandrashekar, A. Raaza, V. Rajendran, and D. Ravikumar, 'Side scan sonar image augmentation for sediment classification using deep learning based transfer learning approach', *Materials Today: Proceedings*, 2021, doi: https://doi.org/10.1016/j.matpr.2021.07.222.

[11] Y.-C. Sun, I. D. Gerg, and V. Monga, 'Iterative, Deep Synthetic Aperture Sonar Image Segmentation', *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, Sep. 2022, doi: 10.1109/TGRS.2022.3162420.

[12] T. Yamada, A. Prügel-Bennett, S. B. Williams, O. Pizarro, and B. Thornton, 'GeoCLR: Georeference Contrastive Learning for Efficient Seafloor Image Interpretation', *FR*, vol. 2, no. 1, pp. 1134–1155, Mar. 2022, doi: 10.55417/fr.2022037.

[13] I. D. Gerg and V. Monga, 'Deep Multi-Look Sequence Processing for Synthetic Aperture Sonar Image Segmentation', *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023, doi: 10.1109/TGRS.2023.3234229.

[14] M. Tan and Q. Le, 'EfficientNetV2: Smaller Models and Faster Training', in *Proceedings of the 38th International Conference on Machine Learning*, Jul. 2021, pp. 10096–10106. Accessed: Mar. 06, 2023. [Online]. Available: https://proceedings.mlr.press/v139/tan21a.html

[15] C.-C. Wang, C.-T. Chiu, and J.-Y. Chang, 'EfficientNet-eLite: Extremely Lightweight and Efficient CNN Models for Edge Devices by Network Candidate Search', *J Sign Process Syst*, Sep. 2022, doi: 10.1007/s11265-022-01808-w.

[16] M. Tan and Q. V. Le, 'EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks', *unpublished*, May 2019, [Online]. Available: http://arxiv.org/abs/1905.11946

[17] O. Russakovsky *et al.*, 'ImageNet Large Scale Visual Recognition Challenge', *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015, doi: 10.1007/s11263-015-0816-y.

[18] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, 'MobileNetV2: Inverted Residuals and Linear Bottlenecks', in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 4510–4520. doi: 10.1109/CVPR.2018.00474.

[19] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, 'Learning Transferable Architectures for Scalable Image Recognition', in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, Jun. 2018, pp. 8697–8710. doi: 10.1109/CVPR.2018.00907.

[20] M. Tan *et al.*, 'MnasNet: Platform-Aware Neural Architecture Search for Mobile', in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 2815–2823. doi: 10.1109/CVPR.2019.00293.

[21] A. Howard *et al.*, 'Searching for mobileNetV3', *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2019-October, pp. 1314–1324, Oct. 2019, doi: 10.1109/ICCV.2019.00140.

[22] D. P. Williams, 'The Mondrian Detection Algorithm for Sonar Imagery', *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 1091–1102, 2018, doi: 10.1109/TGRS.2017.2758808.

[23] D. P. Williams, 'Fast Unsupervised Seafloor Characterization in Sonar Imagery Using Lacunarity', *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 11, pp. 6022–6034, 2015, doi: 10.1109/TGRS.2015.2431322.