# Deep Semi-supervised Semantic Segmentation in Multi-frequency Echosounder Data

Changkyu Choi,
*Member, IEEE*, Michael Kampffmeyer,
Nils Olav Handegard,
*Member, IEEE*, Arnt-Børre Salberg,
and *Senior Member, IEEE*, Robert Jenssen

**Abstract**

Multi-frequency echosounder data can provide a broad understanding of the underwater environment in a non-invasive manner. The analysis of echosounder data is, hence, a topic of great importance for the marine ecosystem. Semantic segmentation, a deep learning based analysis method predicting the class attribute of each acoustic intensity, has recently been in the spotlight of the fisheries and aquatic industry since its result can be used to estimate the abundance of the marine organisms. However, a fundamental problem with current methods is the massive reliance on the availability of large amounts of annotated training data, which can only be acquired through expensive handcrafted annotation processes, making such approaches unrealistic in practice. As a solution to this challenge, we propose a novel approach, where we leverage a small amount of annotated data (supervised deep learning) and a large amount of readily available unannotated data (unsupervised learning), yielding a new data-efficient and accurate *semi-supervised* semantic segmentation method, all embodied into a single end-to-end trainable convolutional neural networks architecture. Our method is evaluated on representative data from a sandeel survey in the North Sea conducted by the Norwegian Institute of Marine Research. The

Changkyu Choi *(changkyu.choi@uit.no)*, Michael Kampffmeyer *(michael.c.kampffmeyer@uit.no)*, and Robert Jenssen *(robert.jenssen@uit.no)* are with UiT The Arctic University of Norway, Tromsø, Norway.
Nils Olav Handegard *(nilsolav@hi.no)* is with Norwegian Institute of Marine Research, Bergen, Norway.
Arnt-Børre Salberg *((salberg@nr.no))* is with Norwegian Computing Center, Oslo, Norway.
Corresponding author: Changkyu Choi

rigorous experiments validate that our method achieves comparable results utilizing only 40 percent of the annotated data on which the supervised method is trained, by leveraging unannotated data. The code is available at *https://github.com/SFI-Visual-Intelligence/PredKlus-semisup-segmentation*.

**Index Terms**

Multi-frequency echosounder data, marine acoustics, acoustic target classification, deep learning, semi-supervised semantic segmentation, convolutional neural networks, deep clustering.

## I. INTRODUCTION

Semantic segmentation is one of the fundamental computer vision tasks, where the aim is to assign each image pixel to a semantic class [1]–[3]. When analyzing echosounder data, the aim is to assign an observed acoustic backscattering intensity to one of several given acoustic classes, often referred to as acoustic target classification [4]–[7]. In practice, semantic segmentation of the echosounder data is still a manual and heuristic process, which is rather vulnerable to human error and bias. It is also expensive in terms of cost and time [8].

There are a few studies that intend to automate the semantic segmentation based on statistical modeling and machine learning techniques [9]–[13]. However, they are exposed to limitations such as relying heavily on handcrafted feature selection and not being able to scale well to large amounts of data. As recent echosounder technology leverages increasing numbers of frequency channels and wider bandwidth [14], automated analysis methods should therefore be scalable in order to cope with increased resolution and multi-frequency data.

Convolutional Neural Networks (CNN) is a framework renowned for excelling at image segmentation tasks [15]. Recent echosounder segmentation studies introduce CNN-based segmentation methods as alternative strategies [5], [16]–[19], where the main advantage is the capacity to learn discriminating features from the training data without requiring a handcrafted process, allowing the analysis to scale to large-sized data. Note that these methods are trained in a fully supervised manner, indicating that the network learns from fully annotated training data. The fully supervised approaches achieve good performance provided that high quality training data and an appropriate choice for the prediction model are assured. However, it is highly challenging for the echosounder data to obtain the class annotation for each backscattering intensity pixel because this relies on the manual annotation process, which is expensive and error-prone.

Hence, a new learning scheme is required to considerably reduce the dependency on the manual annotation process while still facilitating powerful deep learning approaches for the segmentation of the echosounder data. As a key step in this direction, we propose a novel deep *semi-supervised* semantic segmentation method that efficiently uses a small amount of manually annotated data by combining it with a large amount of readily available unannotated data in the learning process [20]–[22].

The key concept invoked in order to train the semi-supervised segmentation network is to alternate between two objective functions, namely an unsupervised clustering objective and a supervised segmentation objective, encapsulated by a single CNN. The unsupervised clustering objective is to search the underlying structure within the training data without using the class annotation. In contrast, the supervised segmentation objective is to map the input echosounder data to the given classes presented in the available annotated data. These two objective functions alternatively optimize the single CNN and gradually integrate the underlying clustering structure to the class decision boundaries presented in the small amount of annotated training data. Our proposed method can create pixel-level prediction maps using the same CNN architecture as [5], [23]. Still, it is data-efficient because it can significantly reduce the use of the annotated data. To the best of our knowledge, our work is the first semi-supervised semantic segmentation method for multi-frequency echosounder data that provides prediction maps on a pixel scale, advancing the existing semi-supervised method of providing patch-scale prediction maps (see Section III-C) [22]. In addition, our proposed method is end-to-end trainable, which refers to a holistic gradient-based learning system where a formulated objective function reflects the principle of a given task without requiring extensive human intervention and prior knowledge [24].

Extensive and rigorous experiments are conducted on the multi-frequency echosounder data collected at the North Sea by the Norwegian Institute of Marine Research. A severe class imbalance in the echosounder data is an ever-present source of bias that prevents training of the neural networks, where 99 percent of the entire acoustic backscattering intensities is occupied by the background class [5], [25]. We introduce a class-rebalancing weight to each learning objective to mitigate the bias, where the weight is calculated with respect to the model prediction without relying on the annotation.

The contributions of the paper are the following:

- To propose a novel deep semi-supervised semantic segmentation method for the multi-frequency echosounder data, which considerably advances the existing methods.

- To achieve comparable results with the fully supervised segmentation method by leveraging a small amount of the annotated data in addition to unannotated data.
- To exploit the underlying structure of the training data using unsupervised deep clustering in a semi-supervised learning manner.
- To demonstrate the innovation potential of the proposed method in a real-world test case.
- To regulate the class imbalance based on the model prediction without leveraging the annotated part of data.
- To operate in an end-to-end and mini-batch training scheme.

## II. Background

Semantic segmentation is the process of partitioning an image into mutually exclusive subsets by assigning a class annotation to each intensity of the data, in which each subset represents a meaningful region of the original image [26]. It thereby provides a comprehensive scene description that includes object class, location, and shape. A wide range of real-world problems require semantic segmentation [27]–[32], such as self-driving vehicles [33], and polyp detection [34], [35], to name a few, all depending on different types of image data.

Semantic segmentation has been considered as a challenging computer vision task due to the large distribution variance as well as the huge class imbalance among objects in the input data [25]. In recent years, however, deep learning has been rapidly advancing and has become a game-changer in many image analysis tasks including semantic segmentation. The CNN [36] is a deep learning framework that has had particular success for grid-structured data such as images. Traditional CNNs consist of convolutional layers and pooling layers, where these layers are stacked in a deep and hierarchical architecture in a particular order, providing unique properties to the analysis. For example, the weight-sharing property of the convolutional filters provides a symmetric transformation between the input space and the output space, referred to as 'equivariance to translation'. The pooling layers help the learned representation becoming approximately invariant to small translations of the input [15], [37]. Another advantage of the CNN is a relatively more straightforward learning process than the conventional methods, where the CNN-based models learn by minimizing a formulated objective function that reflects the strategies of a given task without requiring extensive human intervention and prior knowledge, referred to as an end-to-end learning.

CNN-based segmentation models are distinguishable through their model architecture. Their architecture consists of a downstream module that extracts the abstracted feature representations of the input data and an upstream module that reconstructs the prediction map exhibiting the class attributes of each intensity in the input data based on these extracted feature representations. Thanks to the dual architecture, those models can make class predictions on arbitrary-sized inputs [38]. Fully Convolutional Networks [1] and U-Net [23] are representative architectures, where the models are composed of (transposed) convolutional layers and pooling layers, and end-to-end trainable depending on their formulation of the objective functions.

### *A. Echosounder Data*

For the sustainable management of commercially harvested marine organisms, reliable information on their abundance is essential. For example, lesser sandeel, a species of fish of interest in this study, is the primary food source in the North Sea food web thanks to its ample population [39], which are the preferred prey of a variety of predators, including marine mammals, seabirds, and piscivorous fishes [40]. Therefore, monitoring sandeel stock is critical for the sustainability of the marine ecosystem and fishery management in the North Sea. The echosounder data can contribute to estimating the abundance, leveraging the characteristics of the backscattered responses and knowledge of the target species [8]. The multi-frequency echosounder data that we use in this study has been collected by multi-frequency Simrad EK60 echosounder systems operating at four different frequency channels on the vessel (18, 38, 120, 200 kHz), where the vessel speed is approximately ten knots. The Norwegian Institute of Marine Research has collected the data through the annual trawl surveys in the sandeel areas in the North Sea [41].

We leverage the data preprocessing protocol from the earlier works [5], [22], for which we share the echosounder data. For each frequency channel, a volume backscattering coefficient $s_v$, an average amount of backscattering intensity per cubic metre [42], is stored in the two-dimensional echosounder data. In the physical context, the horizontal and vertical lengths of a single backscattering coefficient are, respectively, one second and 19.2 centimeters based on the pulse duration of 1.024 milliseconds with respect to a common time-range grid based on the resolution of the 200 kHz echosounder data. All the volume backscattering coefficients $s_v$ are first converted to a decibel unit (dB re $1\text{m}^{-1}$). We set the minimum value as -75 dB re $1\text{m}^{-1}$. The coefficients less than -75 dB re $1\text{m}^{-1}$ or missing coefficients are imputed to the minimum values.

For segmentation of the echousounder data, one common approach is a manual annotation method, which relies on the operators' domain expertise of the acoustic properties, such as relative frequency response [43], [44], echo traces [45], and trawl sampling [46]. For that reason, the manual process is vulnerable to bias from the operators. In extreme cases, the systematic error associated with the manual method can be as high as $\pm$ 80 percent [8]. Hence, more structured and automated approaches are required to apply consistent criteria to the analysis while reducing dependency on human intervention. To this end, post-processing systems, including the Large Scale Survey System (LSSS) [9], are developed to facilitate the manual process. The systems support thresholding, error-checking, noise removal, and manipulation of the echosounder data. By adjusting the threshold of backscattered intensities, the post-processing systems visualize the corresponding morphology of the fish schools to enable the operators to detect and delineate the most plausible morphology. In addition, these post-processing systems enable relatively consistent criteria for the analysis by leveraging their acoustic feature libraries. The library consists of a selected part of the backscattered responses and their manually annotated class attributes. By comparing the statistical properties of the collected data to the feature library, the post-processing system predicts the class attribute of the fish school, where the prediction is verified by the scattering model for the corresponding marine organism if available [47], [48].

The sandeel data in this study is manually annotated with the aid of LSSS, where expert operators determine the class of each backscattering coefficient as sandeel (SE), other fish species (OT), or background (BG) class. The primary frequency for LSSS is chosen to 200 kHz considering the highest sandeel signal-to-noise ratio [49]. The operators alter the detection threshold centered at -63 dB at the primary frequency to discover the fish school boundaries visually. The delineated boundary is refined using binary morphological closing to have smoother and pragmatic edges [5]. However, the final decision for both morphology and species is still a manual process, which is time-consuming and requires tacit knowledge that can be potentially biased as with any expert system.

Therefore, recent studies have focused on the automated identification of the fish species using machine-learning methods while leveraging the conventional detection algorithm to detect and delineate the morphology of the schools. SHAPES (Shoal Analysis and Patch Estimation System) [50], [51] is often chosen for the fish school detection algorithm, which extracts a feature vector from each fish school leveraging a single frequency channel of 38 kHz. A random forest based classifier [12] is introduced to classify feature vectors of silver cyprinid from the other

species in Lake Victoria. [52] proposes a classifier leveraging a shallow feedforward network and classify the pelagic Mediterranean fish schools such as anchovy, sardine, and horse mackerel. Those studies show that the automated identification can save time and cost while also achieving robust performance. However, they have limitations in generalizability and scalability because the SHAPES algorithm only exploits a single channel of the echosounder data, and a handcrafted feature selection is required to improve the performance.

Deep learning based models generalize and scale well on various types of data using their flexibility [15], [37]. Among them, the fully supervised deep learning approaches, approaches that learn from the fully annotated training data, achieve a good level of performance provided a high quality of the training data and an appropriate choice of the prediction model are assured. In order to take advantage of supervised deep learning in the analysis of echosounder data, CNN-based semantic segmentation model [5] is introduced to segment the schools of lesser sandeel from the other species leveraging the U-Net architecture [23]. Without relying on the deterministic school detection algorithms and the feature vectors as input, the model constructs the prediction map directly from the input echosounder data.

