D4SC: Deep Supervised Semantic Segmentation for Seabed Characterization and Uncertainty Estimation for Large Scale Mapping

Yoann Arhant, *Graduate Student Member, IEEE*, Olga Lopera Tellez, Xavier Neyt, and Aleksandra Pižurica, *Senior Member, IEEE*

***Abstract*—Seabed characterization consists in the study of the physical and biological properties of the of ocean floors. Sonar is commonly employed to capture the acoustic backscatter reflected from the seabed. It has been extensively used for automatic target recognition (ATR) within mine countermeasures (MCM) operations in shallow waters. However, conventional machine learning (ML) and deep learning approaches face challenges in automatically mapping the seabed due to noise and limited labels. Thus, this article introduces the Deep Supervised Semantic Segmentation model for Seabed Characterization (D4SC), tailored for addressing challenges associated with sonar data. D4SC employs convolutional neural networks, specific high-resolution (HR) synthetic aperture sonar (SAS) data preprocessing and data augmentation methods, including the novel boundary pixel label rejection, and moves from the low-label regime. Performance comparisons against standard methods in the literature are conducted, demonstrating D4SC's superiority on challenging HR SAS survey datasets from real-world MCM exercises at sea. In addition, this work thoroughly explores the effect of the quality of the datasets, the robustness of training models on Out-of-Distribution data, and the estimation of epistemic uncertainty to refine predictions at large scale.**

***Index Terms*—Deep learning (DL), image segmentation, synthetic aperture sonar (SAS), uncertainty.**

I. Introduction

SEABED characterization involves the comprehensive examination of the physical and biological attributes of submerged terrains. This encompasses various methodologies ranging from direct observation by divers and remotely operated vehicles, to more indirect approaches like seabed sampling. Sonar technology, capable of capturing acoustic backscatter over wide areas, has emerged as a particularly valuable tool in delineating boundaries within relatively homogeneous seabeds [1], [2], [3], [4], [5] and automatic target recognition (ATR) within mine countermeasures (MCM) contexts. ATR methods for MCM have achieved remarkable detection accuracies on behalf of by recent developments in autonomous underwater vehicles (AUVs) and high-resolution (HR) synthetic aperture sonar (SAS) sensors [6], [7], [8]. Despite exceeding 95% in benign environment [9], [10], these works have also reported severe performance drops in challenging environments including sandwaves, rocky terrains or fields of seagrass. In these scenarios, increasing the sensor's ground resolution, to enhance target detection and recognition via low-altitude surveying, increases the amount of acoustic shadows possibly concealing targets. Consequently, accurate seafloor characterization is paramount for MCM operations, not only to qualify ATR confidence in specific areas but also to guide mission planning and enhance the autonomy of AUVs.

Additionally, the interpretation of the single look complex (SLC) SAS data obtained with the interferometric sensor presents significant challenges owing to the complicated propagation of sound underwater and the presence of various sources of noise, most notably aperture synthesis errors originating from imperfect motion compensation and speckle noise arising from coherent imaging. Other sources of uncertainty originate from distribution shifts such as different acquisition parameters employed on the sensor at survey time or seasonality. Moreover, the seabed characterization task is hardly discrete and seabed are in fact a mixed composition of sediments of different grain size [11], potentially mixed with organic derived matter. Furthermore, the intricate responses from objects smaller than the sensor's ground resolution such as pebbles or shells exacerbate the difficulty of seabed characterization with single frequency sonar. In light of these challenges, this work aligns with prior research efforts [12], [13], [14], [15], [16], decoupling the seabed characterization task over smaller Regions of Interest (RoI) on which perform semantic segmentation. This consists in feeding the acoustic backscattered intensities, into models producing segmentation maps by assigning a label to each pixel.

However, none of the aforementioned conventional artificial intelligence (AI) approaches, either machine learning (ML) or deep learning (DL), have addressed the automatic seabed characterization at large scale. Therefore, this work introduces the Deep Supervised Semantic Segmentation model for Seabed Characterization (D4SC). Its contributions are as follows.

1) It leverages the work of [17] within the low-label regime by extending the DL end-to-end pipeline, with novel boundary label rejection, specific data augmentation methods

Received 1 May 2024; revised 31 August 2024; accepted 6 September 2024. Date of publication 19 September 2024; date of current version 11 October 2024. This work was supported by the Belgian Royal Higher Institute for Defence under Grant DAP/21-11. *(Corresponding author: Yoann Arhant.)*

Yoann Arhant is with the Department of Communications, Information, Systems and Sensors, Royal Military Academy, 1000 Brussels, Belgium, and also with the Department of Telecommunications and Information Processing, Ghent University, B-9000 Gent, Belgium (e-mail: yoann.arhant@mil.be, yoann.arhant@ugent.be).

Olga Lopera Tellez and Xavier Neyt are with the Department of Communications, Information, Systems and Sensors, Royal Military Academy, 1000 Brussels, Belgium (e-mail: olga.loperatellez@mil.be; xavier.neyt@mil.be).

Aleksandra Pižurica is with the Department of Telecommunications and Information Processing, Ghent University, B-9000 Gent, Belgium (e-mail: aleksandra.pizurica@ugent.be).

Digital Object Identifier 10.1109/JSTARS.2024.3465231

and tayloring existing Bayesian uncertainty prediction refinement for SAS data.

2) It focuses on real-world HR SAS MCM operational datasets, characterized by significant variability in survey acquisition parameters, extensive survey areas in shallow waters worldwide and most detailed annotations. D4SC achieves state-of-the-art results on those datasets, with one left unseen at training for Out-of-Distribution (OoD) discussions. Finally, this work also thoroughly analyses the biases induced by the annotation process of SAS training data over the proposed end-to-end pipeline.
3) It follows [18], [19], and [20] on Bayesian uncertainty estimation over synthetic aperture radar (SAR) and SAS datasets, respectively, and it proposes an in-depth evaluation of those uncertainty estimation methods over SAS datasets.

The rest of this article is organized as follows. The related work is introduced in Section II, followed by a description of the SAS dataset and associated challenges in Section III. Details of the D4SC model and its end-to-end pipeline are provided in Section IV. The results of D4SC over the dataset are presented in Section V, further validated through an extensive ablation experiment. Section VI explores multiple experiments such as the effect of pretraining and the effectiveness of the epistemic uncertainty estimation method based on Monte Carlo Dropout (MCD), Deep Ensemble (DE) and both of them, to refine prediction mappings. Finally, Section VII concludes this article.

II. Related Work

A. ML Tailored to Sonar Data

Despite SAS being inspired from SAR, the complete end-to-end processing has to be tailored from the sensors to the final post-processing steps due to the intrinsic differences of the propagation, namely the nature of waves, mediums, absorptions, etc. For example, the popular phase gradient autofocus (PGA) was adapted for SAS in [21]. Hence to harness SAS data, prior studies [14], [16], [22], [23] have employed the multiple-instance learning via embedded instance selection, which involves computing features at the superpixel level and subsequently applying possibilistic clustering or possibilistic classification for acoustic segmentation. However, these methodologies assumes that both the distribution and the model are well suited to describe all relevant clusters, a condition unlikely to hold in real-world scenarios. Alternatively, [12] has directly grouped pixels into target label maps, leveraging lacunarity and a simple maximum likelihood estimator. In addition, [20], [24], and [25] have first concurrently addressed the challenges of large-scale automatic mapping of the seabed by crafting standard Computer Vision (CV) features to feed to a Gaussian process classification and an adaptive hierarchical Dirichlet process clustering, respectively. However, despite these efforts, these methods have led to unsatisfactory results since the resulting unprocessed classification have reached 80% pixel-accuracy and 90% RoI accuracy, respectively, over simple real-world SAS datasets.

B. Standard DL Approaches

Conversely, DL culminates with the application of Transformer models [26], [27], [28], [29], reaching billions of parameters and capable of handling the training on datasets comprised of billions of images [30], [31]. This paradigm is the high-data and high-label regime leading to models being highly robust to noise and outliers. However, when confronted with smaller datasets comprising few annotated images, falling within the low-label regime, alternative learning strategies become necessary such as unsupervised, semisupervised, and self-supervised learning (SSL) with convolutional neural networks (CNN). Standard CV SSL approaches often rely on Contrastive Clustering [32] or more generalized contrastive learning (CL). Those methods consist in learning similar feature representations from different augmented views of a same image. Some examples of SSL with CL can be found in [26], [27], [32], [33], and [34] for image classification, or even [35] for semantic segmentation. While standard CL methods depend on many negative examples to repel the feature representations of different images, [34] train instead an online model trying to predict, with an extra convolutional layer, the same feature representations as the one of a different augmented view of the same image but inferred by the target network. This network is defined as an exponential moving average of the online model weights. This approach has shown remarkable performance while reducing the self-supervised training time. It was extended to semantic segmentation in [35] by aligning the different feature representations at the deepest layer of the feature extractor, which in a CNN is a low-resolution grid of features, and applying a similarity loss between the part of the grid in common. This method has outperformed the existing image classification SSL pretraining for object detection and semantic segmentation downstream tasks. In contrast, other approaches exploiting the high-data regime can adopt active learning (AL). They encompass iterative human-in-the-loop methods to optimize annotation resources such as in [36], [37], [38], and [39].

C. DL Tailored to RS

The remote sensing (RS) community gathers different types of sensors such as Optical, SAR, Multispectral, or Hyperspectral acquired either airborne or by satellites. For instance, [40] and [41] merged multiple of those Satellite-based imagery into classification datasets. The authors in [40] also proposed a multimodal and multiresolution approach with domain adaptation to address the semantic segmentation of the different geographic zones. This variety in data and tasks leads to a overwhelming diversity of methods addressing them. For example, [42], [43], [44], [45], [46], and [47] focused on unsupervised tasks. The authors in [42], [44], and [46] employ low rank matrix representation methods for multispectral data based on the deep unrolling of the Alternating Direction Method of Multipliers algorithm. In addition, [45], [47], and [48] adapted successful DL methods from CV to RS. Specifically, [47] extended the work of [29], to multitemporal and multispectral data to train a Vision Transformer. The authors in [48] harnessed [49], which was primarily designed for explaining the interest in knowledge distillation [50], to assess the quality of retrained models on SAR data for classification. Despite its efficiency, this methods necessitates the instance segmentation labels of the object in order to count the number of informational points projected on the image.

The existing literature on seabed characterization from SAS data has also explored several of the aforementioned DL methodologies derived from CV, with studies such as [51] and [52], respectively, investigating semisupervised learning and a CL term based on geographic distance. In scenarios characterized by both low-label and low-data regimes, successful DL approaches often incorporate domain adaptation with transfer learning strategies including [51] and [53]. These methods also leverage deep AutoEncoder architectures to exploit the inherent regularization effect of dimensionality reduction in the embedded space, which also serves as an unsupervised pretraining method with a reconstruction loss in [54]. Specifically, [13] and [55] and [51] and [53] have trained a plain AutoEncoder, a ladder network, and U-Net based models, respectively. Furthermore, [51] have explored unsupervised pixel-wise segmentation using superpixels and transfer learning, aiming to leverage knowledge learned from natural images.

In [53], MLSP-Net proposed as a differentiable angle of arrival decomposition method employing fast Fourier transform and a filter-bank learned by a CNN model. This methodology incorporates multiple azimuth pass-band filters in k-space, effectively simulating a time decomposition as the AUV progresses forward during data acquisition. This approach aids in the recovery of information from the movement of fishes within the sensor's line of sight and addresses some shadow effects arising at long ranges. Inspired by the principles of SAR azimuth multilook processing, the work of [53] tailored the conventional CNN architecture to accommodate SLC SAS data. Unlike conventional methods that mitigate noise by averaging the looks, this method fed the CNN output features maps from each look into a convolutional long short-term memory network. This represents the third attempt to characterize the seabed with SLC data, rather than discarding them in favor of intensities, as in [56], [57].

D. Uncertainty Estimation in DL

While DL models achieve high performance for a wide variety of tasks, they often exhibit high confidence in misclassified predictions or OoD patterns, then requiring recalibration for critical tasks [58]. The confidence of DL models is defined as the maximum of their output, particularly in classification tasks where it is represented by softmaxed logits corresponding to probability vectors. Hence, numerous approaches have been developed to address uncertainty estimation of DL models, for which a comprehensive survey can be found in [59].

For uncertainty estimation, [60] have introduced the MCD, which involves keeping dropout activated during inference to produce nondeterministic prediction vectors. This approach provides a Bayesian Interpretation as Variational Inference by enabling the estimation of the posterior distribution of the model's weights given the input data distribution. By averaging a sufficiently high number of nondeterministic predictions, MCD allows for the quantification of epistemic uncertainty inherent in the end-to-end learning pipeline and the limited coverage of the input data distribution over the distribution of the real-world task. In [61], this method has been extended to semantic segmentation and aleatoric uncertainty estimation, which accounts for uncertainty arising from the data acquisition process in regression tasks. The latter has been investigated for RS in [18] and [62].

Alternatively, other approaches rely on DE [63], [64], [65] to estimate the posterior distribution of the model's weights. Conversely in [66] and [67], the CNN models are trained with specific optimizers to learn a Laplace approximation of the posterior. Moreover in [68], the CNN derived Dirichlet Prior Network learns the conjugate prior on the softmaxed distribution by considering it as a Dirichlet distribution. To achieve this, this method employs a dual loss and train over both in-distribution and OoD datasets to learn the boundary between them, effectively distinguishing between in-domain uncertainty and OoD uncertainty. [37] compares aforementioned Bayesian derived methods with the CV AL task, while [69] and [70] extended such a benchmark for RS, respectively for change detection and semantic segmentation. In addition, [19] also performed Bayesian uncertainty estimation for SAR-based road segmentation and OoD.

In contrast, other uncertainty explanation methods often rely on Grad-CAM [71] which analyzes the gradients of the final convolutional layer for each class, creating a heatmap that highlights the contribution of each pixel to the final classification. In addition, [72] extended it for semantic segmentation. Conversely, Grad-CAM [71] being unable to address missed object detections, [73] extended [74] for uncertainty explanation in terms both of regression of the localization and classification.

