URS: An Unsupervised Radargram Segmentation Network Based on Self-Supervised ViT With Contrastive Feature Learning Framework

Raktim Ghosh, *Student Member, IEEE*, and Francesca Bovolo, *Senior Member, IEEE*

***Abstract*—Radar sounders are air and space-borne nadir-looking sensors operating in high-frequency (HF) or very high-frequency (VHF) bands and collect subsurface backscattered returns by transmitting electromagnetic pulses. The backscatter echoes are coherently integrated to generate radargrams for investigating and identifying geophysical characteristics of subsurface targets. While recent efforts have been made to develop supervised or semisupervised deep learning models for segmenting radargrams, obtaining accurate labeled information is often a challenging task. Therefore, it is of paramount importance to develop automatic unsupervised semantic segmentation methods to characterize the subsurface targets without labeled information. Unsupervised segmentation methods learn to discover meaningful semantic contents and decompose them into distinct semantic segments with known ontology. Here, we propose an unsupervised radargram segmentation network that uses a convolution-based expansive network as a proxy decoder and a progressive stepwise reconstruction strategy of the input signal from the latent space to measure the spatial similarity with the input radar sounder signal. After designing a unique training strategy by bootstrapping the randomness inside the minibatch and combining the spatial similarity loss along with the contrastive correlation loss, the proposed architecture outperformed the state of the art in fully unsupervised settings. Experiments were conducted on the multichannel coherent radar depth sounder to test the robustness of the proposed method. We carried out a comparative analysis with the state-of-the-art unsupervised and supervised segmentation methods. MIoU is improved by 23.47%.**

***Index Terms*—Multichannel coherent radar depth sounder (MCoRDS), radar sounder, semantic segmentation, sequence-to-sequence model, TransFuse, TransUNet, transformers.**

I. Introduction

RADAR sounders (RSs) are sensors with active sensing capabilities operating on nadir-looking geometry to transmit linearly modulated electromagnetic (EM) pulses and receive the reflected echoes from the subsurface targets depending upon geophysical characteristics [1]. The operating frequency of these sensors ranges from high frequency (HF) to very high frequency (VHF) [2]. To generate radargrams, the received echoes are coherently integrated by several processing tasks ranging from synthetic aperture radar (SAR) focusing to correct for platform instability, elevation variations, etc. [1]. The radargrams are exploited to characterize the subsurface geophysical targets by automatic techniques. The advantage of RS data is given by the possibility to capture subsurface information up to several hundred meters in the collaborative medium as ice, where SAR sensors show limited penetration capability of a few centimeters in comparable situations. In contrast, optical sensors do not show subsurface penetration capabilities. Thus, RS sensors have unique properties, allowing for the characterization of subsurface features which is not feasible with other remote sensing sensors.

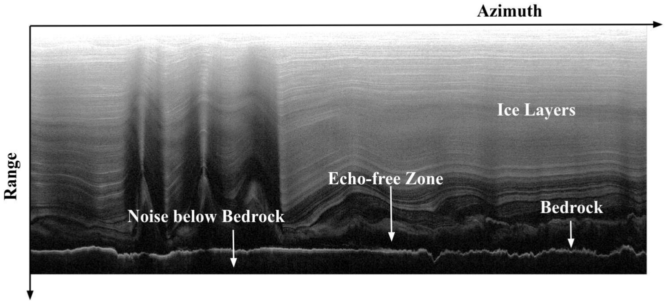


Fig. 1. Sample MCoRDS Radargram with different subsurface targets.

In recent years, the Earth's climate change generated a great deal of attention with miscellaneous research directions to characterize the status of environmental parameters for tackling alarming situations [3]. A significant loss of polar ice sheets has been observed due to global temperature rise along with the unprecedented fluctuations of climate variables contributing to climate change [4]. The subglacial hydrology directly influences the motion of the ice along with the topographic characteristics [2]. The land surface temperature affects the internal stability of the ice layers and is intricately related to the dynamics of the basal temperature [5], [6]. The ice layers depict quasi-linear homogeneous characteristics in the radargram (see Fig. 1). Another important geophysical subsurface component is bedrock, often referred to as the deepest scattering area.

Received 23 May 2024; revised 19 July 2024; accepted 19 August 2024. Date of publication 26 August 2024; date of current version 10 September 2024. This work was supported by the Italian Space Agency through the “Missione JUICE - Attivitã dei team scientifici dei Payload per Lancio, commissioning, operazioni e analisi dati,” Agenzia Spaziale Italian-Istituto Nazionale di Astrofisica (ASI-INAF), under Contract ASI-INAF n. 2023-6-HH.0. *(Corresponding author: Francesca Bovolo.)*

Raktim Ghosh is with the Center of Digital Society, Fondazione Bruno Kessler, 38123 Trento, Italy, and also with the Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy (e-mail: raghosh@fbk.eu).