### *B. Deep Clustering*

We here discuss *deep clustering* since our novel CNN-based semi-supervised semantic segmentation for echosounder data, presented in Section III, relies heavily on this concept. Deep clustering refers to unsupervised deep learning based approaches, that aim to cluster data into underlying groups without requiring the class attributes of the data [53]. Deep clustering leverages the representation power of the neural network in conjunction with clustering algorithms, and partitions the input data into clusters with respect to the learned representation. As clustering performance heavily depends on the underlying structure of the data, deep clustering leverages the neural network to encode the training images in the feature representations where the clustering task becomes much easier [54].

Our proposed method is inspired by a well-known deep clustering framework, referred to as DeepCluster [53], which explicitly models the density of datapoints leveraging the $k$-means clustering algorithm. For a given image dataset, the $k$-means algorithm partitions the feature representation into $K$ different densities, where each density refers to an image descriptor or a visual feature. This has the advantage that it is easy to increase the capacity of more visual features by simply increasing the number of clusters $K$, leading to all-purpose visual features. The

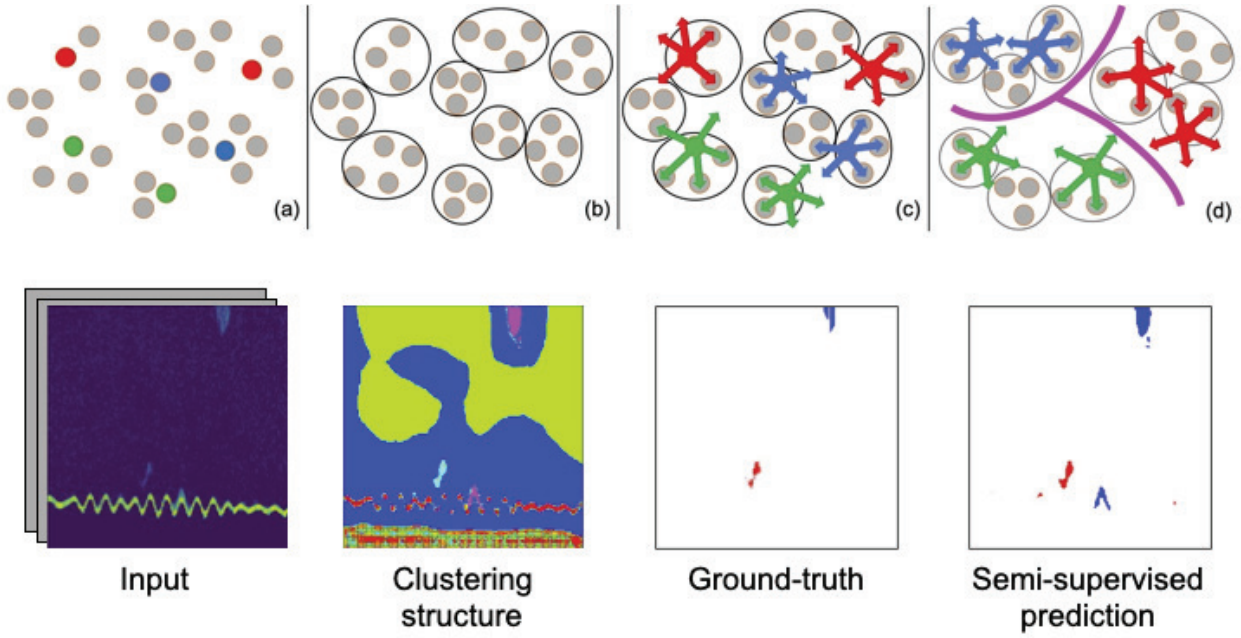


Fig. 1. Overview of the proposed method. Each backscattering intensity in the input is mapped into an arbitrary space shown in (a). The point in gray is unannotated while the point in color (red, green, or blue) indicates the annotated one with respect to the class. (b) shows the clustering structure incorporated by the unsupervised clustering objective without leveraging the annotation. The clustering structure becomes the pseudo-label to train the model in an unsupervised manner. (c) indicates that the annotated data (ground-truth where available) and the supervised segmentation objective optimize the CNN in a supervised manner. (d) indicates that the iteration of (b) and (c) constructs the decision boundary with respect to given classes, where the unannotated points take their place inside the boundary according to their own clusters.

neural network produces cluster indices that can be thought of as clustering-induced annotations for the training data. The network is then updated in a supervised manner to learn the clustering structure. This annotation technique is referred to as pseudo-labeling, allowing the supervised deep learning approach to be applied to unannotated training data [55].

## III. Proposed Method

In this paper, we propose a novel semi-supervised semantic segmentation method, PredKlus, that enables a CNN to simultaneously learn from large amounts of unannotated data and a few annotated data, all in the same network.

The major novelty of our work is the methodology of how the network learns in a semi-supervised manner, illustrated in Figure 1. Our proposed segmentation network operates for two different goals: a) searching for the internal structure of the training data without relying on

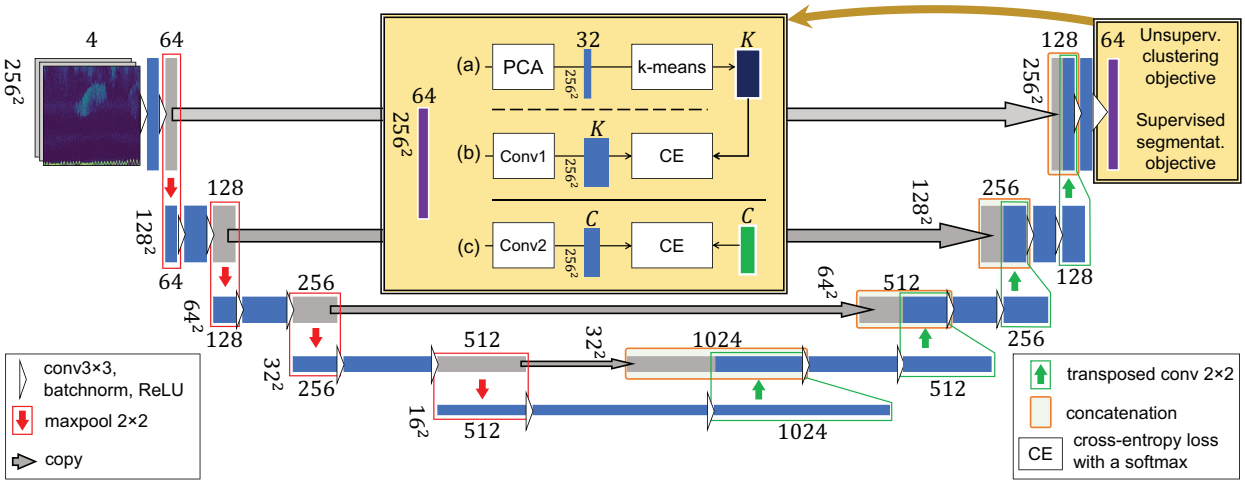


Fig. 2. Proposed model architecture. The application of the two objective functions takes place at the yellow box at the end of the decoder. The unsupervised clustering objective involves in the first two steps: (a) creating pseudo-label using $k$-means, and (b) updating model to learn the clustering structure with the pseudo-label using *conv1*. The supervised segmentation objective involves in (c) training with the partially available annotation using *conv2*. The rectangular bars in blue or gray represent feature maps, where the size of each feature map is specified around it, e.g. $256^2$ or $16^2$. We omit to specify the depth for a few feature maps, as the depth is the same as the feature map on its right, e.g. $64$ or $512$.

external information, e.g., ground truth; b) mapping input echosounder data to given classes. The former goal can be achieved by an *unsupervised clustering objective*, which clusters every pixels in the input based on their features to reveal a clustering structure of the input data in an unsupervised manner. Figure 1(b) illustrates the clustering structure. A *supervised segmentation objective*, on the other hand, aims to map the input to given classes by leveraging the annotated part of training data, albeit in a small amount. Figure 1(c) illustrates this. As these two objective functions alternately optimize the network using gradient descent, the segmentation network gradually learns the class decision boundaries (supervised) with respect to the clustering structure (unsupervised), as illustrated in Figure 1(d). We implement the entire learning process in an end-to-end manner and a mini-batch setting, which are additional novelties of our method.

### A. Model Architecture

Figure 2 describes the model architecture of our proposed method. The encoder-decoder architecture with the skip connections is inspired by U-Net [23] and the recent segmentation study of the echosounder data [5]. The encoder part extracts the abstracted feature map of the echosounder input with a shape of $256 \times 256 \times 4$ over five stages, where the area of the feature

map is reduced to one-fourth at each stage due to a $2 \times 2$ max-pooling layer. By processing two sets of a $3 \times 3$ convolutional layer, a batch-normalization layer [56], and a Rectified Linear Unit (ReLU) [57] at each stage, we abstract the feature map by doubling the depth. The encoder eventually creates five feature maps of different area sizes and depths, where the shape of the last feature map is $16 \times 16 \times 1024$.

The decoder part reconstructs the prediction map leveraging five feature maps from the encoder. At each stage, a $2 \times 2$ transposed convolutional layer and the concatenation of the feature maps along the depth axis play an important role. The $2 \times 2$ transposed convolutional layer increases the area of the feature map fourfold while halving the depth. The halved feature map is concatenated with the feature map in the same shape from the encoder. The concatenated feature map is processed by two sets of a $3 \times 3$ convolutional layer, a batch-normalization layer, and a ReLU, where the depth becomes halved.

The novelty in our architecture is to introduce a convolutional layer for each objective function at the end of the CNN to employ two objective functions in one network. The alternation of the two objective functions takes place at the end of the decoder, where the decoder reconstructs the feature map with a shape of $256 \times 256 \times 64$. To alternately leverage two objective functions, we append a $1 \times 1$ convolutional layer at the end of the network for each objective function, namely *conv1* for the unsupervised clustering objective and *conv2* for the supervised segmentation objective. Note that the number of filters in *conv1* matches the number of clusters or pseudo-classes $K$. Similarly, the number of filters in *conv2* is equal to the number of classes $C$.

### *B. Two Objective Functions*

Our proposed method leverages two objective functions, where those objectives alternately optimize the model. Through the alternating optimization, the CNN indirectly incorporates the class annotations (supervised) to a structured representation (unsupervised) and eventually discovers a structured representation consistent with the available annotations. The yellow box in the middle of Figure 2 shows the overview of our semi-supervised segmentation method. The first two steps, (a) creating pseudo-labels using $k$-means and (b) updating the model to learn the clustering structure with the pseudo-labels using *conv1*, contribute to learning the structured representation in an unsupervised manner. The next step, (c) training with the partially available annotation using *conv2*, represents how the CNN learns in a supervised manner using

the supervised segmentation objective and the available class annotations. Note that a cross-entropy loss (CE) is leveraged to update the model, as depicted in Figure 2(b) and (c).

*a) Unsupervised clustering objective:* The unsupervised clustering objective exploits the underlying structure of the data using the unsupervised clustering algorithm, such as $k$-means, to create pseudo-labels with respect to the clustering structure [53]. Defining the number of clusters $K$ beforehand, the proposed model partitions the feature map $\mathcal{Z} = \left\{\mathbf{z}^{(i)}\right\}_{i=1}^{N}$ located at the end of the decoder into $K$ clusters in a way to find the best assignment by minimizing the $k$-means loss:

$$\mathcal{L}_{kmns} = \frac{1}{N}\sum_{i=1}^{N}\min_{\mathbf{c}_k} d(\mathbf{z}_{PC}^{(i)}, \mathbf{c}_k). \tag{1}$$

In this expression, $N$ is the number of feature vectors in a mini-batch of the feature map. If the batch size $B_s$ is equal to one, $N$ becomes 65,536 as each feature map consists of 65,536 vectors (256 $\times$ 256). The function $d(\cdot,\cdot)$ measures the $L_2$ distance between two vectors, where $\mathbf{c}_k \in \mathbb{R}^{32}$ is the centroid of cluster $k$, and $\mathbf{z}_{PC}^{(i)} \in \mathbb{R}^{32}$ is the dimensionality-reduced training set consisting of the feature vectors $\mathbf{z}^{(i)} \in \mathbb{R}^{64}$. For dimensionality reduction, we use Principal Component Analysis (PCA) [58], which computes the principal components and use only the first few principal components corresponding to the largest eigenvalues for manageable computational complexity.

The clustering result creates the pseudo-labels, having $K$ different pseudo-class attributes according to the $K$ cluster indices. The CNN learns the structured representation from the pseudo-labels using the cross-entropy loss. The unsupervised clustering objective is depicted as:

$$\mathcal{L}_{cls} = \frac{1}{N}\sum_{i=1}^{N} w_{cls,k}^{(i)} CE\{g_\theta(\mathbf{z}^{(i)}), \hat{\mathbf{y}}^{(i)}\}, \tag{2}$$

where $CE(p,q) = -\sum_k q_k \log(p_k)$ is the cross-entropy loss of the probability distribution $p$ for the one-hot encoded label $q$, $\hat{\mathbf{y}}^{(i)} \in \{0,1\}^K$ is the one-hot encoded pseudo-label, and $g_\theta(\mathbf{z}^{(i)})$ is a probability distribution of the output from the CNN, where *conv1* is appended at the end of the decoder. The scalar $w_{cls,k}^{(i)}$ indicates the class-rebalancing weight to penalize the class imbalance of the pseudo-labels. How to obtain this scalar will be explained in III-C. Once updating the CNN with the unsupervised clustering objective, we assign the current centroids of $K$ clusters to the initial centroids for the next clustering to provide consistency of the pseudo-labels over the mini-batches.