III. Description of the Datasets

Over the past decade, the Centre for Maritime Research and Experimentation has conducted numerous sea surveys using the MUSCLE AUV, which is equipped with a high-frequency side-looking interferometric stripmap SAS sensor operating at a central frequency of 300 kHz. Given the low altitude of the AUV, about 10 m, it gathered a large amount of data at a high ground resolution up to of 1.5 cm acrosstrack, i.e., in ground range. The surveys employed in this study, comprehensively described in [12], are decomposed into to 18 627 SAS images, hence falling under the high-data regime. They are characterized below as follows with emphasis placed on patterns not addressed in the final classification, highlighted in italics.

1) Arise 1 (ARI1): Conducted in the Mediterranean Sea, Italy, the survey is comprised of *Megaripples*, Sandwaves, Sand Ripples, Fine Sand, Medium Sand, *Sand with Shells and Shell Debris*, *Mud*, Alive Posidonia and its *Dead Matte*. In addition, the survey showcases human *activity traces* including trawling marks, small underwater objects like trash metal oxygen cylinders or fishing gear, *shipwrecks*, underwater *cables*, and *pipelines*.

2) Colossus 2 (COL2): Conducted in the North Sea, Latvia, the survey is comprised of *Megaripples*, Sandwaves, Sand Ripples, Fine Sand, Medium Sand, *Coarser Sand with Granule*, Pebbles, Medium sized Rocks, Boulder, and Rock Outcrops. In addition, the survey presents *trawling marks*.

3) Minex 18 (MNX18): Conducted in the Mediterranean Sea, Spain, the survey is comprised of *Megaripples*, Sandwaves, Fine Sand, Medium Sand, *Mud*, Alive Posidonia, and its *Dead Matte*. With different acquisition parameters and settings as distribution shift, the Minex 18 survey displays different aperture synthesis errors and low-constrast noise.

In the absence of in-field analyses and due to the inherent challenges to SAS data, reliably characterizing the seabed, especially solely based on SAS images, is highly challenging. Thus,

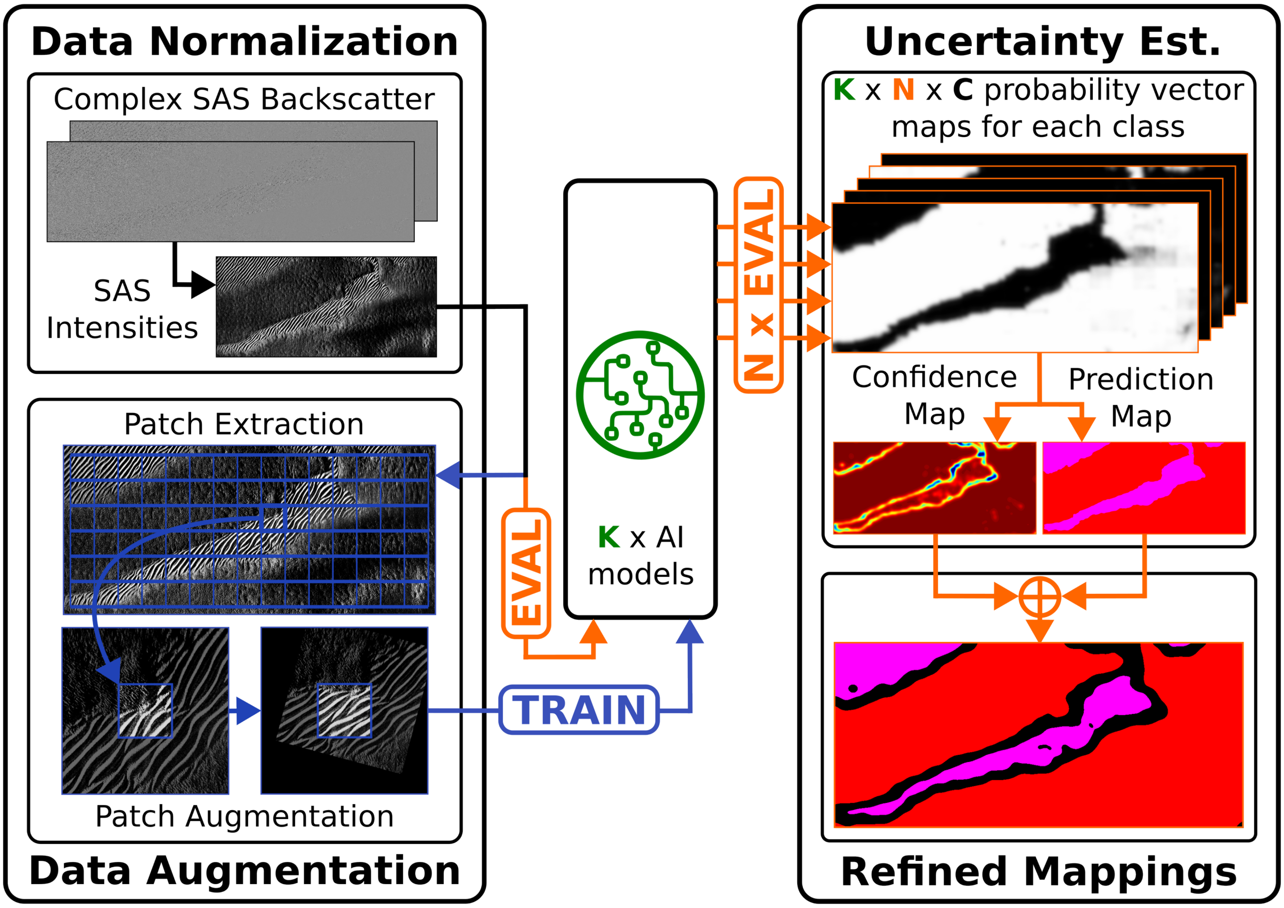


Fig. 1. Workflow of the proposed end-to-end pipeline encompassing data normalization, data augmentation for training the **K** AI models. They are evaluated **N** times for uncertainty estimation producing both a high resolution predictions, as a probability vector map of **C** dimensions per pixel, and a confidence map associated. Then, they can be used to refine the predictions.

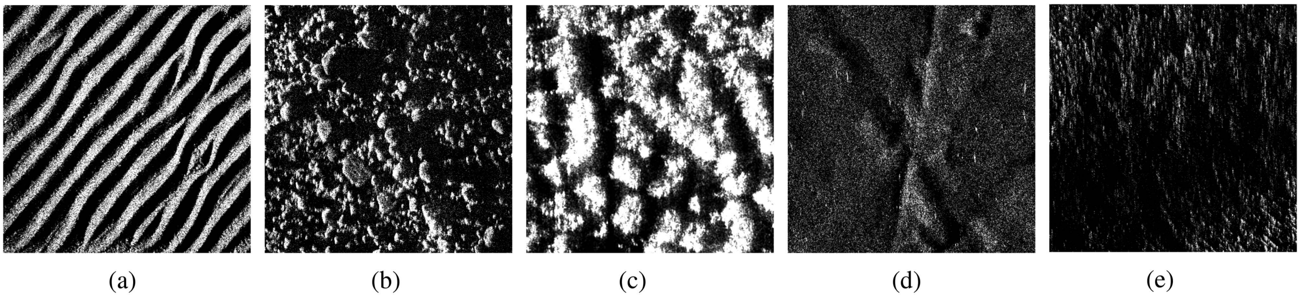


Fig. 2. Examples of pure class patch of size 512 × 512, where all pixels get assigned the same class, sampled from the COL2 and ARI1 datasets sorted by increasing subjective operator's difficulty to annotate. While (a) and (b) are well defined, (c), (d), and (e), respectively, show the effects of time and seasonality with the growing Posidonia, human activity with Trawling Marks engraved in Flat Sediments, and finally long range shadows and motion estimation errors on Sand Ripples. (a) Sandwaves (SW). (b) Rocks (RK). (c) Posidonia (PS). (d) Flat Sediments (FS). (e) Sand Ripples (SR).

aforementioned characterizations are simplified into five operational semantic classes for MCM applications [12]: Locally Flat Sediments (FS), Posidonia (PS), Rocks (RK), Small Sand Ripples (SR), and Sandwaves (SW). Examples of pure class patch samples can be found in Fig. 2. In addition, an extra Unknown (U.K.) class was introduced to accommodate any pattern impossible to fathom without more global context, either resulting from loss of contrast at long range or shadows cast by tall formations, as illustrated in Fig. 4(a) and (b). Similarly to the background class in Semantic Segmentation in CV, errors on such pixels are disregarded both during backpropagation at training and in the computation of evaluation metrics during testing. All the aforementioned difficulties contribute to low quality annotations accounting for a data distribution with a lot of label noise.

The degenerate input data distribution refers to cases where identical data points are assigned different labels. In machine learning (ML), this can arise from annotation errors or hash collisions within the feature extractor. While this concept is straightforward, it becomes more complex in the context of deep

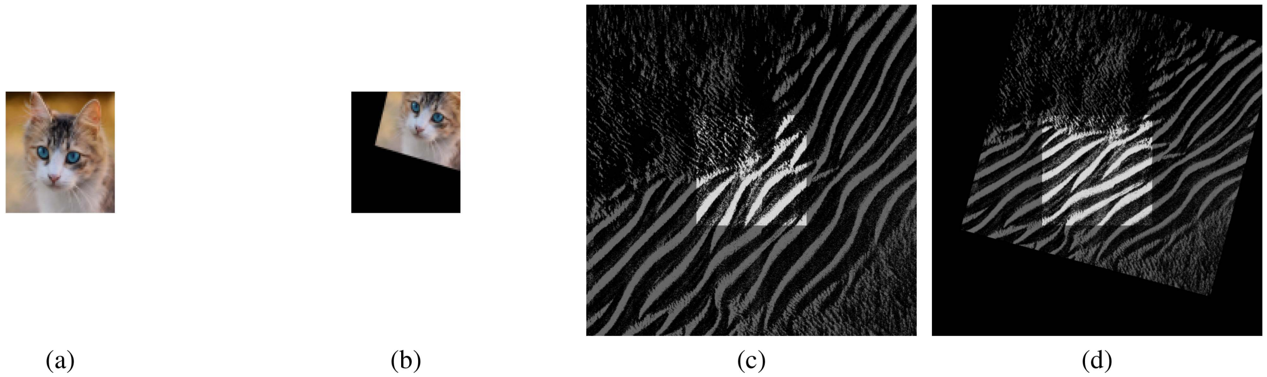


Fig. 3. Diagram illustrating the random data augmentation pipeline consisting of Flips, Rotation, Scaling, Translation, Intensity and Contrast Jittering applied to (a) following classical data augmentation schemes from CV for classification and the proposed one (c). The resulting cropped patch represented by the size of the image in (a) only retains existing pixel information producing the zero-padded 256 × 256 pixels patch in (b), whereas the presented method fills the cropped patch with natural input information (d), represented by the lighter square also in (c).

TABLE I
DATASETS DESCRIPTION AND DISTRIBUTION OF THE TRAIN AND TEST DATASET OVER COL2 AND ARI1

Dataset	Full-size 2000 × 4400 images				256-size patches				512-size patches				Labeled Pixels
	Total		Labeled		Total		Labeled		Total		Labeled		
	COL2	ARI1	COL2	ARI1	COL2	ARI1	COL2	ARI1	COL2	ARI1	COL2	ARI1	
IGARSS 2023 [17]	8219	7613	16	0	1.4 M	1.3 M	2736	0	411 k	381 k	800	0	141 M
Train			70	84			11965	14086			3500	4129	1232 M
Test			13	14			2379	2078			698	609	237 M

TABLE II
PIXEL CLASS LABEL DISTRIBUTION IN THE DIFFERENT SURVEYS:

Survey	UK	SR	RK	SW	FS	PS
COL2	1.2%	24.7%	14.8%	5.2%	54.1%	–
ARI1	1.7%	–	–	20.1%	71.3%	6.9%
Train Set	1.5%	11.3%	6.8%	13.3%	63.4%	3.7%
Test Set	6.2%	13.7%	6.5%	18.9%	49.6%	5.0%
MNX18	1.7%	–	–	47.2%	19.0%	32.1%

learning, where the feature extractor is integrated in the model, and tasks like semantic segmentation are more sophisticated. In these scenarios, it's more appropriate to focus on the quality of the decision boundaries in relation to their support at the ultimate layer, which separates classes, as the result of the optimization process over the model. Moreover, with deep feature extractors like those in [75], where the receptive field often surpasses the size of the output image, defining a degenerate input distribution is impractical; the issue is more accurately described as label noise.

Table I summarizes the repartitions of the different train and test sets across COL2 and ARI1. Particularly, the Train Set is obtained by annotating manually 154 randomly selected images from ARI1 and COL2. This Train Set is an extension of the low-label regime dataset from [17]. To constitute the Test Set, one of each pattern sample not addressed in the final characterization altogether with images encompassing a high number of boundary pixels are manually labeled from ARI1 and COL2. Such an ambitious Test Set measures the performances of models across an extensive variety of seabed configurations for reliable large-scale mapping. Given that locally Flat Sediments are predominant in shallow waters, this class is the most represented in the Train Set, accounting for up to 63.4% of total pixel labels, thereby exacerbating the class imbalance issue, as depicted in Table II. Moreover, the distribution of the pixel class labels in MNX18 differs significantly from that of the Train Set, enabling further validation of the model's robustness to imbalanced data through evaluation on this dataset.

In summary, this study focuses on real-world HR SAS MCM operational datasets, characterized by significant variability in survey acquisition parameters and extensive survey areas in shallow waters worldwide. The annotations associated with these challenging datasets are the most detailed in the seabed characterization literature with SAS data, as shown in Fig. 4(b), Appendix B, and Fig. 13(c) and (d), compared to [20], [51], and [76]. These details include instance segmentation of semantic patterns smaller than a patch, such as individual rocks, spot of posidonia, dunes composing sandwaves, as well as finely localized boundaries between classes.

IV. Method

In this section, the complete end-to-end pipeline encompassing the training and inference of the CNN is introduced. It consists in data preprocessing and augmentation, CNN model architecture and initialization, CNN model training schemes, and some postprocessing methods for inference. It is summarized in the Fig. 1. While the training updates the weights by backpropagating the gradient of the loss function, which is computed from the forward pass, during the backward step, the inference corresponds the forward evaluation of the model.