Francesca Bovolo is with the Center of Digital Society, Fondazione Bruno Kessler, 38123 Trento, Italy (e-mail: bovolo@fbk.eu).

Digital Object Identifier 10.1109/JSTARS.2024.3447879

Due to the significant attenuation of the transmitted wave, the radar measures the noise at a depth greater than the bedrock. Another subsurface region showing a noise-like signature is the echo-free zone (EFZ) often situated above the bedrock region (see Fig. 1). EFZ can be utilized as a proxy for paleoclimatic research, change in flow behaviors, and ice sheet dynamics [7]. Therefore, the semantic segmentation of radargrams has miscellaneous application in the context of geophysical research, such as glaciological dynamics of subsurface targets (e.g., ice layers, bedrock, noise, etc.). Further, radargrams can be exploited for postsemantic segmentation tasks like tracking ice layers, measuring the thickness, etc. The major difficulty lies in investigating the subsurface dynamics of these targets due to the inaccessibility of the subsurface environment [3]. In this regard, nadir-looking RS sensors and radar depth sounder (RDS) provide reliable scientific information to monitor and investigate the subglacial environment without an in situ survey. The RS signals depict sequential linear structures (associated with different subsurface targets) often corrupted by the multiplicative nature of noise (see Fig. 1). Therefore, modeling the local spatial contexts (i.e., local spatial correlations among the small neighbourhoods in radargrams) and the global spatial contexts (i.e., long-range back-and-forth information processing system between different spatial locations in radargrams [8]) is pivotal for the deep learning architectures to segment radargrams.

Very popular CNN-based methods, such as U-Net [9], FPN [10], FCN [11], SegNet [12] dominated the supervised segmentation architectures in the domain of Earth Observation (EO). Considering the RS sensors for EO, the authors in [13], [14], [15], [16], and [17] developed CNN-based architectures for applications, such as segmenting subsurface targets, tracking ice layers, estimating ice thickness, etc. Wang et al. [13] developed a joint multi-GAP RNN augmented with the triple-task CNN architecture for detecting and numerically quantifying the ice layers and the corresponding thickness. Cai et al. [14] utilized a SegCaps network architecture to segment the radargrams. Varshney et al. [15] utilized different FCN-based architectures for tracking and estimating the thickness of ice layers. Liu-Schiaffini et al. [16] developed a novel CNN architecture (refined with continuous CRF) to identify the ice bed interface in the radargrams. Donini et al. [17] utilized an attention UNet architecture for segmenting the radargrams. Recently, the authors in [18] and [19] utilized the semisupervised segmentation models for SAR segmentation. However, convolutions capture the local spatial contexts, whereas radar sounder depicts global spatial contexts with respect to the depth which is often difficult to model by CNNs. To mitigate the problems, Ghosh and Bovolo [20] proposed a hybrid CNN-Transformer architecture for the semantic segmentation of radargrams by utilizing TransUNet [21] and TransFuse [22] architecture. To reduce the number of parameters in Transformers, Ghosh and Bovolo [23] proposed an FFT-based unparameterized self-attention mechanism to segment the radargrams. Recently, Ibikunle et al. [24] developed a Echogram Vision Transformer with miscellaneous patchifying scheme with learnable positional embedding to track the individual ice layers in radargrams. Despite the success in supervised settings, these architectures require a significant amount of labeled samples. However, acquiring the labeled information in radargrams requires drilling and/or strong domain expertise which are often lacking.