*b) Supervised segmentation objective:* To enforce consistency of predictions with regard to the given classes, we train the CNN using the partially available annotated data. The supervised segmentation objective is involved here, where *conv2* layer, another $1 \times 1$ convolutional layer, replaces the *conv1* layer to allow end-to-end training. The supervised segmentation objective is depicted as:

$$\mathcal{L}_{seg} = \frac{1}{N} \sum_{i=1}^{N} w_{seg,c}^{(i)} CE\{f_\theta(\mathbf{x}^{(i)}), \mathbf{y}^{(i)}\}, \tag{3}$$

where $C$ represents the number of given classes, $\mathbf{y}^{(i)} \in \{0,1\}^C$ represents the one-hot encoded vector of the available annotation. $f_\theta(\mathbf{x}^{(i)})$ a probability distribution of the output from the CNN, where *conv2* replaces *conv1*.

*c) Training procedure:* In addition to end-to-end learning, the proposed method operates in a mini-batch training manner, indicating that the network is updated once after each objective processes information in each mini-batch [59]. We form two training subsets for each objective function to facilitate the alternating mini-batch training. The training subset for the unsupervised clustering objective consists of the entire training input data, whether annotated or not, and does not include any class annotation of the data. On the other hand, the training subset for the supervised segmentation objective includes the annotated part of the training data, which takes a small amount of the entire training data in the semi-supervised learning scheme. Algorithm 1 illustrates the semi-supervised training procedure with two training subsets.

### *C. Advance on the Semi-Supervised Image Classification for Echosounder Data [22]*

The problem of being able to obtain manual annotations is much more severe for semantic segmentation compared to image classification, since in the former case annotations refer to the pixel level and not the entire image. The semi-supervised method we propose in this paper therefore solves a much more challenging problem compared to our previous preliminary work on semi-supervised echosounder data patch classification [22], which is only able to classify whole image patches and not do proper segmentation. Some elements of the new segmentation method resembles the previous classification method, however with significant differences due to the completely different aims of the two approaches. For the benefit of the reader, and since we use [22] as one of the comparison models in experiments (referred to as SemiClf, Section IV-D), we will elaborate on these differences in this section.

**Algorithm 1** Training by alternating two objectives

**Input:**

$\mathcal{X}$: training input data

$\mathcal{X}^A \subset \mathcal{X}$: the annotated part of the training input data

$\mathcal{Y}^A$: class annotation of $\mathcal{X}^A$

$\mathbf{X}$: an unannotated mini-batch of $\mathcal{X}$

$(\mathbf{X}^A, \mathbf{Y}^A)$: an annotated mini-batch of $\mathcal{X}^A$ and $\mathcal{Y}^A$

**Output:**

$\mathbf{Z}$: feature map of the mini-batch $\mathbf{X}$ at the end of the decoder

$\hat{\mathbf{Y}}$: created pseudo-label of the mini-batch $\mathbf{X}$

$\mathbf{P}^A$: class prediction of the mini-batch $\mathbf{X}^A$

**Procedure:**

**for** $(\mathbf{X}, \mathbf{X}^A, \mathbf{Y}^A) \in (\mathcal{X}, \mathcal{X}^A, \mathcal{Y}^A)$ **do**

– Compute $\mathbf{Z}$ by processing $\mathbf{X}$ through the model

– Create pseudo-label $\hat{\mathbf{Y}}$ by clustering the principal components of $\mathbf{Z}$

– Compute $w_{cls}$ with respect to $\hat{\mathbf{Y}}$

– Append *conv1* at the end of the decoder

– Update the model end-to-end using ($\mathbf{X}$, $\hat{\mathbf{Y}}$) and the unsupervised clustering objective in Eq.(2) with gradient descent

– Replace *conv1* by *conv2*

– Compute $\mathbf{P}^A$ by processing $\mathbf{X}^A$ through the model

– Compute $w_{seg}$ with respect to $\mathbf{P}^A$

– Update the model end-to-end using ($\mathbf{X}^A$, $\mathbf{Y}^A$) and the supervised segmentation objective in Eq.(3) with gradient descent

**end for**

SemiClf [22] is an image classification method, which is also semi-supervised by design, built around two alternating objective functions. However, this semi-supervised algorithm has some critical drawbacks. The minimum patch size that the method can classify is $32 \times 32$ intensity pixels. This is far too coarse-grained to provide information at a pixel level. Second, the training procedure is inefficient. During training, the method samples the patches to tackle the imbalance

in the cluster size. The sampling hinders mini-batch training, degenerating training efficiency. We highlight benefits of our new semantic segmentation method below.

*a) Obtaining fine-grained segmentation maps:* SemiClf [22] classifies echosounder patches with a shape of $32 \times 32 \times 4$ into three classes using the modified architecture of VGG-16 [60], where 4 in the patch shape indicates the number of frequency channels. The architecture corresponds to an encoder of the neural networks. The result can be interpreted as a coarse-grained segmentation, where the minimum resolution of prediction is equal to the patch shape. On the contrary, our method leverages the modified U-Net architecture [23], providing a fine-grained segmentation where the minimum resolution is $1 \times 1 \times 4$.

Training the CNN for semantic segmentation is much more challenging than the one for classification because the large and sophisticated architecture may hinder the backpropagation of the gradient to the other end of the network. We leverage the coupled architecture of encoder and decoder using dilations and concatenation functions to facilitate the backpropagation of the gradient, as suggested in U-Net. In addition, we simplify the data preprocessing by avoiding applying the criteria for determining which class each patch belongs to, which is required for the classification task.

*b) Annotation-free class-rebalancing weight:* Our method utilizes the cross-entropy loss for both the unsupervised and supervised learning schemes. However, the cross-entropy loss does not account well for imbalanced classes as it sums over all the intensities [61]. A common approach to tackle the class imbalance problem is to allocate class importance to mitigate the imbalance based on the class distribution. This includes rebalancing the class weights [62]–[64] and regulating the learning frequency by sampling [22], [53], [65]. Table I shows that the echosounder data is severely class-imbalanced to the given classes, where more than 99 percent of the backscattering intensities belong to the background (BG) class consisting of the water and seabed features. The supervised segmentation objective, therefore, should deal with the class imbalance problem in the echosounder data.

The unsupervised clustering objective should also tackle the class imbalance problem. The clustering approaches based on DeepCluster [53] can result in a trivial solution, such as empty clusters or immensely larger clusters than their average size. This causes the imbalance among the pseudo-classes, hindering the CNN to address the structured representation. To tackle the imbalance, approaches based on DeepCluster [22], [53] purposely equalize the cluster size by sampling to uniformly distribute the pseudo-classes. For the segmentation task, however,

sampling pixels to create the class-balanced pseudo-labels is not a strategic choice in terms of the learning efficiency as the discarded pixels may create a mask in the pseudo-label, hindering end-to-end mini-batch training.

Hence, we apply the class-rebalancing weight technique to the objective functions to bypass the sampling procedure. The weight leverages the number of predictions to each pseudo-class or class attribute instead of leveraging the available class annotation, differentiating our method from the previous studies [5], [22]. The class-rebalancing weight $w_{cls,k}$ for the unsupervised clustering objective $\mathcal{L}_{cls}$ in Equation (2) is depicted as:

$$w_{cls,k} = \frac{\hat{w}_{cls,k}}{\sum_{k \in K} \hat{w}_{cls,k}}, \text{ where } \hat{w}_{cls,k} = \frac{N}{KN_k}. \tag{4}$$

In this expression, $N$ represents the total number of pseudo-labels in a mini-batch. $K$ represents the number of pseudo-classes or clusters that we predefined. $N_k$ represents the number of pseudo-labels of the pseudo-class $k$, where the sum over the $K$ pseudo-classes is equal to $N$ ($N = \sum_{k \in K} N_k$). Equation (4) indicates that the pseudo-classes larger than the average size $N/K$ are penalized by the smaller weight than the other classes.

In this study, rather than forcing the balance in a few available annotations, we introduce the class-rebalancing weight $w_{seg,c}$ for the supervised segmentation objective $\mathcal{L}_{seg}$ in Equation (3) depicted as:

$$w_{seg,c} = \frac{\hat{w}_{seg,c}}{\sum_{c \in C} \hat{w}_{seg,c}}, \text{ where } \hat{w}_{seg,c} = \frac{N}{CN_c}. \tag{5}$$

In this expression, $C$ represents the number of classes in the annotated data. $N_c$ represents the number of prediction of the class $c$, where the sum over the C classes is equal to $N$ ($N = \sum_{c \in C} N_c$). Note that we count $N_c$ from the prediction of the model rather than the available annotation to avoid the deterministic weight values, resulting in the annotation-free class rebalancing weight.

## IV. Experiment

The purpose of the experiments is to explore the robustness of the proposed method in the semi-supervised learning environment that exploits limited annotations and, at the same time, the contribution of the unannotated data. We evaluate our method by comparing it with other segmentation models applied for the analysis of the echosounder data, where the evaluation metrics include prediction accuracy, F1-score, confusion matrix, Cohen's kappa [66], and AUC-ROC (Area Under the Curve - Receiver Operating Characteristics) [67].

TABLE I

OVERVIEW OF THE ECHOSOUNDER DATA USED FOR TRAINING AND TEST/VALIDATION

| Year | Training set (2016-2017) | Test set (2019) |
|---|---|---|
| No.patches | 200 | 60 |
| The number of backscattering intensities per class (proportion) | | |
| BG | 12,995,258 (0.9914) | 3,904,023 (0.9928) |
| SE | 61,018 (0.0047) | 11,776 (0.0030) |
| OT | 50,924 (0.0039) | 16,361 (0.0042) |
| Total | 13,107,200 (1.0000) | 3,932,160 (1.0000) |

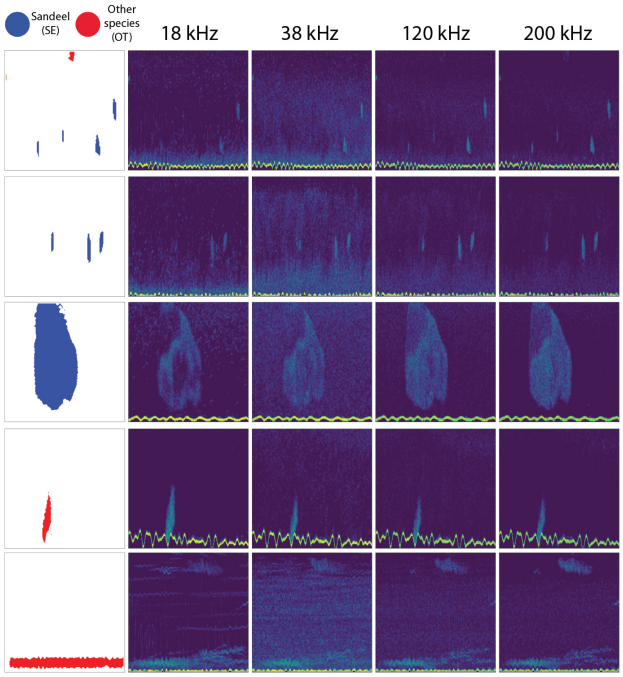


Fig. 3. Five pairs of the training patches. The annotation map (leftmost) and the echosounder data for each frequency channel are shown. The colors in the annotation map indicate the classes: background (BG) in white, other fish species (OT) in red, and sandeel (SE) in blue. The horizontal yellow line at the lower part of the echosounder data is the boundary between water and the seabed. Note that some patches do not include any fish pixel as a result of random patch extraction.

## A. *Data setup*

We leverage the echosounder data from 2016-2017 to train the CNN-based segmentation model and the trained model is evaluated using the echosounder data from 2019. The size of the input echosounder patches is $256 \times 256 \times 4$, where 4 indicates the number of echosounder channels

(18, 38, 120, 200 kHz). We randomly extract the echosounder patches from the echosounder data. 200 patches from the echosounder data between 2016-2017 is used for the training set, and 60 patches from the echosounder data in 2019 is used for the test set. In addition to those sets, we extract 30 patches for the validation set from the echosounder data between 2016-2017 to tune the hyperparameters. There is no overlap among the patches. The model output is the segmentation map of the corresponding input, segmented by the three given classes. Table I and Figure 3 show respectively a subset of the training patches and the general information of the training and test sets.

### *B. Annotation ratio*

To explore the impact of our semi-supervised method, we compute the annotation ratio, which measures the ratio of the number of annotated patches to the number of the entire set of training patches. Six ratios are studied, namely, 1.00, 0.40, 0.35, 0.30, 0.25, and 0.20. The annotation ratio of 1.00 represents a fully supervised setting, where 200 training patches are fully annotated. The annotation ratio of .20 takes the extreme semi-supervised case in this study, where 40 out of 200 training patches are annotated while the remaining 160 patches are unannotated.