A. Data Preprocessing

To preprocess the SLC data acquired by the SAS sensor, this article converts the backscatter coefficients to intensities, apply the normalization *a)*, compress the dynamic range of the SAS data *b)*, interpolate the intensity maps *c)*, optionally reject the

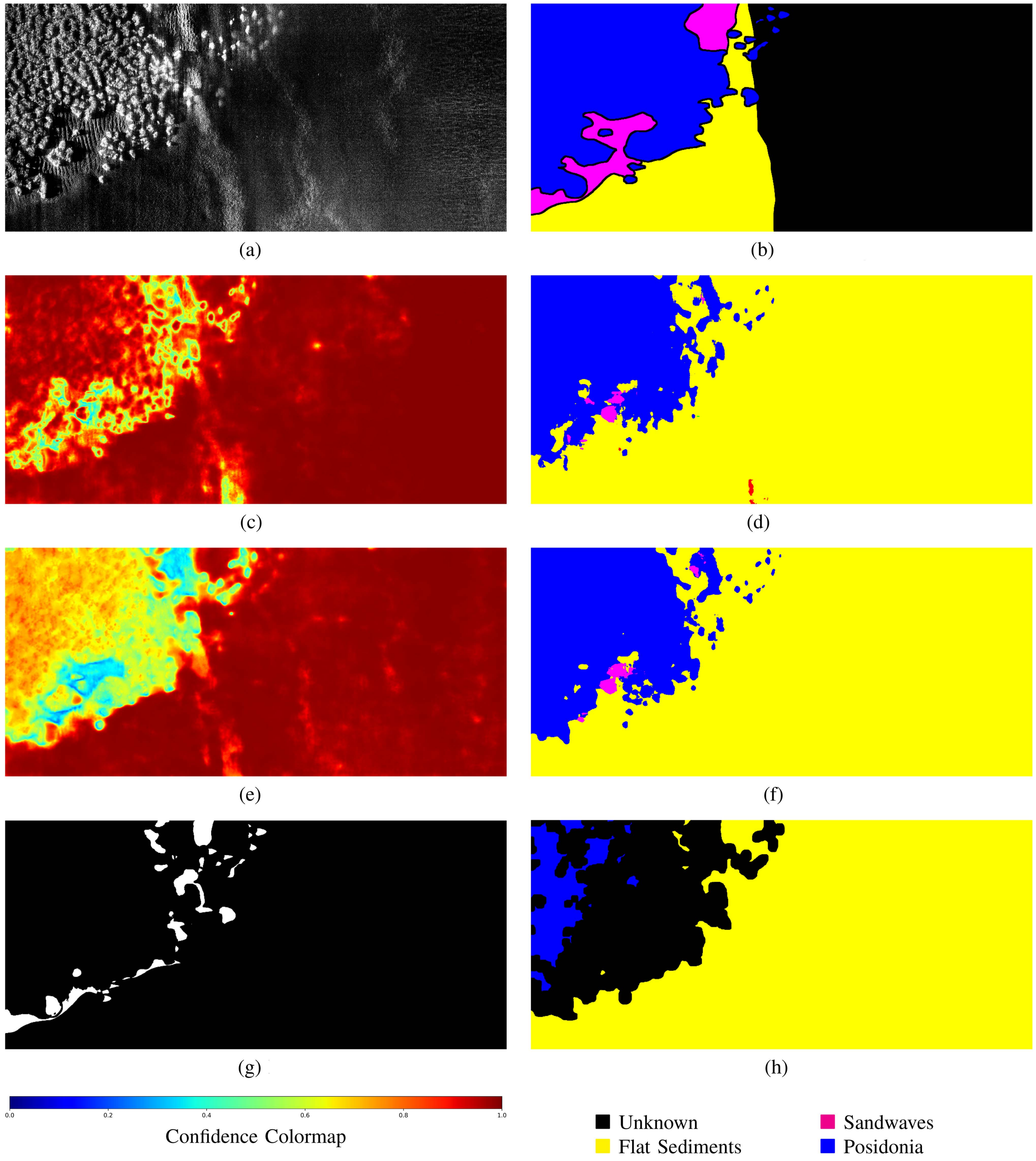


Fig. 4. Example of normalized SAS Backscatter Intensities (a) from a difficult SAS image from the Test set. It displays a boundary between Soft Mud, locally considered flat, with Medium Sand arranged in Sandwaves, where growing Alive Posidonia also arranges in ripples pattern. Its associated ground truth is depicted in (b). At annotation time, the long-range loss of contrast, due to motion compensation issues in the aperture synthesis, resulted in patterns impossible to fathom, hence the annotator labeled such pixels as Unknown. In addition, the Unknown pixels at the boundaries between two classes corresponds to the Boundary Label Rejection (BLR) step. Comparison between deterministic inference standard confidence (c), associated predictions (d), and the DE average probability vector confidence derived from entropy (e), and associated predictions (f) are also presented. The DE average probability vector OACC is 90.8%. This corresponds to summing out the error map (g). The refined predictions (h) from thresholding the associated confidence map (e) raise even more this OACC up to 91.8% for 74.6% PPC by rejecting sandwaves being predicted as Posidonia, recovering from misclassifications as shown in the error map (g). (a) Normalized Backscatter Intensities. (b) Ground truth map. (c) Deterministic inference standard CNN confidence. (d) Deterministic inference predictions. (e) DE confidence from Entropy. (f) DE average predictions. (g) DE average error map. (h) DE refined predictions.

boundary labels of the ground truth *d)* and extract patches for minibatch training *f)*.

a) Median Normalization: This work follow the procedure described in [77] to take into account the absorption of sound in range, the grazing angle and the ensonification pattern, by normalizing acrosstrack vectors by their median before normalizing alongtrack vectors by their normalized median values.

b) Logarithmic Compression: Typical SAS intensities in a single image can span across nine orders of magnitude which can be cumbersome to disentangle for residual CNNs models that learn slight differences from identity [78], therefore images are compressed in logarithm scale as shown in Fig. 4(a). In addition, the log-scale scattering average strength of the seabed response is close to linearly correlated with the sediment grain size at fixed grazing angle as reported in [79].

c) Square Grid Interpolation: As in [77], images at resolution 1.5 cm in ground range and 2.5 cm in azimuth are interpolated to a square grid of 2.5 cm. This enables us to avoid introducing elastic deformations with the rotational part of the data augmentation.

d) Boundary Label Rejection (BLR): Although polygons were drawn to preserve boundaries between classes, as stated by [12], the annotation task is hard and sometimes the resulting borders are rather arbitrary. Therefore, a morphological Erosion operation is applied with circular kernel of 15 pixels to all class label maps, effectively rejecting boundaries into the Unknown class, as the weak labels of [53]. The effect of this preprocessing step can be observed in Fig. 4(b). This approach differs from [80], which enhances semantic segmentation results in CV by predicting a boundary map. Subsequently and within the model, postprocessing operations are conducted to refine the class label maps with this boundary map.

e) Patch Extraction: At the end, the large sonar images of size 4400 × 2000 and their corresponding ground truth maps are split into squared patches of size either 256 × 256 or 512 × 512. This method, using a sliding window, divides large images into smaller patches, potentially overlapping depending on the stride. This simplifies data augmentation and DL training with minibatches, which both help regularize the convergence of the model.

B. Data Augmentation (DA)

As neural networks grow deeper, they often require more data to achieve satisfactory performance. Augmenting datasets extensively can be one way to address this need. While traditional computer vision approaches to data augmentation typically involve affine and color-space transformations, they are extended with a specific emphasis on the physical characteristics of sonar data, as illustrated in Fig. 3. However, directly applying these augmentations the same way as in CV tasks can lead to undesirable outcomes in seabed semantic segmentation. This is because such augmentations may introduce nonreal looking textures, create nonnoisy seabed representations, or generate unrealistic shadows in the input distribution. To address this concern, on-the-fly augmentations are performed over larger input patches, ensuring that the final crop only contains natural input information [see Fig. 3(d)]. However, this approach comes with a tradeoff as patches closer to the edges than one patch size cannot be augmented in this manner. In addition, since shadows are only cast in range, the proposed method limits the application of rotations to small random angles, under 15°, to keep the physical meaning of sonar images.

C. Architecture Design

In contrast to [17], the labeled Train Set was significantly expanded by annotating ten times more images, as depicted in Table I. Consequently, this dataset no longer falls within the low-label regime, enabling the use of deeper and more reliable CNNs. Thus, D4SC's CNN architecture builds upon the foundation of Deeplab [81], with a replaceable feature extractor as the encoder. Feature extractors such as those proposed by Tan et al. [75] replaces the U-Net based feature extractor.

The decoder, which includes bilinear upsampling operations and skip-connections, is automatically generated based on the number of decimations introduced in the encoder. Moreover, convolution layers, excluding the initial one, are substituted with the residual blocks of EfficientNet [75]. These blocks have been demonstrated to be state-of-the-art on embedded systems with limited cache memory, as evidenced by the work of [75] which demonstrated their effectiveness across various sizes and applications. The resulting D4SC model is another Fully Convolutional Network (FCN) [82], virtually capable of processing images of any size and aspect ratio.

D. Model Initialization and Training

When initializing the weights of CNN models, this article follows the procedure outlined in [83]. This initialization method, originally developed for residual networks, has demonstrated faster convergence and improved performance for a wide range of applications and has consequently been adopted.

According to [84], under the assumption of infinite time and a nondegenerate input data distribution, over-parameterized models always converge to the global minimum, thereby yielding the same predictions regardless of different initialization. However, this ideal scenario is hardly encountered in the seabed characterization datasets due to their arbitrary boundaries and loss of contrast at long range, this is the reason why the proposed method trains multiple D4SC models randomly initialized and then selects the best converged one, achieving peak performance, for large scale mapping.

In addition, the Train Set is partitioned into training and validation sets, with a split of 90% and 10%, respectively. To prevent the apparition of outlier results due to insufficient training of the CNN, an early stopping strategy with patience is employed. This strategy continues backpropagating the balanced cross-entropy loss until no improvement over the validation set can be observed, ensuring greater replicability of the complete training pipeline. Furthermore, as standard approach to address class imbalance in the input data distribution, the contribution of pixels in the loss are weighted by the inverse of class label frequency in the ground truth. The initial learning rate is set at 0.0015 and is employed within an Adam Optimizer strategy, eliminating the need for manual adjustment. In addition, small dropout rate of 0.05 and weight decay of 10^{-6} are introduced as regularizers. To insert dropout layers in the residual models, this work adapts the findings of [85] to different feature extractors.

The training phase, which typically ranges from one to a few hours, is sufficiently fast to conduct ablation experiments testing different metaarchitecture parameters, resulting in D4SC architecture, and to evaluate the consistency of model convergence. This is validated by the standard deviation of the mean Accuracy (mACC) across the 256-size patch Test Set of models randomly initialized repeatedly trained, as statistical dispersion metric accounting for the reliability of a model convergence. This can also be visualized using box-and-whisker plots, as shown in Figs. 8 and 9.

Otherwise, pretrained models are also retrained, as a kind of diverging initialization in the domain adaptation paradigm. This article either keeps the pretrained weights of the feature extractor models from ImageNet [86], or pretrains the model over the complete, labeled and unlabeled, datasets with BYOL [34] and [35]. Then, pretrained models are finetuned with tailored and different initial learning rates given the depth of the layers. While the layers computing the high-level features, which corresponds to the second part of the model after the second decimation operation, get assigned 0.05, the low-level features get assigned 0.0025. Indeed in transfer learning, their weights are considered representative enough of the downstream task. The rest of the training remains unchanged for pretrained feature extractors.

E. Postprocessing

To evaluate D4SC and facilitate the automatic mapping of the seabed, several postprocessing steps are employed that differ from both the training process and standard inference methods used in DL CV.

1) Full-size image inference: Since the patch extraction method is used for training, which involves dividing large images into smaller patches using sliding windows, the aggregation of classification outputs from the model to the size of the SAS image necessitates a reconstruction strategy. To tackle this issue, one approach could involve selecting a stride of half the patch size during patch extraction. However, this would entail four times the initial computations and would require deciding whether to discard the predictions at the edges of each patch or to implement another fusion strategy. In addition, discrepancies are inevitable, particularly in extreme cases where pixel predictions at the corners of a patch, drawn from a nonoverlapping set, may result in receptive fields filled with padded features, thus missing at least three-quarters of the information from the full-size image prediction. Instead of delving into such complex procedures that would be difficult to compare and validate, this work leverages Fully Convolutional Networks (FCNs) yielding semantic results for images of any size and aspect ratio [82]. Therefore, inference is performed over full-size images, avoiding the need for patch aggregation.