In contrast, unsupervised segmentation methods learn the inherent semantic contents from the training samples without using any label. A decomposition is then performed on the learned semantic contents with the known ontology across the image corpora [25], [26]. In the domain of computer vision, unsupervised semantic segmentation has been performed based on contrastive clustering [27], mutual information minimization [25], independent information clustering (IIC) [28], generative modeling approaches [29], etc. Caron et al. [30] proposed a novel clustering approach by utilizing the unsupervised training and iterative refinement of weights from pretrained CNN-based networks (convnets). Several variants of clustering methods have been proposed that incorporate miscellaneous pretext tasks for unsupervised feature learning [31], [32], [33]. Contrary to the pretext hypothesis, the contrastive methods leverage the learning of discriminative pixel-level embeddings by maximizing the agreement between positive pairs and minimizing between the negative pairs [34]. Gansbeke et al. [27] proposed a two-step framework by utilizing a-priori contrastive object mask proposal for generating the meaningful pixel-level discriminative contexts for the downstream segmentation tasks. These approaches showed good performance in computer vision, however, less research works have been carried out in segmenting the EO data. Among them, Saha et al. [35] proposed an unsupervised segmentation method (US^{4}EO) by combining deep clustering with the contrastive learning approach by refining weights iteratively from a CNN-based two-stream deep feature learning framework for segmenting EO dataset. In radargrams unsupervised segmentation, Donini et al. [36] established a two-stream teacher–student network by incorporating two parallel UNet-like networks. In the domain of hyperspectral remote sensing, Pérez-García et al. [37] utilized a novel spectral loss function by incorporating a 3-D convolutional encoder to segment the hyperspectral images. Mou et al. [38] utilized a fully conv–deconv framework in the unsupervised spectral-spatial feature learning paradigm. However, hyperspectral images contain very high-dimensional spectral information which can be exploited for the unsupervised tasks with miscellaneous framework associated to the spectral-spatial domain, whereas radagrams do not contain spectral information, thereby making the unsupervised segmentation task challenging. Further, it is often difficult for the unsupervised CNN-based approaches to address the multiplicative nature of noise embedded in radargrams. Also, convolutions amplify the HF components during training and often obfuscate the low-frequency components embedded in radargrams. Although the authors in [35] and [36] demonstrated the capability of unsupervised segmentation methods in EO, however, in the radargrams, the difficulty lies in modeling the global contexts as convolutions inherently impose spatial locality constraints. Further, these methods require a significant number of parameters while training from scratch.

In contrast to CNN-based unsupervised methods in computer vision, Caron et al. [39] established a novel self-supervised feature learning framework (named DINO) to demonstrate that the self-supervised Vision Transformers (ViTs) explicitly embed rich semantic contents which are not often emerged in the supervised ones. By utilizing the aforementioned DINO as a pretrained frozen visual featurizer, Hamilton et al. [25] established an unsupervised semantic segmentation method named self-supervised Transformer with energy-based graph optimization (STEGO) by incorporating a novel contrastive loss function and exploiting the interfeature and intrafeature correlation across discriminative contexts extracted from the positive and negative samples. In STEGO, the k-nearest neighbors (KNN) have been incorporated to select the samples as positive descriptors for generating the correlation volume. However, using KNN as positive descriptors creates instability during training with the radargrams while establishing the discriminative embeddings. Melas-Kyriazi et al. [26] established a novel framework of deep spectral segmentation methods by discretely segmenting the eigenvectors of a Laplacian derived from a feature affinity matrix of intermediate deep features extracted from self-supervised pretrained networks. Overall, the authors in [25], [26], and [39] demonstrated that deep features extracted from self-supervised ViTs can be potentially utilized for the unsupervised semantic segmentation tasks with light-weight parameterizations.

In summary, while considering unsupervised learning framework, several drawbacks can be highlighted in the context of segmenting the radargrams with the existing architectural settings. First, while convolutions in the encoder accurately capture local spatial contexts, they often fail to establish global spatial contexts within the radargrams. Despite amplifying HF components in feature tensors, convolutions may overlook crucial low-frequency components during training. Although the ice layers exhibit linear sequential features across the azimuth, the backscattering response with associated linearity becomes corrupted due to the multiplicative nature of noise as the signal attenuates along the range direction in radargrams. Consequently, the heterogeneity increases in the linear characteristics of the ice layers as the depth increases. Therefore, effectively modeling the rich contextual information embedded in radargrams necessitates careful consideration of the varying transitions between HF and low-frequency regions. Second, while considering the unsupervised contrastive learning framework in computer vision [25], [27], the negative samples depict class separability and spatial variability for the surface features in optical images. The challenge arises when the negative radargram samples are not heterogeneous in terms of subsurface features, thereby impending the difficulty in establishing an attractive–repulsive contrastive optimization strategy in purely unsupervised settings. Thus, addressing the issue of inadequate spatial heterogeneity between the negative and positive descriptors associated with the radargrams is crucial in designing contrastive learning framework. Third, miscellaneous randomness (random shuffling of samples, selecting positive and negative pairs, etc.) is introduced into the unsupervised contrastive framework during training. Therefore, it is equally important to assess and constrain the degree of randomness while establishing the attractive–repulsive discriminative contexts between the positive and negative pairs. By addressing the miscellaneous limitations of convolutions, addressing the properties of noise in radargrams, and the issues related to incorporating contrastive framework in radargrams segmentation, leveraging self-supervised ViTs can be a crucial step forward in the domain of RS. Concretely, no work explored the potential of self-supervised ViTs for unsupervised semantic segmentation in radargrams.