### *C. Training configuration*

The following training configuration is shared for all experiment setups. The model learns by mini-batch training, where the batch size $B_s$ is set to 2 considering the computational resource. Thus the number of feature representations in a mini-batch $N$ is 131,072 ($2 \times 256 \times 256$). The Adam optimizer [68] with learning rate $3 \times 10^{-5}$, beta $(0.9, 0.999)$, and weight decay $10^{-5}$ is applied. The training is iterated to 500 epochs for all experiments, applying early stopping [69] on the condition that the accuracy is not improved for 100 epochs. For PCA, we choose the first 32 principal components shown in Equation (1) as they capture most of the variance of the data. Three prediction classes are given ($C = 3$); background (BG), sandeel (SE), and other fish species (OT).

Regarding the choice of the number of clusters $K$, we choose $K = 512$ after testing a set of different $K$s. Table II exemplifies one of the tests when the annotation ratio is 0.20, where the AUC-ROC value of SE class (0.8306), prediction accuracy of BG and SE classes (BG accuracy 0.9861; SE accuracy 0.5312), Cohen's kappa (0.3449), and F1 score (0.9856) achieve the highest when $K = 512$. As addressed in the DeepCluster work [53], the number of cluster $K$

TABLE II

PERFORMANCE COMPARISON REGARDING DIFFERENT $K$S AT THE ANNOTATION RATIO OF .20

| 0.20 | No.clusters (K) | 256 | 512 | 1024 | 2048 |
|---|---|---|---|---|---|
| AUC-ROC | BG | 0.7582 | 0.8672 | **0.9407** | 0.9258 |
| | SE | 0.7033 | **0.8306** | 0.7075 | 0.7585 |
| | OT | 0.7873 | 0.7851 | **0.8559** | 0.6523 |
| Accuracy | BG | 0.9850 | **0.9861** | 0.9809 | 0.9847 |
| | SE | 0.4628 | **0.5312** | 0.4731 | 0.5166 |
| | OT | 0.4657 | 0.5224 | **0.5817** | 0.4881 |
| Kappa | | 0.2991 | **0.3449** | 0.3045 | 0.3374 |
| F1 | | 0.9844 | **0.9856** | 0.9828 | 0.9849 |

does not have a significant impact on the performance if we cluster the feature representations with a sufficiently large number of clusters compared to the number of classes. We tune those hyperparameters using the validation set. All the code is implemented in PyTorch [70].

### *D. Validation Methods*

Our proposed method, PredKlus, is designed specifically to exploit the intrinsic nature of unannotated data, as well as to enforce class structure by supervision, all while handling the inherent class-imbalance of echosounder data by class-rebalancing weights. One could envision other approaches for exploiting unannotated data in semantic segmentation for acoustic target detection.

As the first comparison model to highlight this, we re-implement a recently published work for generic semi-supervised semantic segmentation [71] for our specific task of acoustic target classification. This method, which we refer to as SemiCPS, also aims to integrate pseudo-class predictions (unsupervised) to the class predictions (supervised) by introducing an additional auxiliary segmentation network mirroring the main segmentation network architecture with different initializations.

SemiCPS intends to encourage high similarity between the predictions of the two networks with different initialization for the same input image. For the annotated input, each network is individually trained in a supervised manner. For the unannotated input, the main network first creates the class prediction map by processing the input. This prediction map becomes the

pseudo-label that will supervise the auxiliary network. Once the auxiliary network is updated by the pseudo-label, the main network is also supervised by the prediction map from the auxiliary network.

With SemiCPS, we implicitly explore how the unsupervised clustering objective affects the predictive performance when data are noisy. Due to the unpredictable underwater nature, the features between the target class and the non-target are visually indistinguishable in some echosounder data. This may lead the mirrored network of SemiCPS to generate incorrect pseudo-labels, which are tied to the supervision of the main network. If it eventually repeats, none of the two networks can make correct predictions. On the other hand, the pseudo-labels in our proposed method are leveraging the internal structure of the dataset and are not tied to the class supervision. This makes our proposed approach more robust against noisy data, such as the echosounder data. As we will show in Section V, SemiCPS does not compare favorably to our approach. We believe this to be due to an inability to exploit the intrinsic nature of the unannotated data leading to a propagation of errors induced by the pseudo-labeling due to the noisy nature of the data. This will be further discussed in Section V-A.

The second comparison model is the semi-supervised patch classification method [22], referred to as SemiClf, where both the annotated and the unannotated parts are involved in the analysis. This model learns from a small input patch of size $32 \times 32 \times 4$, and classifies each patch to given classes leveraging the architecture of the modified VGG-16 [60]. We train SemiClf using the same training set, after splitting one provided echosounder input ($256 \times 256 \times 4$) into 64 small patches. In the inference phase, on the other hand, we extract the small patches with stride of one pixel only, resulting in a fine-grained prediction map. A voting mechanism determines the class for each pixel, which is based on the class prediction frequency among the overlapping small patches. This significantly increases the computational complexity of SemiClf, but provides a pixel-level comparison between all methods.

The third comparison model is the fully supervised segmentation method [5], referred to as SupSeg in this study. This utilizes the same CNN architecture and the supervised segmentation objective as our proposed method, and provides the class prediction of each backscattering intensity. But it does not exploit either the unannotated part of the data or the unsupervised clustering objective. For semi-supervised settings where the annotation ratios are smaller than one, this fully supervised method ignores the unannotated part and learns from the annotated part of the training set, which is partially available.

TABLE III

MODEL PERFORMANCE COMPARISON WITH RESPECT TO AUC-ROC VALUE AND CLASS ACCURACY

| Anno. ratio | Class | AUC-ROC | | | | Accuracy | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Ours | SupSeg | SemiClf | SemiCPS | Ours | SupSeg | SemiClf | SemiCPS |
| 0.20 | BG | **0.8672** | 0.8331 | 0.7870 | 0.6369 | **0.9861** | 0.9833 | 0.7105 | 0.8472 |
| | SE | 0.8306 | 0.6576 | **0.8496** | 0.6229 | 0.5312 | 0.4813 | **0.6867** | 0.3253 |
| | OT | **0.7851** | 0.6816 | 0.4668 | 0.1991 | **0.5224** | 0.4906 | 0.0146 | 0.0000 |
| 0.25 | BG | **0.8499** | 0.8457 | 0.8390 | 0.6510 | **0.9880** | 0.9877 | 0.7748 | 0.9851 |
| | SE | **0.7952** | 0.7251 | 0.7606 | 0.5792 | **0.5290** | 0.5120 | 0.1208 | 0.1416 |
| | OT | 0.7879 | 0.7762 | **0.8387** | 0.3886 | 0.5340 | 0.5271 | **0.6726** | 0.0000 |
| 0.30 | BG | **0.9148** | 0.8763 | 0.9019 | 0.7468 | **0.9856** | 0.9851 | 0.8530 | 0.9155 |
| | SE | **0.8387** | 0.8052 | 0.8240 | 0.6005 | **0.6282** | 0.6080 | 0.6639 | 0.3295 |
| | OT | **0.8423** | 0.7744 | 0.7792 | 0.6834 | **0.5326** | 0.5231 | 0.0501 | 0.2115 |
| 0.35 | BG | **0.9385** | 0.8474 | 0.8666 | 0.7945 | 0.9842 | **0.9857** | 0.7444 | 0.8859 |
| | SE | **0.8687** | 0.7977 | 0.7770 | 0.8159 | **0.6609** | 0.6128 | 0.5938 | 0.5670 |
| | OT | **0.8930** | 0.8856 | 0.8103 | 0.5836 | **0.6419** | 0.6399 | 0.1329 | 0.1792 |
| 0.40 | BG | **0.9097** | 0.9015 | 0.8446 | 0.8455 | 0.9811 | **0.9857** | 0.9534 | 0.8846 |
| | SE | **0.8840** | 0.8103 | 0.8256 | 0.8539 | **0.6304** | 0.6238 | 0.3769 | 0.5671 |
| | OT | **0.8621** | 0.8128 | 0.7968 | 0.7572 | **0.7307** | 0.6029 | 0.1748 | 0.3226 |
| 1.00 | BG | **0.9262** | 0.8696 | 0.8651 | 0.9088 | **0.9888** | 0.9886 | 0.8602 | 0.8687 |
| | SE | **0.8705** | 0.8619 | 0.8221 | 0.8634 | **0.6779** | 0.6076 | 0.3420 | 0.5247 |
| | OT | 0.9025 | 0.8285 | 0.8135 | **0.9045** | **0.7461** | 0.7180 | 0.4489 | 0.6548 |

TABLE IV

MODEL PERFORMANCE COMPARISON WITH RESPECT TO COHEN-KAPPA AND F1 SCORE

| Anno. ratio | Cohen-kappa | | | | F1 score | | | |
|---|---|---|---|---|---|---|---|---|
| | Ours | SupSeg | SemiClf | SemiCPS | Ours | SupSeg | SemiClf | SemiCPS |
| 0.20 | **0.3449** | 0.3267 | 0.0191 | 0.0183 | **0.9856** | 0.9843 | 0.8208 | 0.9047 |
| 0.25 | **0.3756** | 0.3747 | 0.0383 | 0.0454 | **0.9869** | 0.9868 | 0.8634 | 0.9778 |
| 0.30 | **0.3579** | 0.3558 | 0.0571 | 0.0560 | **0.9857** | 0.9854 | 0.9107 | 0.9436 |
| 0.35 | 0.3774 | **0.3887** | 0.0306 | 0.0515 | 0.9855 | **0.9862** | 0.8448 | 0.9276 |
| 0.40 | 0.3565 | **0.3878** | 0.0951 | 0.0664 | 0.9841 | **0.9863** | 0.9638 | 0.9277 |
| 1.00 | **0.4796** | 0.4540 | 0.0596 | 0.0770 | **0.9889** | 0.9885 | 0.9143 | 0.9194 |

## V. Result and discussion

Our method and three comparison models, e.g. SupSeg [5], SemiClf [22] and SemiCPS [71], are evaluated by the various performance measures using the test echosounder data specified in Table I. The measures include AUC-ROC value and the class prediction accuracy for each class and annotation ratio (Table III), Cohen's kappa (kappa), and F1 score regarding each annotation ratio (Table IV). The area under the ROC curve is AUC, where a higher AUC indicates better segmentation performance. Regarding the class prediction accuracy, note that the SE class achieves the lowest prediction accuracy than any other class for the many setups. This indicates that the SE class is a conservative estimate [22].

In addition to these measures, the confusion matrix and the corresponding ROC curve for each setup are computed for the comparison, as shown in Figures 4-9. For the confusion matrices, each row of these confusion matrices sums to one, indicating the ground truth of the prediction. Each column illustrates the class prediction of the method. The first column and row indicate the BG class, the second and the third columns and rows denoting the SE class and the OT class respectively. For the ROC curves, the vertical axis indicates a true-positive rate while the horizontal axis shows a false-positive rate. For the visual comparison, we provide the prediction map of the test data in Figures 10-12, where four parts of the echosounder data in 2019 and their prediction maps are visualized. Overall, the results show that our semi-supervised method outperforms the comparison models throughout annotation ratios.

### *A. Comparison to semi-supervised segmentation method using pseudo-labels (SemiCPS)*

Tables III-IV show that our proposed method outperforms SemiCPS through the entire evaluation metrics in the semi-supervised setups containing the annotation ratios of 0.20-0.40. The greatest performance difference is observed at the annotation ratio of 0.20, which is the most extreme semi-supervised setup. Our method achieves the kappa score of 0.3449, which is 18.8 times greater the kappa score of SemiCPS (0.0183).

The prediction maps in Figure 10 also visually validate the outperforming results of our proposed method. SemiCPS does not make predictions close to the fish patterns for the annotation ratios of 0.20-0.25, but tends to capture the fish class patterns from the annotation ratios of 0.30 and higher. However, quite a few fish patterns are still misclassified to the BG class, yielding a smaller prediction area and underperforming results than our proposed method. Our proposed method, in contrast, tends to capture most of the major fish patterns on the prediction map from

the annotation ratio of 0.20. Although the prediction map appears noisy due to misclassification of small clutter patterns at low annotation ratios, the noise is filtered out as the annotation ratio increases and shows a good prediction map close to the ground truth and the input. We discover the same visual trends in Figures 11-12.

### *B. Comparison to semi-supervised patch-based segmentation (SemiClf)*

Compared to SemiClf [22], our proposed method outperforms throughout the measures and setups. We argue that the novelties of our method, such as the learning mechanism for the fine-grained segmentation and the annotation-free class-rebalancing technique, contribute to achieving the surpassing result by addressing the shortcomings of patch-based SemiClf. The kappa scores contrast the difference nicely, where ours achieves 18.3 times greater scores than SemiClf with the annotation ratio of 0.20 (ours 0.3449; SemiClf 0.0191).

In addition to the poor prediction maps shown in Figures 10-12, another critical drawback of SemiClf is misclassification of the seabed feature, which is known for a considerably higher intensity than the other fish classes [5]. The seabed feature is marked with a distinct yellow horizontal line in the input echosounder data. As shown in the prediction maps, SemiClf and SemiCPS predict the seabed as one of the fish classes (blue or red) throughout the annotation ratios. In contrast, our method learns the seabed feature and correctly predicts it to BG class in white as intended.