2) Estimating Epistemic Uncertainty: The presented method introduce a prediction consistency step in the end-to-end pipeline which uses MCD [60]. Specifically, it keeps dropout activated during inference to yield N nondeterministic predictions vectors $\mathbf{Y}_{n,p} = [Y_{n,p,0}, \ \ldots, Y_{n,p,C-1}]^\intercal$, for each pixel p, $n < N$ and class $c < C$. Alternatively, this work also compares it to DE and a combination of MCD and DE, making good use of repeated training, as another way of yielding differing prediction vectors specific to a model. Those methods allow us to compute the Epistemic Uncertainty (EU) as the entropy (H) of the predictive distribution. Contrary to aleatoric uncertainty present in the single inferences, the EU is computed from their average probability vectors

$$\mathbf{Y}_{N,p} = \frac{\sum_k \mathbf{Y}_{k,p}}{N}. \tag{1}$$

While [60] employed it as a measure of EU for classification, this metric can effortlessly be extended to each pixel p of the semantic segmentation task to generate maps of EU

$$EU_p = H(\mathbf{Y}_{N,p}) \tag{2}$$

$$EU_p = -\sum_c Y_{N,p,c} \cdot \log(Y_{N,p,c}). \tag{3}$$

The average of the relative entropy of the probability vector to the single nondeterministic distributions is also considered in this work as another disagreement divergence derived from entropy-based uncertainty sampling. This divergence is widely used in the literature on AL methods such as query-by-committee [87], where it is employed to select the most uncertain samples for annotation. It has also been applied to deep ensemble active learning in [88]. In this case, the different nondeterministic inferences can be considered CNN weights sampling, effectively creating a correlated ensemble of smaller models. In the end, the disagreement score is computed with the Kullback–Leibler divergence, which penalizes harder the variance in probability vectors and denoted KL, as follows:

$$Disagreement_p = \frac{1}{N} \sum_n KL\left(\mathbf{Y}_{n,p} \| \mathbf{Y}_{N,p}\right) \tag{4}$$

$$= \frac{1}{N} \sum_n \sum_c Y_{n,p,c} \cdot \log\left(\frac{Y_{n,p,c}}{Y_{N,p,c}}\right). \tag{5}$$

As it is not inherently a true metric of uncertainty, and therefore challenging to calibrate for model confidence derived from uncertainty estimation, its absolute and symmetrized version is employed instead. The Absolute Symmetric Disagreement (ASD) is

$$ASD_p = \frac{1}{N} \sum_n \sum_c \left| (Y_{n,p,c} - Y_{N,p,c}) \cdot \log\left(\frac{Y_{n,p,c}}{Y_{N,p,c}}\right) \right|. \tag{6}$$

3) Uncertainty or Disagreement Refined Mappings: Although those pixel-level disagreements can be summed out for comparison to hard thresholds, they can reject predictions from uncertain pixels in the mappings such as in Fig. 1 and 4(h). An optional label map refinement step is added, where each individual class label map rejected from the uncertainty got reduced by morphological filters. Specifically, individual class label map undergoes Closing, Opening, and Erosion operations to filter out impossible predictions of too small seabed textures. In addition, the final prediction map is also reduced by Erosion using the same circular kernel of size 15 pixels, similarly to the BLR operation introduced in Section IV-A-d). This work believe it is preferable to err on the side of caution by rejecting slightly more of automatic predictions, especially at boundaries. Therefore, the Pixel Prediction Coverage (PPC) can be defined as

TABLE III

Comparison With the State of the Art on the Test Set of Each Best Performing Model, Called Peak, Selected After ten Repeated Trainings Corresponding to Different Random Initialization, in Terms of Deterministic Inference Across the 256-Size Patch Mean Accuracy (mACC) and Pixel-Wise Overall Accuracy (OACC) Over the Full-Size Image

Model	Trainable parameters	Full size image inference GFLOPs	Best model evaluated over the Test Set		256-size patch Test Standard Deviation over mACC
			256-size patch mACC	fullsize image OACC	
IGARSS 2023 [17]	958k	1035	82.4%	85.0%	2.5%
MLSP-Net [53]	52.3M	16800	80.0%	–	–
D4SC (ours)	11.4M	1080	**87.6%**	**88.5%**	1.6%
Resnet50-UNet	25.8M	1120	87.4%	87.2%	**1.0%**

The number of trainable parameters and Giga Floating-point Operations (GFLOPs) accounts for the model's efficiency.
Bold values indicate the best results for the different performance metrics.

the percentage of pixels which get assigned a label in the refined prediction mapping to compare between different uncertainty rejections approaches.

V. Results and Ablations

A. Comparison With SotA

For DL-based segmentation, prior arts applied domain adaptation with transfer learning or retrained models from scratch, without tailoring the architecture neither to the data nor to the seabed characterization task. To compare the proposed model the MLSP-Net model from [53] is implemented with its two U-Net backbones [89], implemented with a ResNet50 feature extractor trained on ImageNet, following standard transfer learning schemes. For a fair comparison, augmented 512 × 512 patches, which are extracted from SLC normalized images with the normalization used in [53], are fed to MLSP-Net but the median normalization is added to avoid fooling the model from the higher dynamic inherent to the full-size images datasets. In contrast to the Keras implementation, a ResNet50-Unet segmentation model from PyTorch pretrained on ImageNet, features larger dimensions in the embedded space, consequently expanding the number of trainable parameters. This is no issue for this model as it is implemented with ten times more labels and data augmentation. However, the seabed characterization scheme with darker sand and shadows is not reproduced, the weak labels employed for the training of MLSP-Net are simulated by increasing BLR up to 50 pixels. In the end, this method is training for two days, preventing this article to train it repeatedly and report its average Overall Accuracy (OACC) standard deviation over the full-size Test Set.

The deterministic D4SC outperforms MLSP-Net, as reported by Table III with 87.6% best model mACC across the full-size images of the Test Set against 80.0%. This suggests that MLSP-Net is more subject to the effect label noise and to the choice of semantic classes, instead of more acoustic classes, than other more conventional CNNs. Despite the higher performance of the deterministic D4SC on the full size image Test Set compared to Resnet50-UNet pretrained on ImageNet, Table III also report less FCN convergence consistency over the mean Accuracy (mACC). This suggests that the pretrained weights or higher size of the models in term of training parameters limit the apparition of models stuck in an underperforming local minimum. The ResNet50-UNet pretrained on ImageNet serves as the backbone for MLSP-Net, and its lower OACC on the Test Set indicates that the reduced performance is attributable the other modifications.

B. Early Performance Boosting by Averaging Predictions

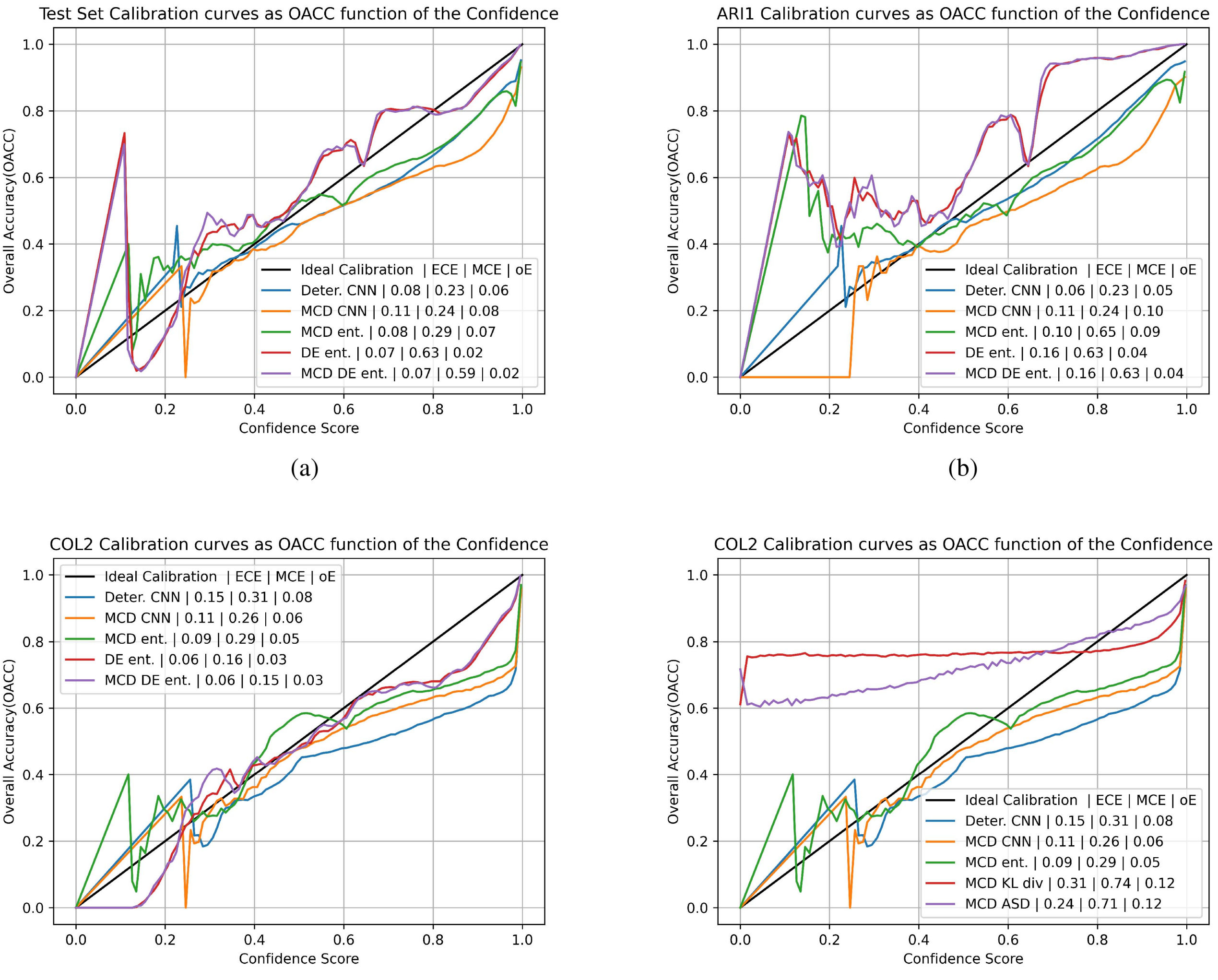
Additionally, the OACC improves significantly with the average probability vector of 30 MCD, 10 DE, and 10 DE 30 MCD, respectively, reaching up to 92.0%, 91.4%, and 89.9%, respectively, to 30, 10, and 300 times the computation cost. To align with the confidence of the CNN models, this work opts to display uncertainty maps as confidences. A visual comparison of confidence maps Fig. 4(c) and (e), respectively, associated with the predictions shown in Fig. 4(d) and (f).

As anticipated, D4SC exhibits high confidence over misclassified pixels and boundaries with the deterministic CNN confidence map. While confidence maps derived from averaging and entropy also exhibit relative confidence in misclassifications, they penalize them more effectively, particularly at boundaries. However, the prediction and confidence maps appear uncorrelated, especially at the boundaries, as if the predictions are occasionally overwhelmed locally by the most prevalent class. This suggests another form of overfitting driven by the semantically homogeneous nature of the training SAS images of the seabed. Thus, this heavily influences the learning process and was further discussed in [90]. Furthermore, despite the performance improvements observed with all predictions derived from averaged probability vectors, the increased complexity leads to lower confidence, indicating negative interactions from averaging. This effect will be examined further in the next section.

C. Identifying Outliers in the Test Set

By analyzing the OACC across individual images in the Test Set for all repeatedly trained models, two images are identified as outliers, one from COL2 and the other from ARI1. These images exhibit significantly fluctuating OACC values, frequently dropping below 50% for several models, a phenomenon not observed with the other images of the Test Set. This variability is also evident in the sampling rates ablations shown in Fig. 6(a) and (b). For instance, in the last repeatedly trained model from DE, the deterministic inference yields OACC values of 27.8% for COL2 and 7.6% for ARI1, while a single inference with dropout activated at inference time improves performance to 89.3% and 95.8% OACC, respectively.

The two outliers from the Test Set illustrate instances of model overfitting caused by label noise, as shown in Appendix B Fig. 13. Due to the bias towards Rocks (RK) of COL2 manual annotations at training but less on this image, the worst performing model from the repeated training succumbs to the bias, while the best model better recover the correct proportion of Sand Ripples (SR). In contrast, the outlier from ARI1, which features the only boundary between Sand Ripples (SR) and Sandwaves (SW) in



(a)

(b)

(c)

(d)

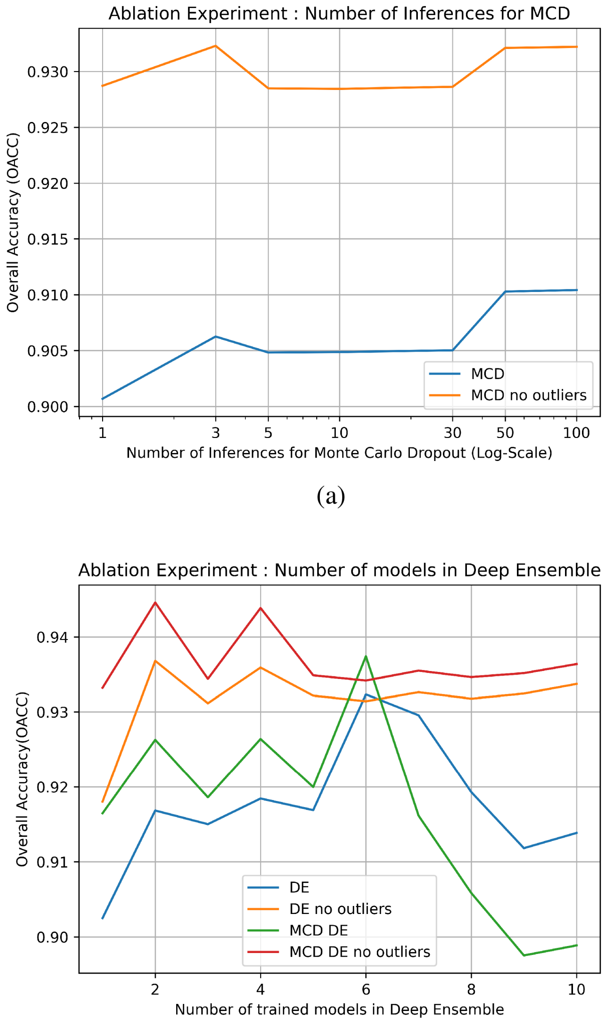
Fig. 5. Calibration curves for the different uncertainty estimation methods compared to the Deterministic CNN and ideal calibrations. The calibration curves from (a), (b) and (c) consider MCD, DE and MCD DE over the images of the Test Set, respectively, from the complete set, ARI1 and COL2. Different mathematical Operators are applied on MCD and analyzed in (d) on COL2. For each calibration curve, the Expected Calibration Error (ECE) Maximum Calibration Error (MCE) and overall Error (oE) are reported as measures of goodness of fit. The calibration curve can be understood with respect to the Ideal Calibration; the closer to the Ideal Calibration, the better, and when the curve is higher and lower, the predictions are respectively underconfident and overconfident. (a) Complete Test Set. (b) ARI1. (c) COL2. (d) COL2.

the entire set of annotated images of ARI1, as detailed in Table II, illustrates what this article terms “survey overfitting.” This is the result of the suboptimal operational parameters and survey configuration in ARI1, leading to poorly focused images. It prevents the annotator to accurately identify Sand Ripples (SR), resulting in all training samples being labeled as Flat Sediments. This is a case of distributional shift and enables the model to differentiate the two surveys based on the residual patterns of the image formation process with the acquisition parameters. Furthermore, given that the two surveys have distinct sets of classes, the best model generalized well across the Test Set, correctly predicting Sand Ripples (SR), whereas the worst model overfitted the survey, predicting only Flat Sediments (FS). This is a minor issue for MCM as the Flat Sediments (FS) and Sand Ripples (SR) have the same operational value.