By incorporating the self-supervised ViTs as a deep featurizer for the input RS signals, the similarity of augmented view, and utilizing a hybrid loss function, we develop an unsupervised radargram segmentation (URS) architecture. The key contributions are twofold as follows.

1) We develop an unsupervised architecture by exploiting the components of STEGO along with an expansive network as a proxy decoder to reconstruct the rich discriminative embeddings of the RS signals. The training strategy is based on bootstrapping the randomness inside the training samples for developing a stable attractive– repulsive coupling between the positive-negative pairs for the radargrams.
2) We design a loss function by integrating the spatial similarity measure between the input RS signals with the reconstructed RS signals from the latent space along with the contrastive correlation loss to steer the learning optimization with stable gradient flow.

The rest of this article is organized as follows. Section II depicts the concrete mathematical details of the proposed methodology. Section III reports the results and corresponding discussions. Finally, Section IV concludes this article.

II. Proposed Methodology

The proposed method develops an unsupervised architecture by utilizing the self-supervised ViTs into a two-stream contrastive learning framework by exploiting jointly the contrastive correlation loss along with the spatial similarity loss.

A. Problem Formulation

We denote a radargram as a 2-D single channel matrix. The backscattered returns from the different spatial positions along with the channel information can be denoted as

$$R=\{R(C,i,j)|C=1, i\in P=[1,\ldots,n_T], j\in Q=[1,\ldots,n_S]\} \tag{1}$$

where C denotes the channel dimension, $[1,\ldots,n_T]$ denotes the acquired number of samples in the along-track direction and $[1,\ldots,n_S]$ denotes the number of samples in the range direction. A concrete mathematical treatment of the proposed architecture is depicted as follows.

Let us denote N as a number of radargram training samples $\mathcal{T}=\{X_1,X_2,\ldots,X_N\}$ where the dimension of each X_i is $C\times H\times W$ $(i\in\{1,2,\ldots,N\})$. The goal of this research work is to perform radargram unsupervised semantic segmentation and to assign each pixel (C,i,j) a distinct class. The unsupervised segmentation method aims at establishing discriminative pixel embedding with deep feature learning framework to cluster pixels associated to targets without using any labeled information. In our case, the known ontology in the radargram can be grouped into three different categories: 1) ice layers, 2)