### *C. Comparison to fully supervised method (SupSeg)*

We compare the result of our method to SupSeg [5], to investigate how the unsupervised clustering objective and the unannotated data improve the predictive performance. Overall, our proposed method outperforms SupSeg through the entire annotation ratios for the entire AUC-ROC values and the SE and OT class accuracies in Table III. The results indicate that the unsupervised clustering objective improves the performance of the segmentation task by effectively exploiting the structured representation from both the unannotated data and the available annotated data.

Note that our proposed method outperforms SupSeg for the annotation ratio of 1.00 (fully-supervised case). With this result, we argue that our proposed method is generic and outperforms the conventional fully-supervised learning methods, such as SupSeg. Alternating two objective functions are applicable to the fully-supervised case, which facilitates the interconnection of the

two objectives to make good use of the annotated data based on the clustering structure. By the iteration, the datapoints in each cluster gradually share the dominant class annotation, and eventually have the same class prediction, approximating the decision boundaries that SupSeg achieves to some extent.

In Table IV, we find two inconsistent cases for the annotation ratios of 0.35 and 0.40, where SupSeg achieves greater Kappa and F1 scores. However, we argue that this result does not undermine the robustness of our proposed method. Instead, we believe that SupSeg is biased to make more predictions for the BG class, where the bias is related to a severe class imbalance in the training data, especially in the increased part of the annotated data. The prediction accuracy of the BG class for these annotation ratios validates our reasoning, where SupSeg achieves better accuracy than our proposed method for these annotation ratios (SupSeg 0.9857, ours 0.9842 with the annotation ratio of 0.35; SupSeg 0.9811, ours 0.9857 with the annotation ratio of 0.40).

On the other hand, the prediction accuracies of two fish classes do not seem to increase as much as it increases in our method (ours 0.6609, SupSeg 0.6128 with the annotation ratio of 0.35 and the SE class accuracy; ours 0.6304, SupSeg 0.6238 with the annotation ratio of 0.40 and the SE class accuracy; ours 0.6419, SupSeg 0.6399 with the annotation ratio of 0.35 and the OT class accuracy; ours 0.7307, SupSeg 0.6029 with the annotation ratio of 0.40 and the OT class accuracy). Through visual inspection of the annotated part of the training data, we are able to obtain other grounds for our argument.

When performing the visual inspection of the increased part of the training set between the annotation ratio of 0.30 and 0.35, where ten input-annotation data pairs are increased, we discover that five out of ten data pairs consist of only BG class pixels without any fish class pixel. Analogously, we discover that six out of ten data pairs consist of only BG class pixels without any fish class pixel between the annotation ratio of 0.35 and 0.40. For the entire training data, the case that no fish intensity pixels are obtained in the input takes about 20 percent of the training data on average. Hence, we argue that the class imbalance found with these annotation ratios is more severe than the other cases and causes the prediction bias towards the BG class for the SupSeg case.

### *D. Confusion matrix and ROC curve*

Figures 4-9 compare our proposed method to other comparison models using confusion matrices and ROC curves. When comparing the diagonal components of the confusion matrices

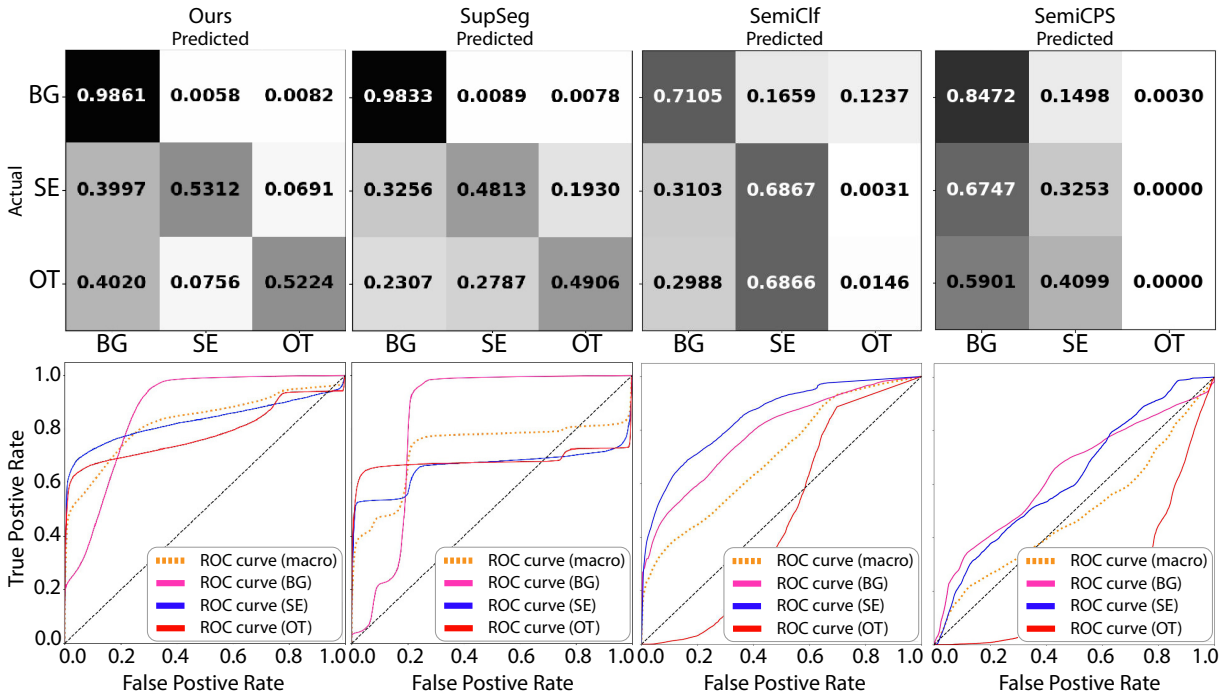


Fig. 4. The confusion matrices and the corresponding AUC-ROC plots of the annotation ratio of 0.20.

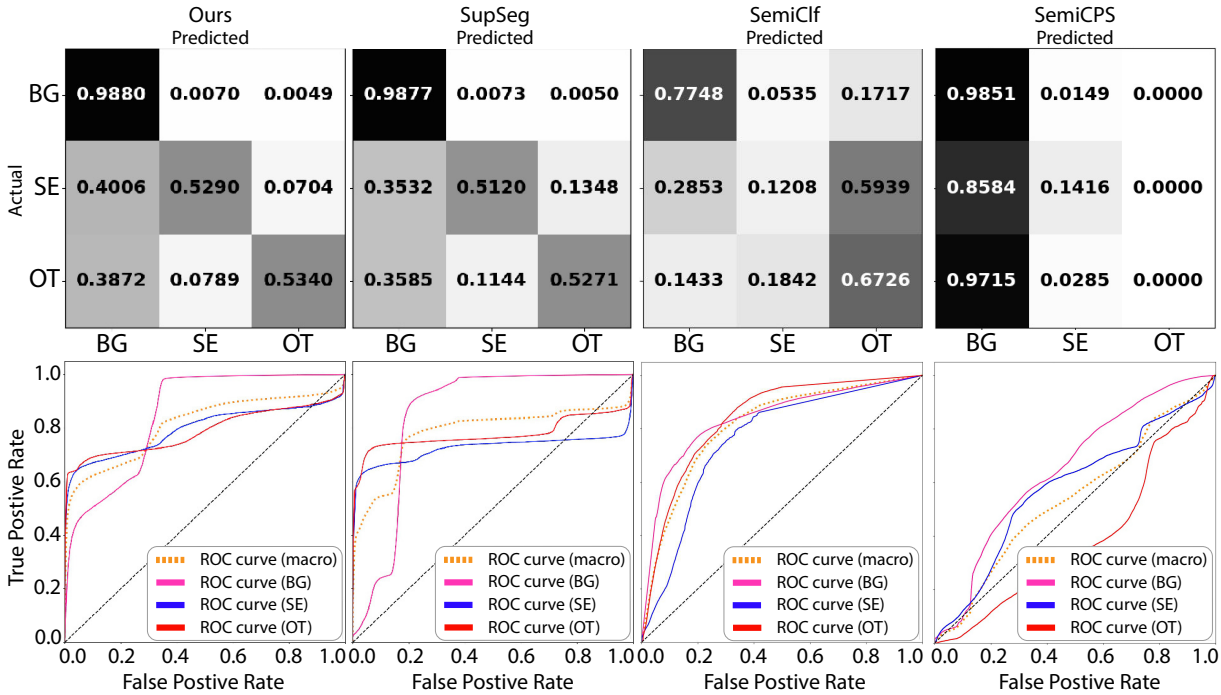


Fig. 5. The confusion matrices and the corresponding AUC-ROC plots of the annotation ratio of 0.25.

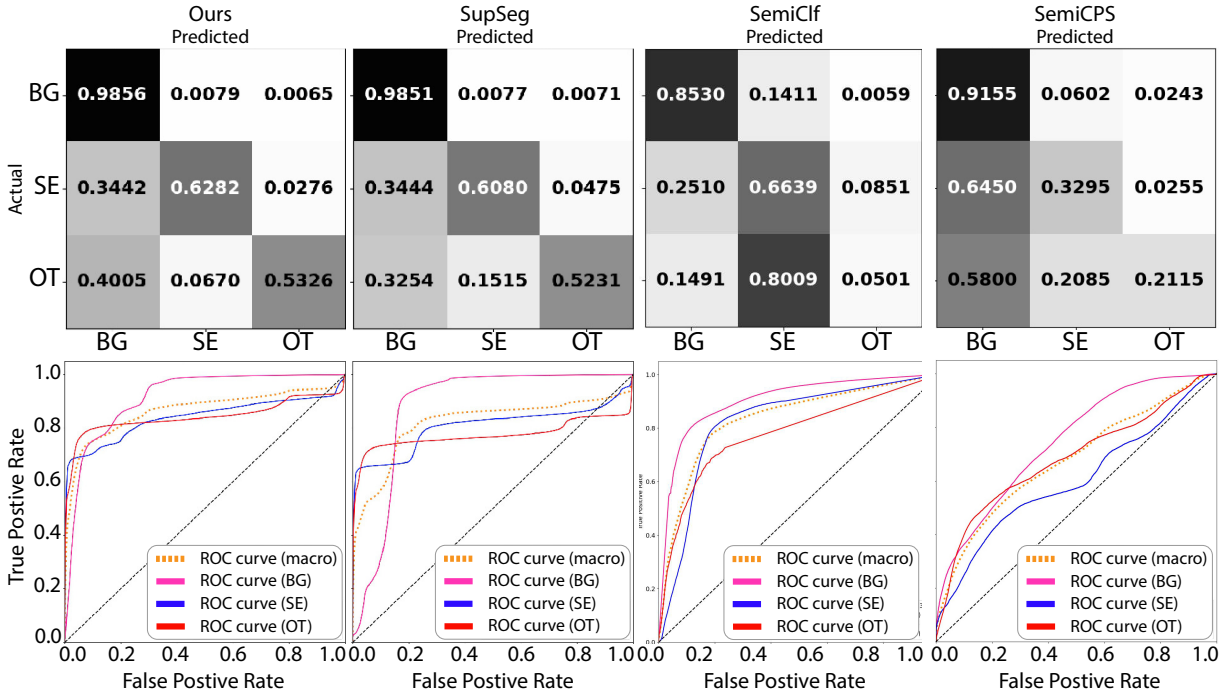


Fig. 6. The confusion matrices and the corresponding AUC-ROC plots of the annotation ratio of 0.30.

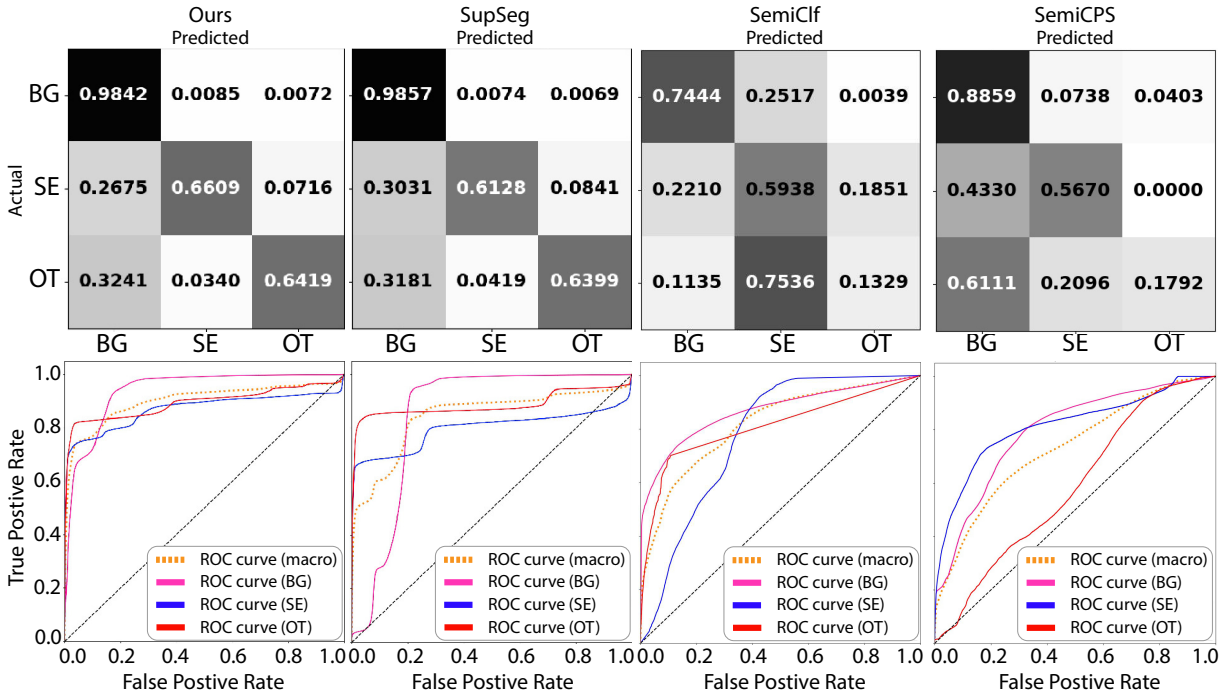


Fig. 7. The confusion matrices and the corresponding AUC-ROC plots of the annotation ratio of 0.35.