After removing the outliers from the test set, the OACC of averaged predictions improves significantly with the average probability vector of 30 MCD, 10 DE, and 10 DE 30 MCD reaching up to 92.9%, 93.4%, and 93.6% OACC, respectively, compared to the previous 92.0%, 91.4%, and 89.9%. In addition, removing the outliers recovers the expected behavior of boosting predictions with MCD or DE, so that as the number of predictions averaged increases, the overall performance also improves.

D. Calibration Curves for Uncertainty Estimation

To assess the effectiveness of the recalibration and uncertainty estimation methods, previous works like [58] often compare calibration curves across different techniques, such as in Fig. 5. This is a representation of the accuracy of the model over the confidence estimated by the different methods. Following [58], this study also evaluates the Expected Calibration Error (ECE)



(a)

(b)

Fig. 6. Ablation experiments discussing the sampling rates in (a) and (b) for the different uncertainty estimation methods.

and Maximum Calibration Error (MCE) to measure the goodness of fit of the calibration curve against the ideal one. The calibration curves and these errors are calculated by binning the confidence scores and determining the corresponding OACC per bin of 1% on the test set. In addition, this work introduces the overall binned confidence Error (oE), representing the overall error rate of the uncertainty estimation process in predicting the correct confidence score relative to the Ideal Calibration per bin, i.e., below 0.5% error. Similarly, ECE and MCE can be interpreted as the mean and maximum error rates per bin, respectively.

The calibration curves for the complete Test Set, depicted in Fig. 5(a), tend to show that all uncertainty estimation methods perform similarly poorly in terms of ECE. In addition, they exhibit even poorer performance in terms of MCE. This paradoxical result can be attributed to this relatively simple seabed characterization task and the often homogeneous nature of local seabed regions. Specifically, the calibration pairs of pixel confidence scores and OACC tend to be skewed towards 100%, as the model frequently makes high-confidence predictions of the correct class. This is where the calibration curves of the different methods are closer to the ideal calibration and is further illustrated in Appendix A in Fig. 12. Therefore, the calibration curves, ECE and MCE fail to effectively assess the quality of uncertainty estimation on our SAS Test Set, as they rely mostly on bin-level information and do not capture the overall distribution. Moreover, since most of the correctness of the calibration occurs at high confidence score, the calibration curves represent the edge effects of the different methods. In contrast, the oE in Fig. 5(a) is lower for DE and MCD DE, i.e., 2% compared to 6% for the deterministic CNN confidence, indicating a more accurate estimation of the confidence with those methods.

Decoupling the analysis over COL2 and ARI1, increases the oE of DE-based methods in both surveys indicating that averaging surveys obscures calibration errors, as reported by Fig. 5(b) and (c). Specifically, Fig. 5(b) further stresses the inadequacy of the calibration curves to correctly reflect the quality of uncertainty estimation methods for images from the ARI1 Test Set. Remarkably, the underconfidence exhibited by DE related methods suggests the presence of some underlying effect with ARI1 annotated images, most probably the lower quality of the image formation process.

Conversely, the different uncertainty estimation methods produce much more stable calibration curves, ECE and MCE over the COL2 images of the Test Set, as shown by Fig. 5(c) and (d). Thus, further analysis can be performed to compare the correctness of uncertainty estimation approaches without much edge effects. In fact, Fig. 5(c) shows superior performance of DE-based methods over the others, by being less overconfident, also showcased in terms of ECE and MCE. Similarly, the differences in calibration of DE and MCD DE are marginal, suggesting that they similarly correctly assess their own confidence, despite the slightly higher OACC observed with MCD DE. Finally, Fig. 5(d) compares mathematical operators for uncertainty estimation over MCD only. Unsurprisingly, the KL divergence derived methods, which are not normalized, exhibit incorrect calibration at low confidence. In contrast, the entropy and the off-the-shelf CNN confidence obtained from MCD both outperform the Deterministic CNN confidence.

In summary, this calibration curve analysis stresses the difficulty of applying existing approches for uncertainty estimation from the literature to SAS data and the importance of distinguishing between the different surveys in SAS imagery to avoid concealing biases by averaging surveys.

E. Sampling and Dropout Rates Ablation for Uncertainty Estimation

In this section, the different sampling and dropout rates are discussed in a thorough ablation experiment. The default parameters in the other experiments are taken from [60] with the MCD rate of 0.05 and the number of MCD inferences of 30. They are sufficient in addressing uncertainty estimation on their regression task. In contrast, while [37], [63], and [69] were considering 5 to 15 models for DE, this work reuses the ten models generated by the repeated training.

Performing the ablation both on the number of inferences for MCD and models in DE, the OACC from a single model with only a single MCD inference surprisingly outperforms the

TABLE IV
COMPARISON BETWEEN DIFFERENT DROPOUT RATES IN TERMS OF OACC OVER THE FULLSIZE IMAGE PREDICTIONS AND STANDARD DEVIATION (STD) OF THE REPEATED TRAINING OVER MACC OVER THE 256 × 256 PATCHES

Dropout rate	Deterministic	Single MCD Inference	30 MCD	10 Deterministic DE	10 DE	10 DE 30 MCD	STD over mACC
0.05	88.5 %	90.5 %	92.9 %	–	93.4 %	93.6 %	1.6 %
0.01	90.6 %	90.4 %	91.6 %	93.4 %	92.7 %	–	0.4 %

Deterministic CNN, for which the dropout is not activated at test time, as reported in Fig 6(a) and (b). Notably, the ablation experiments are adversely affected by the highly varying predictions associated with the two outliers.

Varying the number of inferences for MCD has only a marginal impact on the average prediction OACC of the Test Set, as demonstrated in Fig 6(a). Similarly, Fig 6(b) suggests that averaging models for DE related methods, with the exception of two and four where some positive interactions can be observed, also minimally affects the resulting prediction OACC. Thus, this article could have reported DE results with fewer models considered rather than opting for the maximum number to err on the side of caution. Additionally and when excluding outliers, while employing 30 MCD in conjunction with DE further enhances the OACC, it also reveals a perfect correlation with DE, highlighting that MCD is incapable of reducing the prediction errors left from DE on the Test Set. In contrast, MCD alone or combined with DE effectively mitigate the impact of the two outliers in the experiments starting from 50 inferences, demonstrating the complementary benefits of using MCD and DE for uncertainty estimation with models overfitting over their training datasets.

Starting from a dropout rate of 0.1, the models trained on SAS datasets consistently failed to converge. Notably, as reported in Table IV, the models trained with a 0.01 dropout rate outperformed those trained with a 0.05 dropout rate when evaluated using deterministic inference—where dropout is not activated during inference. This suggests that higher dropout rates may negatively impact training by causing the models to overfit the dataset, leading to poorer performance. Conversely, when considering average predictions, the models trained with a 0.05 dropout rate exhibited superior performance, suggesting that the variability introduced by a higher dropout rate is beneficial for uncertainty estimation.

F. Interest in Monte Carlo Dropout Probability Vector

To prove the performance increase of the MCD probability vector, where some weights are randomly discarded, instead of the deterministic one, where the model is intact, the distribution of 1500 different nondeterministic inferences over the best performing model is studied in term of OACC over the test set, considering both single inference and their average probability vectors over 30 samples. The results in Fig. 7(a) illustrate the improvement of OACC when performing MCD inference compared to the deterministic counterpart, even in the case of averaging a single inference. Notably, averaging 30 single probability vectors further enhances the reliability of predictions on the Test Set as its distribution has a standard deviation of 0.05% while the one of single inference is 0.32%.

The analysis is further extended by examining the effect of using the MCD probability vector on segmentation results across two surveys, COL2 and ARI1, respectively, compiled in Fig. 7(b) and (c). As the probability distribution of nondeterministic predictions with MCD always significantly increases the OACC over the deterministic ones, about 3% on average as shown in Fig. 7(b), MCD acts as a strong inference time regularization over COL2, similarly to test-time data augmentation. This suggests that the model is strongly overfitting. Specifically the predictions the deterministic model exhibits significant overfitting towards Sand Ripples where the model is not confident, a pattern observed not only in COL2, but also in ARI1. Conversely, the distributions obtained over the ARI1 survey exhibits less overfitting in Fig. 7(c), which aligns more closely with the expected effect of MCD over the OACC over a nondegenerate dataset. However, averaging predictions over 30 inferences of MCD decreases OACC over the images of the Test Set of ARI1 compared to the single inference MCD, indicating that averaging reinforces the detrimental effect from Section V-D. Indeed, where MCD fails to recovers for the Sand Ripples overfitting in ARI1, averaging exacerbates the "survey overfitting" as discussed in Section V-C. To conclude, MCD probability vector can assist in identifying and mitigating overfitting over the SAS datasets.

G. Importance of Refined Mappings

Consequently, the uncertainty of the predictions can be estimated reliably enough to compute the uncertainty or disagreement refined mappings similarly to [20] and [24]. This consists in only keeping the predictions of the pixels which value in the confidence map is higher than the percentile value corresponding to a PPC of 8%. Effectively, the PPC diminished all the more so as the extra morphological operations from Section IV-E-b) are applied to stronger penalize boundaries and predictions corresponding to areas smaller than a MCM target.

The results of different approaches to estimate uncertainty, namely MCD, DE and the combination of MCD and DE, and scores, namely Entropy and ASD, to refine mappings are reported in Table V on the Test Set without the outliers. This work also compares them in terms of OACC improvements to the baselines MCD, DE, and MCD DE probability vectors, the deterministic D4SC and its refined predictions based on standard CNN confidence. Notably, models tend to show higher mACC than OACC compared to standard CV applications due to the limited number of classes and the fact that the models are likely to misclassify Flat Sediments, which is the most represented class.

Remarkably, the Entropy and ASD scores yield similar accuracies. As expected by results in the literature such as [91] and the calibration curves of Section V-D, the DE average probability vector of ten models outperforms the 30 repeated inferences of MCD average probability vector for uncertainty

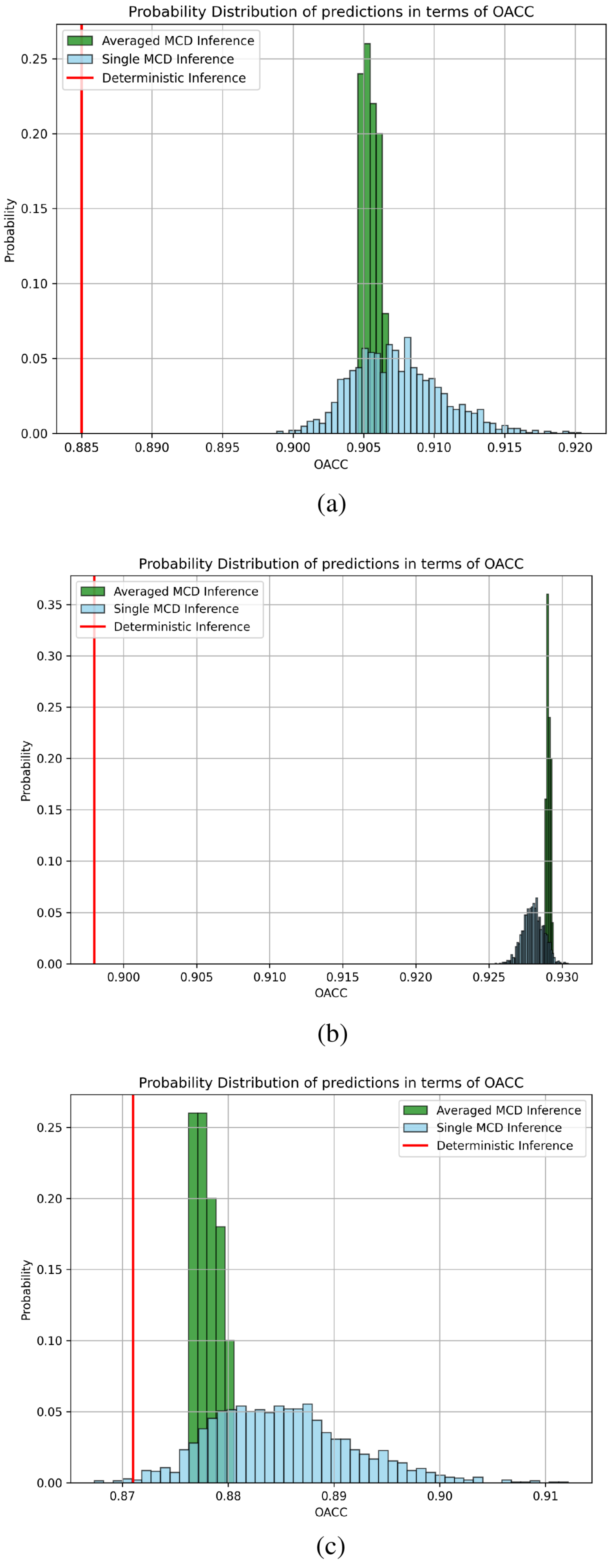


Fig. 7. Comparison of approximate probability density function, in terms of OACC, of both single MCD inference and the average of 30 MCD inferences corresponding to the approximate probability vector respectively evaluated on the complete Test Set, the images from COL2 in the Test Set and the images from ARI1 in the Test Set in (a), (b), and (c). As a reference the OACC of the Deterministic inference is provided for comparison where dropout is not activated at test time. (a) Complete Test Set. (b) COL2. (c) ARI1.

TABLE V

COMPARISON OF DIFFERENT COMBINATIONS OF UNCERTAINTY APPROACHES AND SCORES IN TERM OF OACC AND MACC PERFORMANCE BOOST WHEN REFINING THE PREDICTIONS OF D4SC

Method	Full-size image Test Set		
	OACC	mACC	PPC
Baseline MCD	92.9 %	93.9 %	100 %
Baseline DE	93.4%	95.2 %	100 %
Baseline MCD DE	93.6 %	95.4 %	100 %
Deterministic CNN refined	92.7 %	96.1 %	76.9 %
MCD Entropy refined	94.3 %	96.7 %	75.9 %
MCD ASD refined	94.6 %	97.2 %	72.2 %
DE Entropy refined	95.7 %	98.0 %	75.5 %
MCD DE Entropy refined	**96.2 %**	**98.3 %**	75.8 %

Bold values indicate the best results for the different performance metrics.

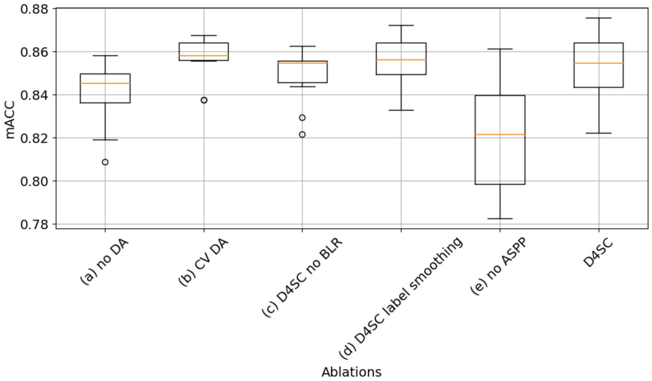


Fig. 8. Box-and-Whisker plot for the ablation experiment over the components of the end-to-end pipeline for 10 repeated trainings in term of mACC over 256-size patch Test Set.

estimation as shown by the refined predictions proxy. For example, the DE confidence map of Fig. 4(e) correctly identifies the Sandwaves where Posidonia starts to grow. In addition, it also underlines the rest of the growing Posidonia pattern which is under represented compared to the grown Posidonia one in the input data distribution. Ultimately, MCD DE outperfoms the DE couterpart at the cost of additional computational time and the necessity of handling larger tensors, similarly to the findings of Sections V-D and V-E. In conclusion, uncertainty estimation effectively improves performances and interpretability of CNN outputs over SAS datasets.

H. Other Ablation Experiments

To ensure the added value of other proposed components of the end-to-end pipeline, ablation experiments are performed and reported in Fig. 8 as Box-and-Whisker plots over the ten repeatedly trained models. Specifically, the performances of D4SC is evaluated with and without Data Augmentation (DA) (a), but also against the common data augmentation pipeline derived from CV for classification (b). Then, BLR is not applied in (c) and compared against label smoothing regularization [92], [93] (d), as another method addressing degenerate labels. Finally, the importance of ASPP is aslo investigated in (e).

Remarkably, removing ASPP in (e) demonstrates its critical role for the reliable convergence of well performing models. Despite the proposed end-to-end pipeline performs best in term

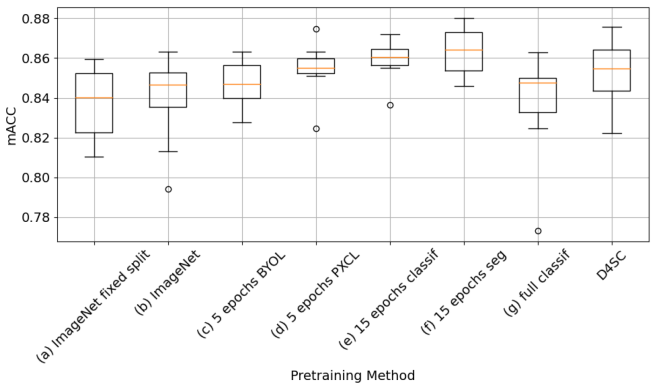


Fig. 9. Box-and-Whisker plot of different pretraining strategies for 10 repeated trainings in term of mACC over 256-size patch Test Set.

of peak mACC according to in Fig. 8, the standard DA derived from CV (b) and label smoothing (c) showcase superior convergence consistency. The improved consistency of the standard DA derived from CV (b) possibly originates from the extra augmentation of patches on the edges of the full-size image.

VI. Discussion

A. Effect of Pretraining

The results presented in Section V underlines a high variability in the mACC across runs of the complete training pipeline, which is a consequence of the unreliability of model convergence. As, this variability is primarily attributed to the degenerate input data distribution, the effect of pretrained feature extractors is investigated in mitigating this variability. Pretraining from ImageNet fixing the weight initialization and dataset splits, leaves only the variability of convergence to random shuffling and dropout. As shown by Fig. 9(a) and the convergence spread of models pretrained with ImageNet weights, their effect is nearly as significant as the complete D4SC model with random initialization. This indicates that the other sources of randomness, such as random dataset splits and weight initialization, have minimal impact.

Then this work explores various pretraining methods, for which results are reported in Fig. 9, including retraining the ImageNet pretrained model (b), and 5 epochs of self-supervised pretraining such as BYOL [34] (c) and [35] noted PXCL (d). In addition, the pretraining of D4SC feature extractor is evaluated with 15 epochs of training using either a classification model (e) or the D4SC model itself (f), both selected out of ten random initializations, and the complete classification training (g). Specifically, the classification pretraining is performed with the standard model from [75], which consists of the feature extractor and a classification head.

Remarkably, this study observes lower performance of the complete classification training from pretrained weights on ImageNet for the seabed patch classification task, where the patch label get assigned the most represented pixel class label. Specifically, a pretrained model on ImageNet achieves up to 90.6% mACC compared to 96.5% mACC for the randomly initialized model retrained from scratch. This is analogous to the findings of [48] which report similar performance gaps in SAR scene classification tasks when comparing ImageNet pretrained models to those trained from scratch. Despite this hindering initialization for classification, the semantic segmentation downstream task over SAS data benefits from the random initialization of the newly created decoder on top of the pretrained feature extractor. This setup provides ample opportunity for the weights to be fine-tuned and hence for the model to be retrained without showcasing significant performance drop in Fig. 9(b).

Moreover, the high variance observed in the convergence of CNN models is mitigated by the improved initialization of the pretrained feature extractor, allowing models to more consistently converge towards a global optimum, as depicted in Fig. 9 sith the other pretraining methods (c), (d), (e), (f), and (g). While (b), (c), and (g) hinders the peak performance compared to the model trained from scratch, alternatively (d), (e), and (f) showcase similar peak performance. Despite their improved reliability of model's convergence, this works refrains from incorporating pretraining methods into uncertainty estimation due to the unknown effect surrounding their impact on model weight sampling.

Additionally and despite applying data-driven pretraining methods, that have been proven successful in different tasks such as in [94], resulting in highly differing CNN initial weight distributions, the different model architectures of Table III and methods of Figs. 8 and 9 still converge to similar mACC values while exhibiting different types of misclassifications. This is another indication that the bottleneck of the end-to-end pipeline lies in the quality of the imperfect data and annotations in the input data distribution, further stressing the importance of uncertainty estimation and refined mappings.

B. Zero-Shot Experiment on MNX18

To further emphasize the utility of the end-to-end pipeline, especially in generating uncertainty refined predictions for OoD data, D4SC is evaluated on the MNX18 survey without retraining. Uncertainty estimation derived from ASD is employed with a hard threshold set to 1 to ensure a PPC superior to 80%, as done in [20]. On average over the MNX18 annotated images, the initial OACC increased from 89.9% to 94.9% with a PPC of 83.5%. The effects of uncertainty estimation with the metrics presented in this work on an OoD survey and derived refined prediction mappings can be observed over Fig 10. It depicts one of the most challenging images of MNX18 with different survey related operational parameters accounting for the normalization issue at short range, which is an OoD pattern. However, despite promising results to account for OoD patterns within individual images, these scores also depend on other intrinsic characteristics such as the amount of boundary pixels and the quality of the SAS imagery. This prevents this work from comparing them between different images and reliably detecting OoD samples as is.

Interestingly, the calibration curves of Fig. 11, associated with the uncertainty estimation of the zero-shot experiment on MNX18, reveal an inverse trend compared to the analysis on the Test Set. Specifically, the quality of uncertainty estimation improves when employing the DE CNN confidence exhibitig

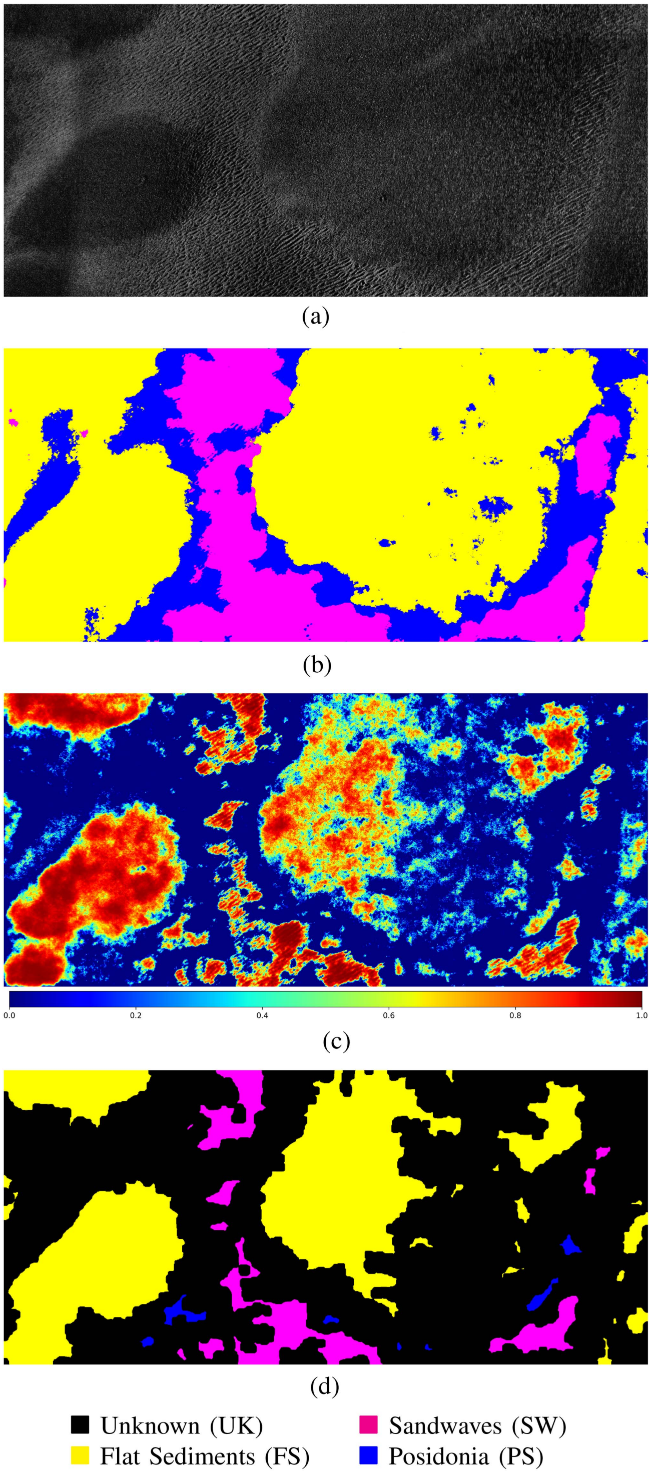


Fig. 10. Challenging OoD SAS image (a) from MNX18 and the prediction mappings obtained with the average MCD probability vector of D4SC (b) associated confidence map derived from the Kullback–Leibler divergence (c) and uncertainty refined predictions (d). The ground range increases from 40 to 150 m within the SAS image from left to right. It depicts mud packs in between fields of sandwaves probably built upon the dead matte of previously alive posidonia. (a) Challenging OoD SAS image from MNX18 (b) Associated MCD averaged predictions (c) Associated MCD ASD confidence (d) Associated MCD ASD refined predictions.

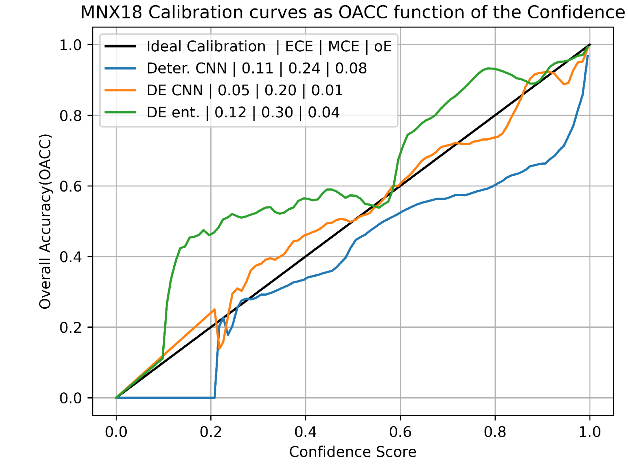


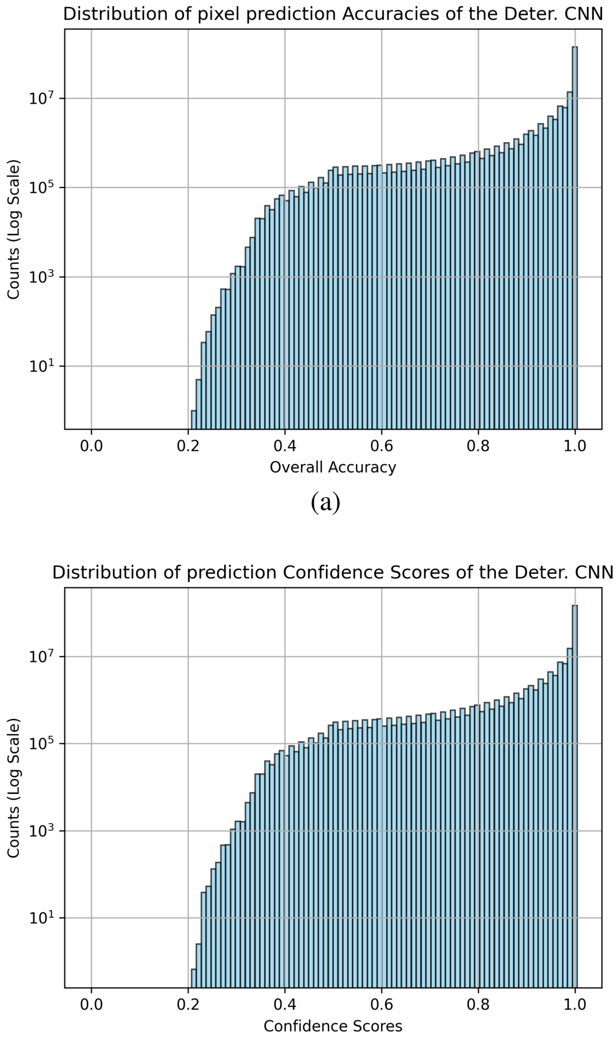
Fig. 11. Calibration curves for models trained on COL2 and ARI1 but evaluated on MNX18. The calibration curve can be understood with respect to the Ideal Calibration; the closer to the Ideal Calibration, the better, and when the curve is higher and lower, the predictions are respectively underconfident and overconfident. Notably, the quality of uncertainty estimation on the Unseen dataset is inversely increased using the DE CNN confidence instead of the entropy, compared to the one from the images from COL2 Test dataset.

almost a perfect fit with a 1% oE compared to the 4% of entropy. This finding suggests that the off-the-shelf CNN confidence, particularly when paired with DE and MCD DE predictions should be favored over entropy for uncertainty estimation on OoD SAS datasets. Despite this, both methods effectively counteract the overconfidence of the trained CNN models, particularly at high confidence and accuracy levels.