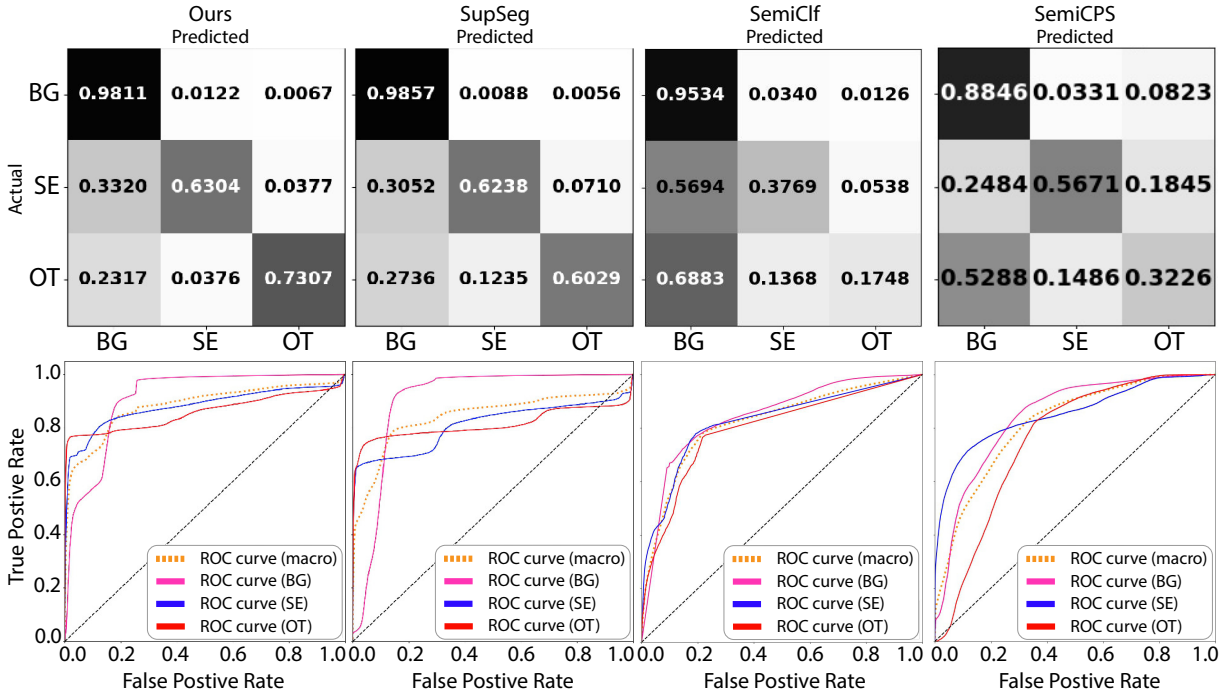


Fig. 8. The confusion matrices and the corresponding AUC-ROC plots of the annotation ratio of .40.

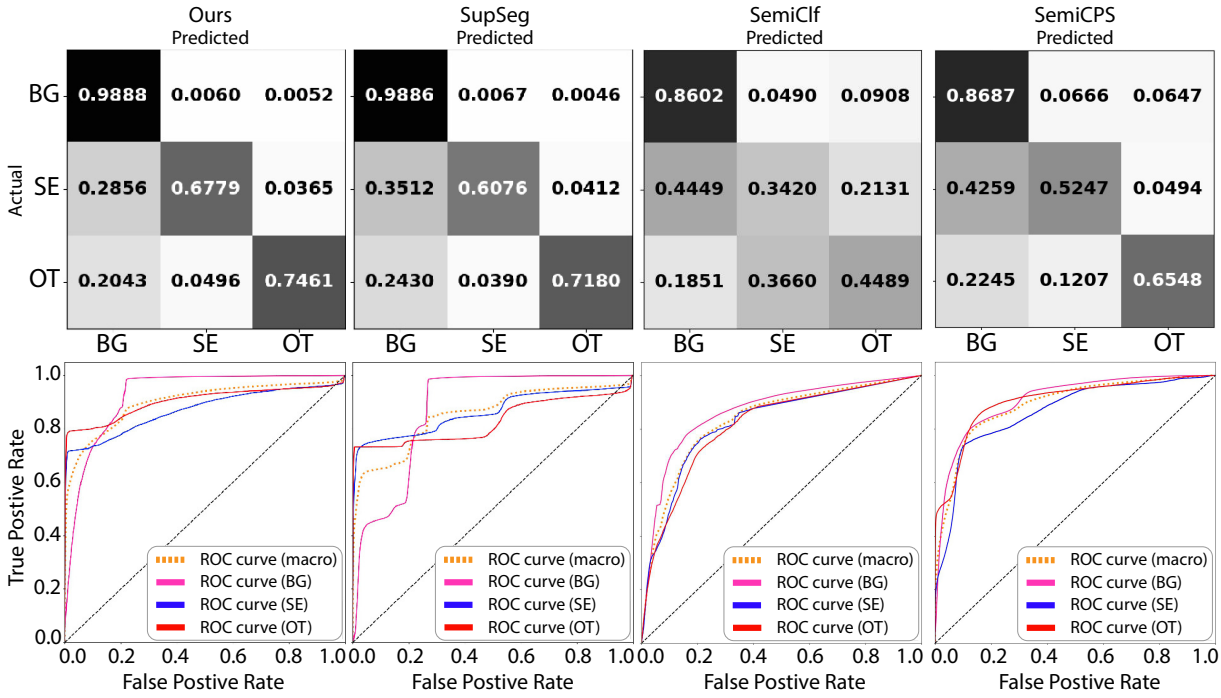


Fig. 9. The confusion matrices and the corresponding AUC-ROC plots of the annotation ratios of 1.00.

visually, our proposed method shows more distinct diagonal components than the other models. This implies that (1) our proposed method can be seen to outperform the comparison model in terms of the class accuracy as illustrated in Table III; (2) our proposed method also achieves lower false-positive rates within fish classes compared to other models when having a deeper look at the diagonal components of the SE and OT classes (second and third row and column). For example, comparing the false-positive rate of SE prediction of the OT class ground truth, shown in the second column and the third row of the confusion matrices, ours achieves lower false-positive rates throughout the semi-supervised setups. This result is consistent with the false-positive rate of OT prediction of the SE class ground truth, shown in the third column and the second row of the matrices.

The ROC curve shows the trade-off between true-positive and false-positive rates. The curves indicate that segmentation models with curves closer to the top-left corner perform better, resulting in greater area under the curve (AUC) as depicted in Table III. The results in the curves and the AUC values validate the outperforming result of our method.

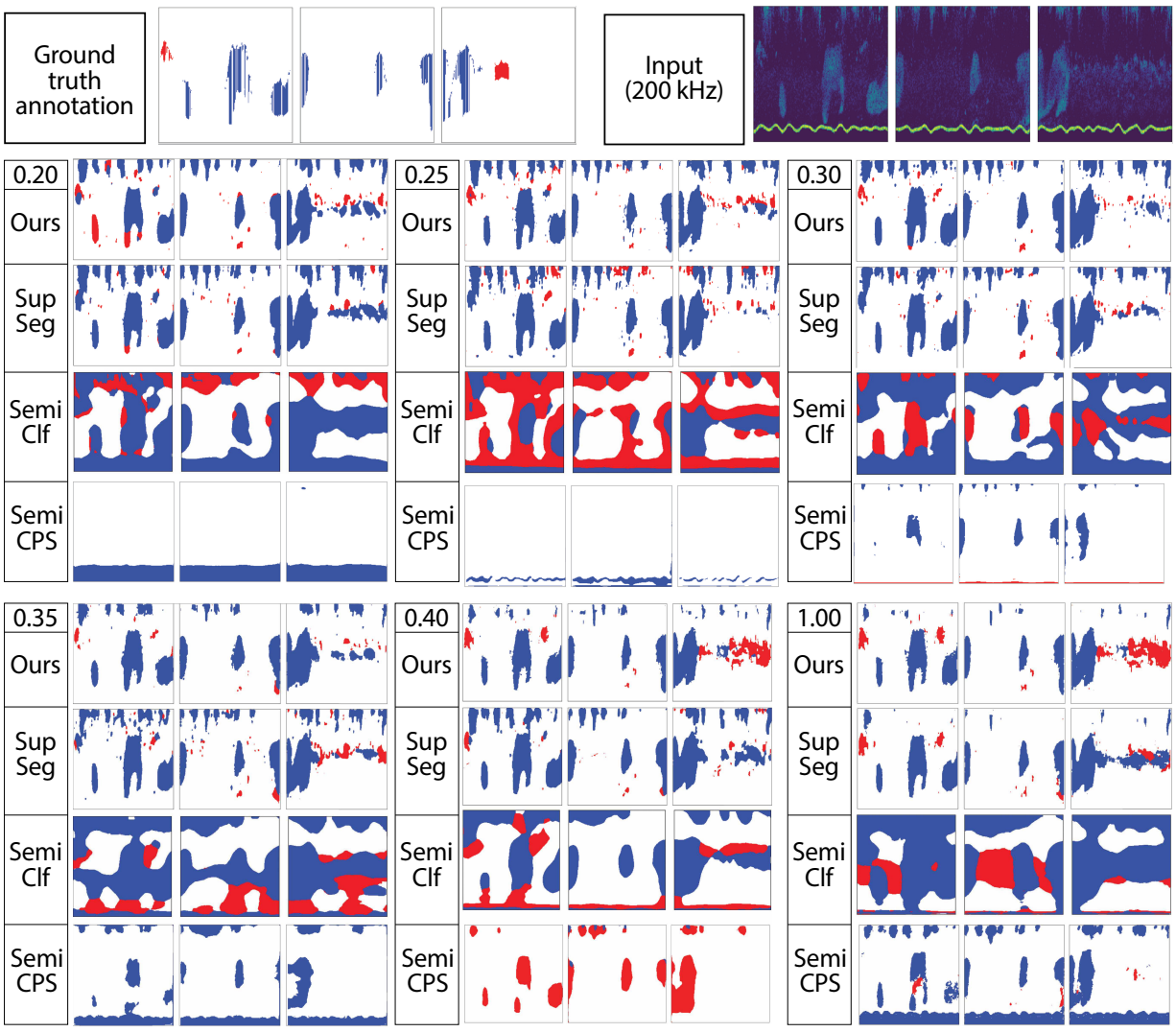


Fig. 10. Prediction maps of the test echosounder data with respect to the annotation ratios. The colors in the annotation map indicate the classes: background (BG) in white, other fish species (OT) in red, and sandeel (SE) in blue.