VII. Conclusion

The careful design of the end-to-end learning pipeline of D4SC, including innovative features like boundary label rejection, advanced data augmentation and Bayesian uncertainty prediction refinement methods, are beneficial in enhancing the semantic segmentation of the seabed. By incorporating these elements, D4SC is capable of learning generic seabed patterns robust to noise, thus significantly improving the accuracy and reliability of seabed characterization. Performance evaluations against established methods in the literature underscore D4SC's superiority across multiple HR SAS survey datasets obtained from real-world MCM exercises at sea. Furthermore, this study delves into the exploration of epistemic uncertainty within CNNs trained on SAS data, altogether with an in depth analysis of SAS data biases and their impact over the end-to-end pipeline. The robustness of D4SC and its uncertainty estimation are further validated by their excellent performance over unseen data and different operational parameters specific to the survey, both in terms of calibration and refined mappings, demonstrating its ability to generalize effectively across diverse underwater environments at large scale. But as also underlined by this work, automatic seabed characterization is a complicated task considering the quality of the data and the difficulty of annotation. Future work might also want to design specific methods to assess or improve the quality of the data and annotation. In conclusion, the more reliable seabed characterization mappings produced by D4SC offer promising implications not only for MCM, but also for various research fields such as environmental monitoring.

APPENDIX A
EXPLAINING THE PERFORMANCE BOOST OF THE DE RELATED UNCERTAINTY ESTIMATION METHODS



(a)

(b)

Fig. 12. Distribution of OACC (a) and Confidence Score (b) for the computation of the calibration curve of the Deterministic CNN over the Test Set. The distributions are roughly the same for all other calibration curves. Notably, almost all the points fall into the highest accuracies and confidence score bins for the visualization of the calibrations. Thus, instead of just looking at ECE and MCE scores to understand the performance boost of the different uncertainty estimation methods, the upper-right part of the calibration curve should be taken into account. This is where the DE related Uncertainty Estimation methods are a better fit than others. This is why this article introduces oE to take into account the distribution of the confidence and accuracy pairs. In addition, the drastically low number of samples at the low confidence and accuracy pairs accounts for its high distance from the Ideal Calibration. (a) Overall Accuracy. (b) Confidence Score.

Appendix B
The Two Outliers SAS Images in the Test Set

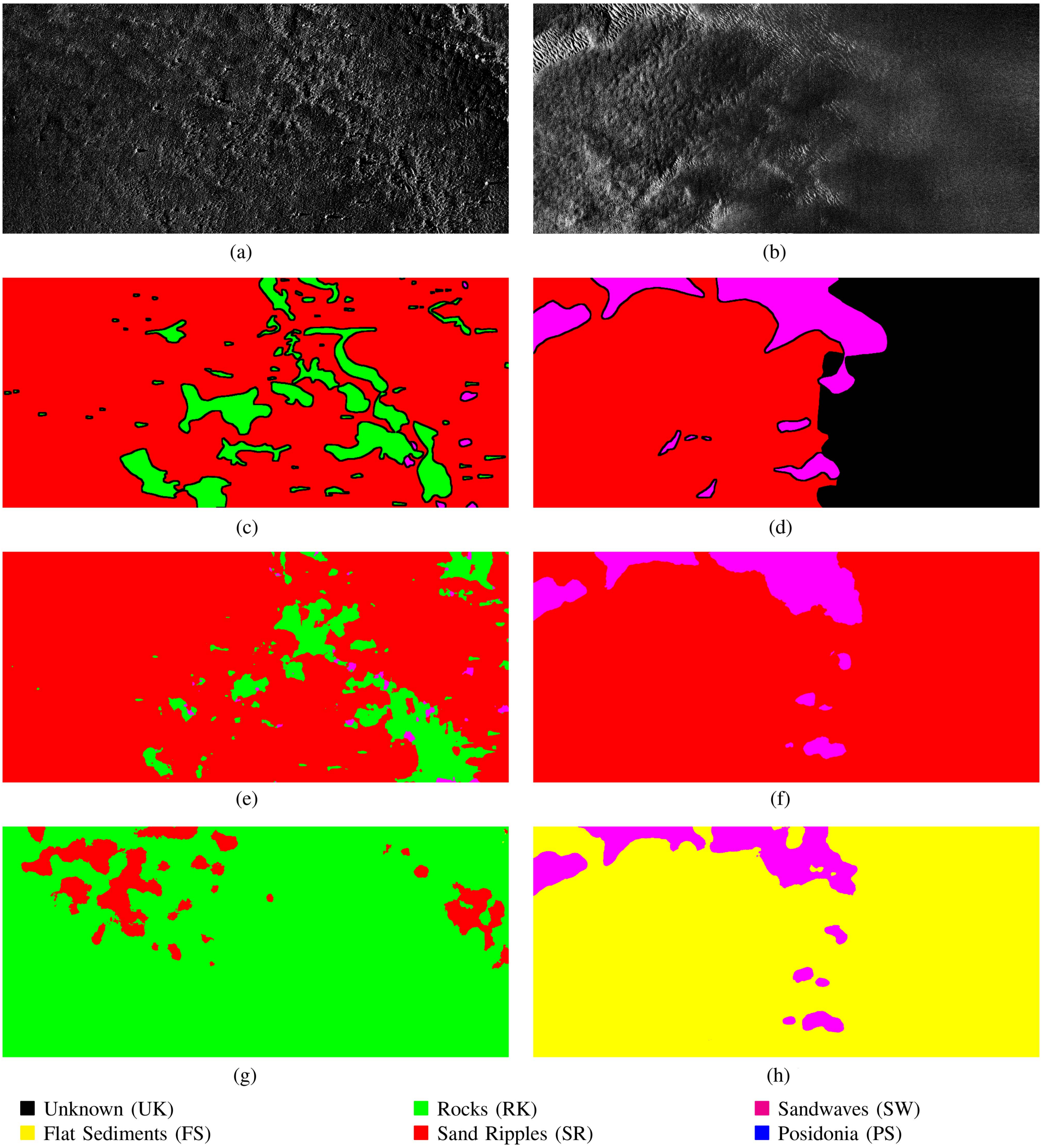


Fig. 13. This figure illustrates the two outlier SAS backscatter Intensity Images from the Test Set, (a) and (b), respectively, from COL2 and ARI1, for which the performances across the family of repeatedly trained models are highly inconsistent. They represent two cases of model overfitting due to noise in the Ground Truth annotations, respectively (c) and (d) for COL2 and ARI1. As (a) depicts a mixed composition of Flat Sediments (FS), Rocks (RK) and Sand Ripples (SR) while COL2 annotations being biased towards Rocks (RK), the best model mostly predicts Sand Ripples (SR) in (e) while the worst one fall into the bias and predicts mostly Rocks (RK) in (g). In contrast, (b) and (d) displaying the only boundary between Sand Ripples (SR) and Sandwaves (SW) in the entire set of annotated images of ARI1, embodies a case of survey overfitting. Indeed, the survey configuration and its suboptimal operational parameters leading to not well autofocus images in ARI1, made it almost impossible for the annotator to recognize Sand Ripples in ARI1, then leading to annotating all training samples as Flat Sediments, which is a distributional shift, but also making it possible for the model to distinguish between the two surveys. In addition, the two surveys having distinct sets of output classes at training, respectively {Rocks (RK), Flat Sediments (FS), Sand Ripples (SR), Sandwaves (SW)} and {Flat Sediments (FS), Posidonia (PS), Sandwaves (SW)} for COL2 and ARI1, the best model generalized well on the dataset in (f) while the worst overfitted the surveys (h) by predicting Flat Sediments (FS). (a) Normalized Backscatter Intensities, COL2 (b) Normalized Backscatter Intensities, ARI1 (c) Ground truth map, COL2 (d) Ground truth map, ARI1 (e) Highest OACC, COL2 (f) Highest OACC, ARI1 (g) Least OACC, COL2 (h) Least OACC, ARI1.

Acknowledgment

ChatGPT 3.5 was used for editing and grammar enhancement.

References

[1] M. Lianantonakis and Y. R. Pétillot, "Sidescan sonar segmentation using texture descriptors and active contours," *IEEE J. Ocean. Eng.*, vol. 32, no. 3, pp. 744–752, Jul. 2007.

[2] D. P. Williams, "Unsupervised seabed segmentation of synthetic aperture sonar imagery via wavelet features and spectral clustering," in *Proc. 16th IEEE Int. Conf. Image Process.*, 2009, pp. 557–560.

[3] D. Williams, "Bayesian data fusion of multiview synthetic aperture sonar imagery for seabed classification," *IEEE Trans. Image Process.*, vol. 18, no. 6, pp. 1239–1254, Jun. 2009.

[4] T. Celik and T. Tjahjadi, "A novel method for sidescan sonar image segmentation," *IEEE J. Ocean. Eng.*, vol. 36, no. 2, pp. 186–194, Apr. 2011.

[5] G. Huo, Q. Li, and Y. Zhou, "Seafloor segmentation using combined texture features of sidescan sonar images," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, 2016, pp. 3794–3799.

[6] M. Pinto, "High resolution seafloor imaging with synthetic aperture sonar," in *Proc. IEEE Ocean. Eng. Soc. Newslett.*. Biloxi: Oceanic Engineering Society of the Institute of Electrical and Electronics Engineers, Inc., 2002, pp. 15–20.

[7] A. Bellettini and M. A. Pinto, "Theoretical accuracy of synthetic aperture sonar micronavigation using a displaced phase-center antenna," *IEEE J. Ocean. Eng.*, vol. 27, no. 4, pp. 780–789, Oct. 2002.

[8] A. Bellettini and M. Pinto, "Design and experimental results of a 300-khz synthetic aperture sonar optimized for shallow-water operations," *IEEE J. Ocean. Eng.*, vol. 34, no. 3, pp. 285–293, Jul. 2009.

[9] D. P. Williams, "Fast target detection in synthetic aperture sonar imagery: A new algorithm and large-scale performance analysis," *IEEE J. Ocean. Eng.*, vol. 40, no. 1, pp. 71–92, Jan. 2015.

[10] I. D. Gerg and V. Monga, "Structural prior driven regularized deep learning for sonar image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.

[11] R. L. Folk, "The distinction between grain size and mineral composition in sedimentary-rock nomenclature," *J. Geol.*, vol. 62, no. 4, pp. 344–359, 1954.

[12] D. P. Williams, "Fast unsupervised seafloor characterization in sonar imagery using lacunarity," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 11, pp. 6022–6034, Nov. 2015.

[13] J. Chen and J. E. Summers, "Deep convolutional neural networks for semi-supervised learning from synthetic aperture sonar (SAS) images," *Proc. Meeting Acoust.*, vol. 30, no. 1, 2017.

[14] X. Du, A. Seethepalli, H. Sun, A. Zare, and J. T. Cobb, "Environmentally-adaptive target recognition for SAS imagery," *Proc. SPIE*, S. S. Bishop and J. C. Isaacs, Eds., vol. 10182, 2017, Art. no. 101820I.

[15] L. Picard, A. Baussard, I. Quidu, and G. Le Chenadec, "Seafloor description in sonar images using the monogenic signal and the intrinsic dimensionality," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5572–5587, Sep. 2018.

[16] J. Peeples, M. Cook, D. Suen, A. Zare, and J. Keller, "Comparison of possibilistic fuzzy local information c-means and possibilistic k-nearest neighbors for synthetic aperture sonar image segmentation," in *Detection and Sensing of Mines, Explosive Objects, and Obscured Targets XXIV*, J. C. Isaacs and S. S. Bishop, Eds., vol. 11012. California, USA: SPIE, 2019, pp. 211–220.

[17] Y. Arhant, O. L. Tellez, X. Neyt, and A. Pižurica, "D4sc: Deep supervised semantic segmentation for seabed characterisation in low-label regime," in *Proc. IGARSS IEEE Int. Geosci. Remote Sens. Symp.*, 2023, pp. 6932–6935.

[18] A. Hartmann, A. Davari, T. Seehaus, M. Braun, A. Maier, and V. Christlein, "Bayesian U-Net for segmenting glaciers in sar imagery," in *Proc. IEEE Int. Geosci., Remote Sens. Symp.*, 2021, pp. 3479–3482.

[19] J. Haas and B. Rabus, "Uncertainty estimation for deep learning-based segmentation of roads in synthetic aperture radar imagery," *Remote Sens.*, vol. 13, no. 8, 2021, Art. no. 1472.

[20] B. Gips, "Texture-based seafloor characterization using Gaussian process classification," *IEEE J. Ocean. Eng.*, vol. 47, no. 4, pp. 1058–1068, Oct. 2022.

[21] H. J. Callow, M. P. Hayes, and P. T. Gough, "Autofocus of stripmap SAS data using the range-variant SPGA algorithm," in *Proc. Oceans Conf. Rec.*, vol. 5, 2003, pp. 2422–2426.

[22] X. Du, A. Zare, and J. T. Cobb, "Possibilistic context identification for SAS imagery," in *SPIE Defense + Security*, S. S. Bishop and J. C. Isaacs, Eds., Baltimore, Maryland, United States: SPIE, 2015, Art. no. 94541I.

[23] J. Peeples, D. Suen, A. Zare, and J. M. Keller, "Possibilistic fuzzy local information C-means with automated feature selection for seafloor segmentation," in *Detection and Sensing of Mines, Explosive Objects, and Obscured Targets XXIII*, J. C. Isaacs and S. S. Bishop, Eds., California, USA: SPIE, 2018, pp. 395–408.

[24] B. Gips, "Bayesian seafloor characterization from sas imagery," in *Proc. UACE2019 Underwater Acoust. Conf. Exhib.*, J. S. Papadakis, Ed., 2019, pp. 235–242.

[25] T. S. Brandes and B. Ballard, "Adaptive seabed characterization with hierarchical Bayesian modeling of SAS imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1278–1290, Mar. 2019.

[26] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[27] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., Red Hook, NY, USA: Curran Associates, Inc., vol. 33, 2020, pp. 1877–1901.

[28] M. Caron et al., "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9650–9660.

[29] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16000–16009.

[30] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.

[31] Z. Liu et al., "Swin transformer v2: Scaling up capacity and resolution," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11999–12009.

[32] P. Goyal et al., "Self-supervised pretraining of visual features in the wild," 2021, *arXiv:2103.01988*.

[33] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," in *Advances in Neural Information Processing Systems*, vol. 33. Red Hook, NY, USA: Curran Associates, Inc., 2020, pp. 22243–22255.

[34] J.-B. Grill et al., "Bootstrap your own latent - A new approach to self-supervised learning," in *Advances in Neural Information Processing Systems*, vol. 33. Red Hook, NY, USA: Curran Associates, Inc., 2020, pp. 21271–21284.

[35] Z. Xie, Y. Lin, Z. Zhang, Y. Cao, S. Lin, and H. Hu, "Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16684–16693.

[36] Y. Gal, R. Islam, and Z. Ghahramani, "Deep Bayesian active learning with image data," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 1183–1192.

[37] W. H. Beluch, T. Genewein, A. Nürnberger, and J. M. Köhler, "The power of ensembles for active learning in image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9368–9377.

[38] V. Rakesh and S. Jain, "Efficacy of Bayesian neural networks in active learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2601–2609.

[39] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 4015–4026.

[40] D. Hong et al., "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sens. Environ.*, vol. 299, 2023, Art. no. 113856.

[41] K. N. Clasen, L. Hackel, T. Burgert, G. Sumbul, B. Demir, and V. Markl, "REBEN: Refined bigearthnet dataset for remote sensing image analysis," 2024, *arXiv:2407.03653*.

[42] C. Li, B. Zhang, D. Hong, J. Yao, and J. Chanussot, "LRR-Net: An interpretable deep unfolding network for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–12, 2023.

[43] X. Li, N. Nadisic, S. Huang, N. Deligiannis, and A. Pižurica, "Model-aware deep learning for the clustering of hyperspectral images with context preservation," in *Proc. 31st Eur. Signal Process. Conf.*, 2023, pp. 885–889.

[44] C. Cui, X. Wang, S. Wang, L. Zhang, and Y. Zhong, "Unrolling nonnegative matrix factorization with group sparsity for blind hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–12, 2023.

[45] C. Li et al., "Casformer: Cascaded transformers for fusion-aware computational hyperspectral imaging," *Inf. Fusion*, vol. 108, 2024, Art. no. 102408.

[46] X. Li, N. Nadisic, S. Huang, and A. Pižurica, "Unfolding admm for enhanced subspace clustering of hyperspectral images," 2024, *arXiv:2404.07112*.

[47] D. Hong et al., "SpectralGPT: Spectral remote sensing foundation model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5227–5244, Aug. 2024.

[48] H. Yang, X. Kang, L. Liu, Y. Liu, and Z. Huang, "SAR-Hub: Pretraining, fine-tuning, and explaining," *Remote Sens.*, vol. 15, no. 23, 2023, Art. no. 5534.

[49] Q. Zhang, X. Cheng, Y. Chen, and Z. Rao, "Quantifying the knowledge in a DNN to explain knowledge distillation for classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 5099–5113, Apr. 2023.

[50] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[51] Y.-C. Sun, I. D. Gerg, and V. Monga, "Iterative, deep synthetic aperture sonar image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

[52] 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, 2022.

[53] I. D. Gerg and V. Monga, "Deep multi-look sequence processing for synthetic aperture sonar image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023.

[54] J. Chen and J. Summers, "Deep neural networks for learning classification features and generative models from synthetic aperture sonar Big Data," *J. Acoustical Soc. Amer.*, vol. 140, 2016, Art. no. 3423.

[55] A. Burguera and F. Bonin-Font, "On-line multi-class segmentation of side-scan sonar imagery using an autonomous underwater vehicle," *J. Mar. Sci. Eng.*, vol. 8, no. 8, 2020, Art. no. 557.

[56] D. P. Williams, "Exploiting phase information in synthetic aperture sonar images for target classification," in *Proc. OCEANS - MTS/IEEE Kobe Techno-Oceans*, 2018, pp. 1–6.

[57] I. Gerg and D. Williams, "Additional representations for improving synthetic aperture sonar classification using convolutional neural networks," 2018, *arXiv:1808.02868*.

[58] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 1321–1330.

[59] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, 2019.

[60] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. 33rd Int. Conf. Mach. Learn.*, 2016, pp. 1050–1059.

[61] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?," in *Proc. Adv. Neural Inf. Process. Syst.*, I. Guyon et al., Eds., 2017, vol. 30. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/2650d6089a6d640c5e85b2b88265dc2b-Paper.pdf

[62] C. Dechesne, P. Lassalle, and S. Lefèvre, "Bayesian U-net: Estimating uncertainty in semantic segmentation of earth observation images," *Remote Sens.*, vol. 13, no. 19, 2021, Art. no. 3836.

[63] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. Adv. Neural Inf. Process. Syst.*, Curran Associates, Inc., vol. 30, 2017.

[64] M. Valdenegro-Toro, "Deep sub-ensembles for fast uncertainty estimation in image classification," 2019, *arXiv:1910.08168*.

[65] R. Rahaman and a. thiery, "Uncertainty quantification and deep ensembles," in *Proc. Adv. Neural Inf. Process. Syst.*, Curran Associates, Inc., vol. 34, 2021, pp. 20063–20075.

[66] H. Ritter, A. Botev, and D. Barber, "A scalable laplace approximation for neural networks," in *Proc. 6th Int. Conf. Learn. Representations*, 2018, vol. 6.

[67] E. Daxberger, A. Kristiadi, A. Immer, R. Eschenhagen, M. Bauer, and P. Hennig, "Laplace redux - effortless Bayesian deep learning," in *Proc. Adv. Neural Inf. Process. Syst.*, Curran Associates, Inc., vol. 34, 2021, pp. 20089–20103.

[68] A. Malinin and M. Gales, "Predictive uncertainty estimation via prior networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Curran Associates, Inc., vol. 31, 2018.

[69] V. Růžička, S. D'Aronco, J. D. Wegner, and K. Schindler, "Deep active learning in remote sensing for data efficient change detection," 2020, *arXiv:2008.11201*.

[70] G. Lenczner, A. Chan-Hon-Tong, B. Le Saux, N. Luminari, and G. Le Besnerais, "DIAL: Deep interactive and active learning for semantic segmentation in remote sensing," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 3376–3389, 2022.

[71] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, 2020.

[72] K. Vinogradova, A. Dibrov, and G. Myers, "Towards interpretable semantic segmentation via gradient-weighted class activation mapping (student abstract)," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 10, 2020, pp. 13943–13944.

[73] Z. Huang, Y. Liu, X. Yao, J. Ren, and J. Han, "Uncertainty exploration: Toward explainable sar target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–14, 2023.

[74] V. Petsiuk et al., "Black-box explanation of object detectors via saliency maps," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11443–11452.

[75] M. Tan and Q. V. Le, "EfficientNet : Rethinking model scaling for convolutional neural networks," 2019, *arXiv:1905.11946*.

[76] J. T. Cobb, "Synthetic aperture sonar seabed environment dataset (SASSED)," *Mendeley Data*, vol. 4, 2022, doi: 10.17632/s5j5gzr2vc.4.

[77] D. P. Williams, "The mondrian detection algorithm for sonar imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 1091–1102, Feb. 2018.

[78] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016, Dec. 2016, pp. 770–778.

[79] X. Lurton et al., "Backscatter measurements by seafloor-mapping sonars," *Guidelines Recommendations*, vol. 200, 2015.

[80] H. Ding, X. Jiang, A. Q. Liu, N. M. Thalmann, and G. Wang, "Boundary-aware feature propagation for scene segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6819–6829.

[81] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 833–851.

[82] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[83] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*. Santiago, Chile, 2015, pp. 1026–1034.

[84] Z. Allen-Zhu, Y. Li, and Z. Song, "A convergence theory for deep learning via over-parameterization," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 242–252.

[85] B. J. Kim, H. Choi, H. Jang, D. Lee, and S. W. Kim, "How to use dropout correctly on residual networks with batch normalization," in *Proc. Uncertainty Artif. Intell.*, 2023, pp. 1058–1067.

[86] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[87] A. McCallum et al., "Employing EM and pool-based active learning for text classification," in *Proc. Int. Conf. Mach. Learn.*, Citeseer, vol. 98, 1998, pp. 350–358.

[88] A. Hein et al., "A comparison of uncertainty quantification methods for active learning in image classification," in *Proc. Int. Joint Conf. Neural Netw.*, 2022, pp. 1–8.

[89] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput., Comput.-Assist. Interv.–MICCAI*, J. W.W.M.F.A. F. Navab Nassirand Hornegger, Ed. Cham: Springer International Publishing, 2015, pp. 234–241.

[90] Y. Arhant, O. L. Tellez, X. Neyt, and A. Pižurica, "Recovering from catastrophic receptive field overflow in semantic segmentation of high resolution images: Application to seabed characterization," in *Proc. IGARSS IEEE Int. Geosci. Remote Sens. Symp.*. Athens, Greece, 2024, pp. 9561–9565.

[91] F. K. Gustafsson, M. Danelljan, and T. B. Schon, "Evaluating scalable Bayesian deep learning methods for robust computer vision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 318–319.

[92] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826.

[93] R. Müller, S. Kornblith, and G. E. Hinton, "WHen does label smoothing help?," in *Proc. Adv. Neural Inf. Process. Syst.*, Curran Associates, Inc., vol. 32, 2019.

[94] J. Li, X. Li, and Y. Yan, "Unlocking the potential of data augmentation in contrastive learning for hyperspectral image classification," *Remote Sens.*, vol. 15, no. 12, 2023, Art. no. 3123.

Yoann Arhant (Graduate Student Member, IEEE) received the Eng. degree in photonics from l'Institut d'Optique, Palaiseau, France, and the M.Sc. degree in computer science from Université de Bordeaux, Talence, France, in 2019. He is currently working toward the Ph.D. degree in computer science engineering with Royal Military Academy, Brussels, Belgium, in collaboration with Ghent University, Ghent, Belgium, since 2021.

His research interests include data science and artificial intelligence applications for signal and image processing.

Olga Lopera Tellez received the Eng. degree in electrical engineering from the University of Los Andes (UA), Bogota, Colombia, in 2001, and the M.Sc. degree in applied sciences and the Ph.D. degree in engineering sciences from the Université Catholique de Louvain (UCL), Louvain-la-Neuve, Belgium, in collaboration with the Royal Military Academy (RMA), Brussels, Belgium, and UA, in 2004 and 2008, respectively.

In 2007, she became a Researcher with the RMA, where she has been a Senior Researcher and Project Coordinator with the Remote Sensing unit since 2021. Her main scientific interests include target detection and target identification using acoustic data for mine countermeasures applications, UXO detection, and seabed monitoring.

Xavier Neyt received the master's degree in engineering (summa cum laude) from the Université Libre de Bruxelles (ULB), Brussels, Belgium, in 1994, the postgraduate degree in signal processing (summa cum laude) from the Université de Liège (ULg), Liège, Belgium, in 2004, and the Ph.D. degree in applied science from the Royal Military Academy, Brussels, Belgium, and ULg, in 2008.

He has been working as a Research Engineer with the Royal Military Academy. In 1996–1997, he was a Visiting Scientist with the French Aerospace Center, Palaiseau, France, and in 1999, with the German Aerospace Centre, Cologne, Germany. In 1997–1999, he was responsible for the design of the image compression module of the European MSG satellite, and in 2000–2007, responsible for the redesign of the ground processing of the scatterometer of the European ERS satellite following its gyroscope anomaly. Since 2008, he has been leading the Scatterometer Engineering Support Laboratory, Royal Military Academy for the European Space Agency. He is an Associate Professor with the Department of Communication, Information, Systems, and Sensors, Royal Military Academy. His research interests include signal processing, radar remote sensing, array processing, bistatic radars, and image processing.

Dr. Neyt was the recipient of the Frerichs Award from the ULB and the special IBM grant from the Belgian National Fund for Scientific Research, in 1995.

Aleksandra Pižurica (Senior Member, IEEE) received the Diploma degree in electrical engineering from the University of Novi Sad, Novi Sad, Serbia, in 1994, the M.Sc. degree in telecommunications from the University of Belgrade, Belgrade, Serbia, in 1997, and the Ph.D. degree in engineering from Ghent University, Ghent, Belgium, in 2002.

She is currently a Professor of statistical image modeling with Ghent University. Her research interests include the area of signal and image processing and machine learning, including multiresolution statistical image models, Markov random field models, sparse coding, representation learning, and image and video reconstruction, restoration, and analysis.

Dr. Pižurica was the recipient of the Scientific Prize "de Boelpaepe" from the Royal Academy of Science, Letters and Fine Arts of Belgium for her contributions to statistical image modeling and applications to digital painting analysis, in 2015, and other recognitions for her work, among which as a co-recipient of the David Hestenes Prize from AGACSE 2018 and the Best Paper Award of the IEEE Geoscience and Remote Sensing Society (GRSS) Data Fusion contest, in 2013 and 2014. She is a Member of the EURASIP Technical Area Committee Signal and Data Analytics for Machine Learning. She served as the TPC Co-Chair for the 30th EUSIPCO Conference (in 2022), an Europe Liaison for IEEE ICIP 2020 and ICIP 2024, the Plenary Co-Chair for EUSIPCO 2024, and elected as the TPC Co-Chair for IEEE ICIP 2026. She served as an Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING from 2012 to 2016 and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY from 2016 to 2019. She is a Senior Area Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING from 2016 to 2019 and since 2022.