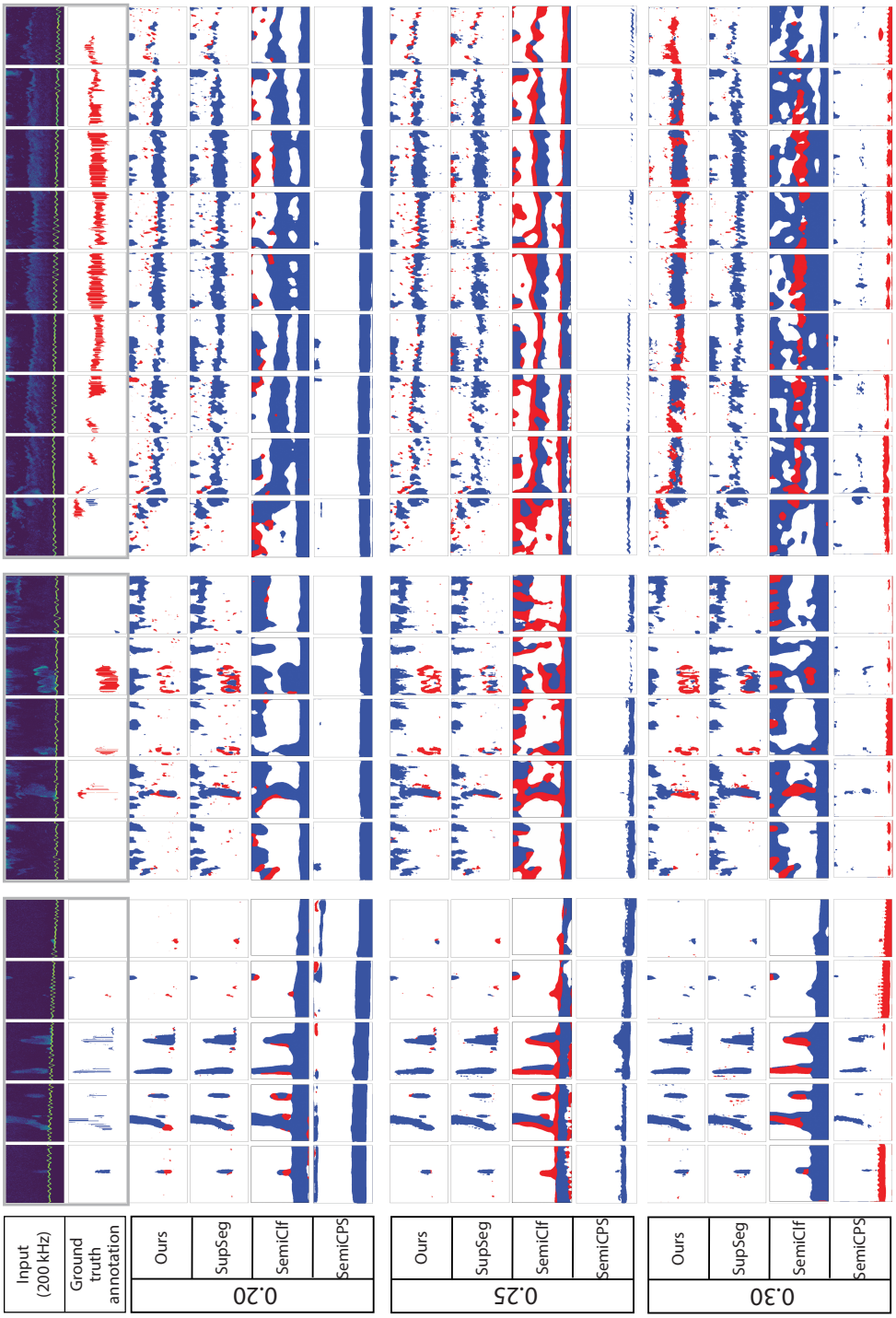


Fig. 11. Prediction maps of the test echosounder data with the annotation ratio of 0.20, 0.25, and 0.30. The colors in the annotation map indicate the classes: background (BG) in white, other fish species (OT) in red, and sandeel (SE) in blue. (a) depicts the case where the SE class is dominant, whereas (b) and (c) shows the case where the OT class is dominant.

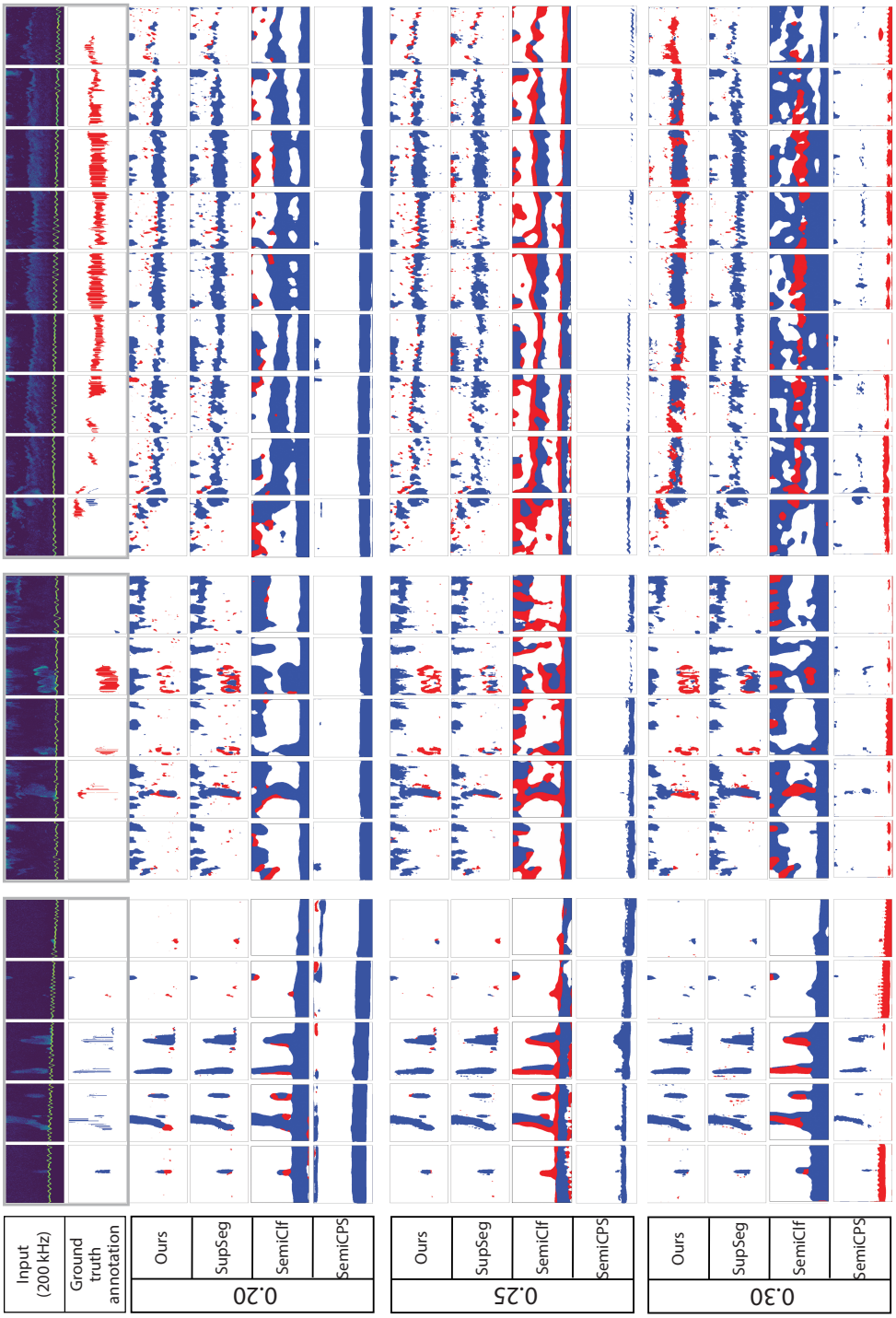


Fig. 12. Prediction maps of the test echosounder data with the annotation ratio of 0.35, 0.40, and 1.00. The colors in the annotation map indicate the classes: background (BG) in white, other fish species (OT) in red, and sandeel (SE) in blue. (a) depicts the case where the SE class is dominant, whereas (b) and (c) shows the case where the OT class is dominant.

## VI. Conclusion

In this paper, we propose a novel semi-supervised deep learning method for semantic segmentation of echosounder data. Our method (1) considerably reduces the dependency on the annotated data, achieving comparable results with the fully supervised segmentation method [5], by leveraging 40 percent of the annotated data in addition to unannotated data. Our method also (2) outperforms the other semi-supervised methods for echosounder data [22], [71]. Our methodological novelty is to (3) take advantage of deep clustering to exploit the underlying structure of the training data regardless of the annotation in a semi-supervised learning scheme. In addition, our method (4) is end-to-end and mini-batch trainable, and (5) regulates the class imbalance based on the model prediction without leveraging the annotated part of data. The rigorous and extensive experiments validate the robustness of the proposed method, where various performance measures are introduced.

Our proposed method is generic and applicable to other fish species with a small amount of annotated echosounder data. To the best of our knowledge, this is the first semi-supervised semantic segmentation paper for the echosounder data analysis based on deep learning. The promising results imply that our proposed method can reduce the expensive costs required for the annotation. The performance can be improved by utilizing semantic information, e.g., a simple classifier that can exclude the background class pixels when collecting the echosounder data.

In future work, we intend to explore the uncertainty of the segmentation results to improve the interpretability of the model prediction. As a further example of future work, we intend to extend our method to take the uncertainty into account in order to create more crisp and clear decision boundaries among the clusters when the pseudo-labels are created.

*Acknowledgements*

This work was financially supported by the Research Council of Norway (RCN), grant no. 270966 ”Ubiquitous Cognitive Computer Vision for Marine Services” (COGMAR). The work was further financed by the RCN grant no. 309439 Centre for Research-based Innovation ”Visual Intelligence” and consortium partners, as well as RCN grant no. 309512 Centre for Research-based Innovation in Marine Acoustic Abundance Estimation and Backscatter Classification (CRIMAC).

# REFERENCES


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

[2] T. Zhou, L. Li, X. Li, C.-M. Feng, J. Li, and L. Shao, "Group-wise learning for weakly supervised semantic segmentation," *IEEE Trans. Image Process.*, vol. 31, pp. 799–811, 2022.

[3] C. Ge, H. Sun, Y.-Z. Song, Z. Ma, and J. Liao, "Exploring local detail perception for scene sketch semantic segmentation," *IEEE Trans. Image Process.*, vol. 31, pp. 1447–1461, 2022.

[4] W.-J. Lee and T. K. Stanton, "Statistics of broadband echoes: Application to acoustic estimates of numerical density of fish," *IEEE J. Ocean. Eng.*, vol. 41, no. 3, pp. 709–723, 2016.

[5] O. Brautaset, A. U. Waldeland, E. Johnsen, K. Malde, L. Eikvil, A.-B. Salberg, and N. O. Handegard, "Acoustic classification in multifrequency echosounder data using deep convolutional neural networks," *ICES J. Mar. Sci.*, vol. 77, no. 4, pp. 1391–1400, 2020.

[6] P. L. D. Roberts, J. S. Jaffe, and M. M. Trivedi, "Multiview, broadband acoustic classification of marine fish: A machine learning framework and comparative analysis," *IEEE J. Ocean. Eng.*, vol. 36, no. 1, pp. 90–104, 2011.

[7] R. J. Korneliussen, "Acoustic target classification," *Coop. Res. Rep. - Int. Counc. Explor. Sea*, no. 344, p. 104, 2018.

[8] J. Simmonds and D. N. MacLennan, *Fisheries acoustics: theory and practice*. John Wiley & Sons, 2008.

[9] R. J. Korneliussen, Y. Heggelund, G. J. Macaulay, D. Patel, E. Johnsen, and I. K. Eliassen, "Acoustic identification of marine species using a feature library," *Methods Oceanogr.*, vol. 17, pp. 187–205, 2016.

[10] M. Woillez, P. Ressler, C. Wilson, and J. Horne, "Multifrequency species classification of acoustic-trawl survey data using semi-supervised learning with class discovery," *J. Acoust. Soc. Am*, vol. 131, no. 2, pp. 184–190, 2012.

[11] M. Peña, "Robust clustering methodology for multi-frequency acoustic data: A review of standardization, initialization and cluster geometry," *Fish. Res.*, vol. 200, pp. 49–60, 2018.

[12] R. Proud, R. Mangeni-Sande, R. J. Kayanda, M. J. Cox, C. Nyamweya, C. Ongore, V. Natugonza, I. Everson, M. Elison, L. Hobbs *et al.*, "Automated classification of schools of the silver cyprinid rastrineobola argentea in lake victoria acoustic survey data using random forests," *ICES J. Mar. Sci.*, vol. 77, no. 4, pp. 1379–1390, 2020.

[13] S. M. Gugele, M. Widmer, J. Baer, J. T. DeWeber, H. Balk, and A. Brinker, "Differentiation of two swim bladdered fish species using next generation wideband hydroacoustics," *Sci. Rep.*, vol. 11, no. 1, pp. 1–10, 2021.

[14] D. Demer, L. Andersen, C. Bassett, L. Berger, D. Chu, J. Condiotty, and G. Cutter, "Evaluation of a wideband echosounder for fisheries and marine ecosystem science," *Coop. Res. Rep. - Int. Counc. Explor. Sea*, no. 336, p. 69, 2017.

[15] P. Baldi, *Deep Learning in Science*. Cambridge University Press, 2021.

[16] A. Ordoñez, I. Utseth, O. Brautaset, R. Korneliussen, and N. O. Handegard, "Evaluation of echosounder data preparation strategies for modern machine learning models," *Fisheries Research*, vol. 254, no. 106411, 2022.

[17] X. Luo, X. Qin, Z. Wu, F. Yang, M. Wang, and J. Shang, "Sediment classification of small-size seabed acoustic images using convolutional neural networks," *IEEE Access*, vol. 7, pp. 98 331–98 339, 2019.

[18] T. P. Marques, M. Cote, A. Rezvanifar, A. B. Albu, K. Ersahin, T. Mudge, and S. Gauthier, "Instance segmentation-based identification of pelagic species in acoustic backscatter data," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 4378–4387, 2021.

[19] S. C. Lowe, L. P. McGarry, J. Douglas, J. Newport, S. Oore, C. Whidden, and D. J. Hasselman, "Echofilter: A deep learning segmentation model improves the automation, standardization, and timeliness for post-processing echosounder data in tidal energy streams," *Front. Mar. Sci.*, vol. 9, no. 867857, 2022.

[20] O. Chapelle, B. Scholkopf, A. Zien *et al.*, "Semi-supervised learning," *IEEE Trans. Neural Netw.*, vol. 20, no. 3, pp. 542–542, 2006.

[21] H. Pham, Z. Dai, Q. Xie, and Q. V. Le, "Meta pseudo labels," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 11 557–11 568, 2021.

[22] C. Choi, M. Kampffmeyer, N. O. Handegard, A.-B. Salberg, O. Brautaset, L. Eikvil, and R. Jenssen, "Semi-supervised target classification in multi-frequency echosounder data," *ICES J. Mar. Sci.*, vol. 78, no. 7, pp. 2615–2627, 2021.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *in Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv. (MICCAI)*, pp. 234–241, 2015.

[24] T. Glasmachers, "Limits of end-to-end learning," *in Proc. Asian Conf. Mach. Learn. (ACML)*, pp. 17–32, 2017.

[25] Y. Li, L. Song, Y. Chen, Z. Li, X. Zhang, X. Wang, and J. Sun, "Learning dynamic routing for semantic segmentation," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020.

[26] S. Hao, Y. Zhou, and Y. Guo, "A brief survey on semantic segmentation with deep learning," *Neurocomputing*, vol. 406, pp. 302–321, 2020.

[27] Q. Liu, M. Kampffmeyer, R. Jenssen, and A.-B. Salberg, "Dense dilated convolutions' merging network for land cover classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 9, pp. 6309–6320, 2020.

[28] A. Vali, S. Comai, and M. Matteucci, "Deep learning for land use and land cover classification based on hyperspectral and multispectral earth observation data: A review," *Remote Sens.*, vol. 12, no. 2495, p. 31, 2020.

[29] L. T. Luppino, M. Kampffmeyer, F. M. Bianchi, G. Moser, S. B. Serpico, R. Jenssen, and S. N. Anfinsen, "Deep image translation with an affinity-based change prior for unsupervised multimodal change detection," *IEEE Trans. Geosci. Remote Sens.*, 2021.

[30] X. Yu, J. Fan, J. Chen, P. Zhang, Y. Zhou, and L. Han, "NestNet: a multiscale convolutional neural network for remote sensing image change detection," *Int. J. Remote Sens.*, vol. 42, no. 13, pp. 4898–4921, 2021.

[31] S. Hansen, S. Kuttner, M. Kampffmeyer, T.-V. Markussen, R. Sundset, S. K. Øen, L. Eikenes, and R. Jenssen, "Unsupervised supervoxel-based lung tumor segmentation across patient scans in hybrid PET/MRI," *Expert Syst. Appl*, vol. 167, no. 114244, 2021.

[32] M. A. Naser and M. J. Deen, "Brain tumor segmentation and grading of lower-grade glioma using deep learning in mri images," *Comput. Biol. Med.*, vol. 121, no. 103758, 2020.

[33] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 3, pp. 1341–1360, 2020.

[34] K. Wickstrøm, M. Kampffmeyer, and R. Jenssen, "Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps," *Med. Image Anal.*, vol. 60, no. 101619, 2020.

[35] D. Jha, S. Ali, N. K. Tomar, H. D. Johansen, D. Johansen, J. Rittscher, M. A. Riegler, and P. Halvorsen, "Real-time polyp detection, localization and segmentation in colonoscopy using deep learning," *IEEE Access*, vol. 9, pp. 40 496–40 510, 2021.

[36] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[37] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1.

[38] Y. Guo, Y. Liu, T. Georgiou, and M. S. Lew, "A review of semantic segmentation using deep neural networks," *Int. J. Multimed. Info. Retr.*, vol. 7, no. 2, pp. 87–93, 2018.

[39] N. Daan, P. Bromley, J. Hislop, and N. Nielsen, "Ecology of north sea fish," *Neth. J. Sea Res.*, vol. 26, no. 2-4, pp. 343–386, 1990.

[40] R. W. Furness, "Management implications of interactions between fisheries and sandeel-dependent seabirds and seals in the north sea," *ICES J. Mar. Sci.*, vol. 59, no. 2, pp. 261–269, 2002.

[41] E. Johnsen, G. Rieucau, E. Ona, and G. Skaret, "Collective structures anchor massive schools of lesser sandeel to the seabed, increasing vulnerability to fishery," *Mar. Ecol.: Prog. Ser.*, vol. 573, pp. 229–236, 2017.

[42] D. N. MacLennan, P. G. Fernandes, and J. Dalen, "A consistent approach to definitions and symbols in fisheries acoustics," *ICES J. Mar. Sci.*, vol. 59, no. 2, pp. 365–369, 2002.

[43] R. Kloser, T. Ryan, P. Sakov, A. Williams, and J. Koslow, "Species identification in deep water using multiple acoustic frequencies," *Can. J. Fish. Aquat. Sci.*, vol. 59, no. 6, pp. 1065–1077, 2002.

[44] R. J. Korneliussen and E. Ona, "Synthetic echograms generated from the relative frequency response," *ICES J. Mar. Sci.*, vol. 60, no. 3, pp. 636–640, 2003.

[45] D. G. Reid, "Report on echo trace classification," *Coop. Res. Rep. - Int. Counc. Explor. Sea*, no. 238, p. 107, 2000.

[46] N. O. Handegard and D. Tjøstheim, "The sampling volume of trawl and acoustics: estimating availability probabilities from observations of tracked individual fish," *Can. J. Fish. Aquat. Sci.*, vol. 66, no. 3, pp. 425–437, 2009.

[47] T. K. Stanton, P. H. Wiebe, D. Chu, M. C. Benfield, L. Scanlon, L. Martin, and R. L. Eastwood, "On acoustic estimates of zooplankton biomass," *ICES J. Mar. Sci.*, vol. 51, no. 4, pp. 505–512, 1994.

[48] D. A. Demer and S. G. Conti, "Reconciling theoretical versus empirical target strengths of krill: effects of phase variability on the distorted-wave born approximation," *ICES J. Mar. Sci.*, vol. 60, no. 2, pp. 429–434, 2003.

[49] E. Johnsen, R. Pedersen, and E. Ona, "Size-dependent frequency response of sandeel schools," *ICES J. Mar. Sci.*, vol. 66, no. 6, pp. 1100–1105, 2009.

[50] M. Barange, "Acoustic identification, classification and structure of biological patchiness on the edge of the agulhas bank and its relation to frontal features," *S. Afr. J. Mar. Sci.*, vol. 14, no. 1, pp. 333–347, 1994.

[51] J. Coetzee, "Use of a shoal analysis and patch estimation system (SHAPES) to characterise sardine schools," *Aquat. Living Resour.*, vol. 13, no. 1, pp. 1–10, 2000.

[52] S. Aronica, I. Fontana, G. Giacalone, G. L. Bosco, R. Rizzo, S. Mazzola, G. Basilone, R. Ferreri, S. Genovese, M. Barra *et al.*, "Identifying small pelagic mediterranean fish schools from acoustic and environmental data using optimized artificial neural networks," *Ecol. Inform.*, vol. 50, pp. 149–161, 2019.

[53] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," *in Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 132–149, 2018.

[54] M. Jabi, M. Pedersoli, A. Mitiche, and I. B. Ayed, "Deep clustering: On the link between discriminative models and k-means," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 6, pp. 1887–1896, 2019.

[55] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," *in Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 3, no. 2, p. 896, 2013.

[56] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *in Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 448–456, 2015.

[57] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," *in Proc. Int. Conf. Mach. Learn. (ICML)*, p. 807–814, 2010.

[58] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemom. Intell. Lab. Syst.*, vol. 2, no. 1-3, pp. 37–52, 1987.

[59] T. Lin, L. Kong, S. Stich, and M. Jaggi, "Extrapolation for large-batch training in deep learning," *in Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 6094–6104, 2020.

[60] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *in Proc. Int. Conf. Learn. Representations (ICLR)*, 2015.

[61] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 1–9, 2016.

[62] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009.

[63] M. Hyun, J. Jeong, and N. Kwak, "Class-imbalanced semi-supervised learning," *in Proc. Int. Conf. Learn. Representations (ICLR)*, 2021.

[64] L.-Z. Guo, Z. Zhou, J.-J. Shao, Q. Zhang, F. Kuang, G.-L. Li, Z.-X. Liu, G.-B. Wu, N. Ma, Q. Li, and Y.-F. Li, "Learning from imbalanced and incomplete supervision with its application to ride-sharing liability judgment," *in Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, p. 487–495, 2021.

[65] C. Wei, K. Sohn, C. Mellina, A. Yuille, and F. Yang, "CReST: A class-rebalancing self-training framework for imbalanced semi-supervised learning," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 10 857–10 866, 2021.

[66] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochem. Med.*, vol. 22, no. 3, pp. 276–282, 2012.

[67] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern Recognit.*, vol. 30, no. 7, pp. 1145–1159, 1997.

[68] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *in Proc. Int. Conf. Learn. Representations (ICLR)*, 2015.

[69] L. Prechelt, *Early stopping-but when? Neural Networks: Tricks of the trade*. Springer. New York City, USA, 1998.

[70] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," *Proc. Neural Inf. Process. Syst. (NeurIPS)*, pp. 8024–8035, 2019.

[71] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 2613–2622, 2021.

Fig. 13. Changkyu Choi

**Changkyu Choi** received the bachelor's degree in Electronic and Electrical Engineering from Sungkyunkwan University, South Korea, in 2011. He received his first master's degree in User Experience (UX) Design from the Graduate School of Information, Yonsei University, South Korea, in 2016. In 2019, he received his second master's degree in Machine Learning from the Department of Physics and Technology, UiT The Arctic University, Norway. Currently, he is a PhD candidate and an assistant professor at UiT The Arctic University. From 2011 to 2013, Choi was with Samsung Electronics, South Korea, where he primarily worked on researching and developing mobile communication devices. In 2022, he worked as a researcher at the Norwegian Computing Center, Norway. He is a design chair of the annual Northern Lights Deep Learning Conference, NLDL. His research interests are in computer vision and deep learning, and in the application of deep learning to marine image analysis.

Fig. 14. Michael Kampffmeyer

**Michael Kampffmeyer** (Member, IEEE) is an Associate Professor and Head of the Machine Learning Group at UiT The Arctic University of Norway. He is also a Senior Research Scientist II at the Norwegian Computing Center in Oslo. His research interests include medical image analysis, explainable AI, and learning from limited labels (e.g. clustering, few/zero-shot learning,

domain adaptation and self-supervised learning). Kampffmeyer received his PhD degree from UiT in 2018. He has had long-term research stays in the Machine Learning Department at Carnegie Mellon University and the Berlin Center for Machine Learning at the Technical University of Berlin. He is a general chair of the annual Northern Lights Deep Learning Conference, NLDL. For more details visit https://sites.google.com/view/michaelkampffmeyer/

Fig. 15. Arnt-Børre Salberg

**Arnt-Børre Salberg** (Member, IEEE) received the diploma degree in applied physics and the Dr.Scient degree in physics from the University of Tromsø, Tromsø, Norway, in 1998 and 2003, respectively. He is currently a Senior Research Scientist in earth observation with Norwegian Computing Center, Oslo, Norway. From February 2003 to December 2005, he had a postdoctoral and research position with the Institute of Marine Research, Tromsø. From December 2005 to October 2008, he was the Head of research and development with Dolphiscan AS, Moelv, Norway. Since October 2008, he has been with Norwegian Computing Center. From August 2001 to June 2002, he was a Visiting Researcher with the U.S. Army Research Laboratory, Adelphi, MD, USA. His research interests are in the area of earth observation, computer vision, machine learning, and statistics.

Fig. 16. Nils Olav Handegard

**Nils Olav Handegard** is a principal research scientist at the Institute of Marine Research, Bergen, Norway. He received his PhD in applied mathematics in 2004, and his main field is within fisheries acoustics. He has also an interest in general e-science, data management and data processing. He has been part of the science leadership group in the International Council for Exploration of the Sea (ICES), and has contributed to a range of working groups, including leading the Fisheries Acoustics Science and Technology group (WGFAST) and the steering group overseeing ICES coordinated scientific surveys. He has had longer-term visiting scholar stays at the University of Washington and at Princeton University, working on acoustic sensors to observe fish behavior. Current interest are applications of new methodology and data processing methods within marine ecology and fisheries oceanography, including machine learning and deep learning algorithms. He is leading the CRIMAC center, a Norwegian Centre for research-innovation funded by the research council of Norway.

Fig. 17. Robert Jenssen

**Robert Jenssen** (Senior Member, IEEE) is Director of Visual Intelligence, a Norwegian Centre for research-based innovation funded by the Research Council of Norway and consortium partners. Visual Intelligence solves research challenges in deep learning to advance image analysis. Jenssen is professor and founder of the Machine Learning Group at UiT The Arctic University of Norway. He is in addition a part time professor at the Pioneer AI Centre, University of Copenhagen, and an adjunct professor at the Norwegian Computing Center, Oslo, Norway. Jenssen received his Dr. Scient (PhD) in 2005 from UiT. He has had long-term research stays at the University of Florida, at the Technical University of Denmark, and at the Technical University of Berlin. Jenssen's research interests are in neural networks, graph and kernel-based learning, and in health and industrial applications of machine learning. Jenssen has been on the IEEE MLSP TC and on the Governing Board of IAPR. He is an editor for the journal Pattern Recognition and a member of the ELLIS Cph unit. Jenssen is general chair of the annual

Northern Lights Deep Learning Conference - NLDL.