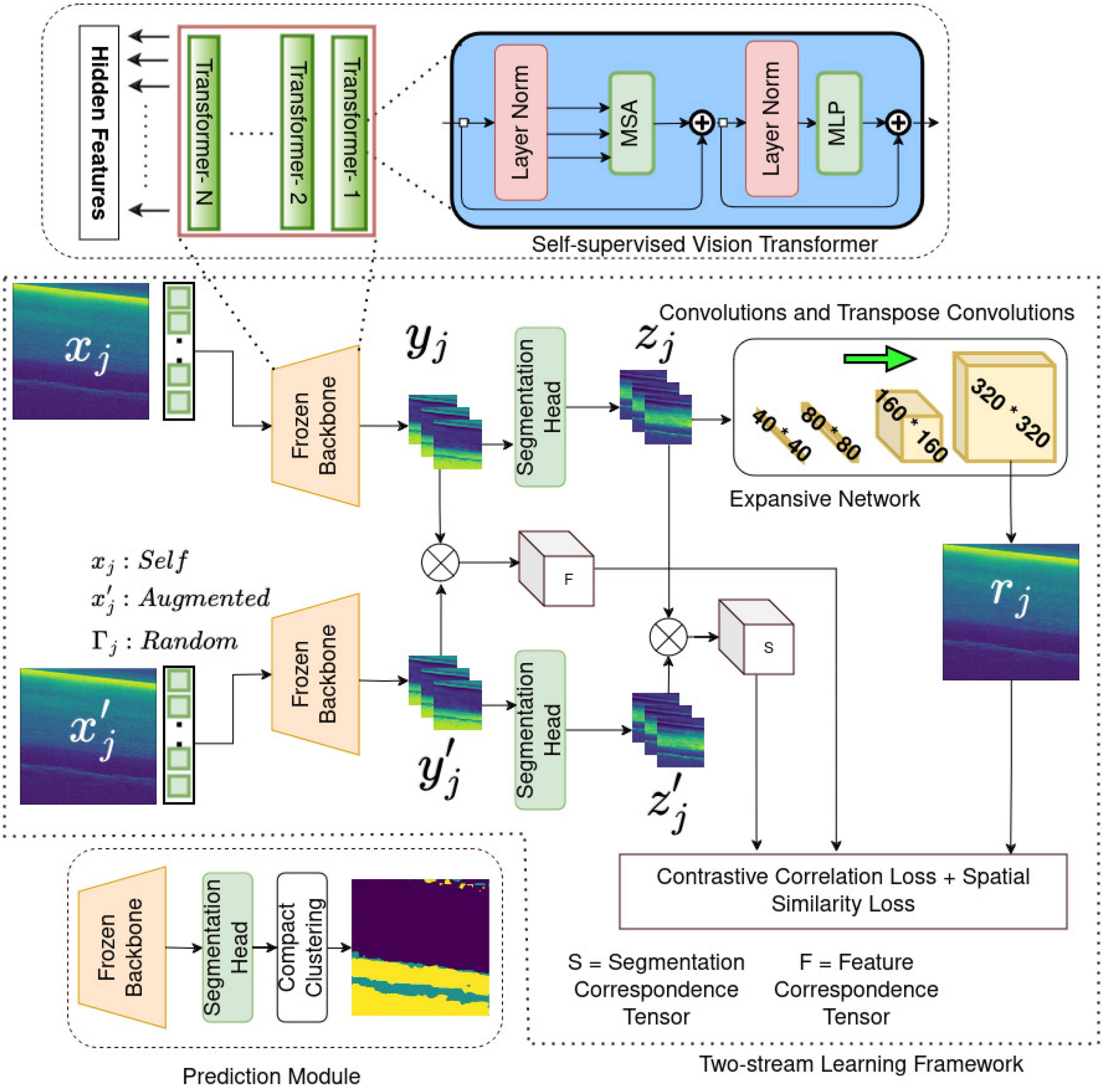


Fig. 2. Schematic layout of proposed architecture.

bedrock, and 3) noise (composed of EFZ above the bedrock region and the noise at the depth greater than the bedrock).

B. Proposed URS Architecture

Fig. 2 depicts a two-stream schematic layout of the proposed unsupervised URS architecture. From one branch, the self-supervised transformer-based pretrained frozen visual backbone extracts rich discriminative semantic contents embedded in tensors from the input radargram patches. From the other branch, the pretrained backbone extracts the semantic contents from the augmented view of the unlabeled training samples. At this stage, the feature correspondence tensor is computed by estimating the 1) self-correlation, 2) correlation with the augmented view, and 3) correlation with the random samples. The outputs from the frozen visual backbones are subsequently fed into the lightweight segmentation head to amplify the patterns embedded in the feature tensors. After that, the segmentation correspondence tensors are computed between the output generated from the segmentation head of two branches. By utilizing feature correspondence tensors and segmentation correspondence tensors, the contrastive correlation loss can be computed for optimizing the network. On the other hand, the expansive network incorporates a progressive upsampling by utilizing the successive convolutions and transpose convolution operations to reconstruct the input tensors from the intermediate feature tensors from the segmentation head. A spatial similarity is measured between the input tensor and reconstructed tensor.

Finally, the contrastive correlation loss and the spatial similarity loss is coupled to steer the gradient optimization for the unsupervised feature learning framework. During the prediction time, the frozen visual backbone along with the weights from the learned segmentation head is utilized to generate the intermediate feature tensors. After that, the clustering operation is performed to create pseudo labels for generating the final prediction maps.

1) Training Strategy: Let the radargram samples in $\mathcal{T} = \{X_1, X_2, \ldots, X_N\}$ be randomly shuffled and the new shuffled sequence be denoted as $\mathcal{T}_{\text{shf}} = \{x_1, x_2, \ldots, x_N\}$ without loss of generality. Let $\mathcal{B}_\alpha$ ($\alpha \in \{1, 2, \ldots, \lceil N/K \rceil\}$) denote a batch consisting of k samples selected from the shuffled training sequence $\mathcal{T}_{\text{shf}}$. For consistency of notations, we will carry out all the mathematical treatment on

$$\mathcal{B}_\alpha = \{x_j \in \mathcal{T}_{\text{shf}} | K(\alpha - 1) + 1 \leq j \leq K\alpha, \\ \forall \alpha \in \{1, \ldots, \lceil N/K \rceil\}\}. \tag{2}$$

Let $\mathcal{A} : \mathbb{R}^{\text{CHW}} \rightarrow \mathbb{R}^{\text{CHW}}$ be a function that creates an invariant augmented view of every x_j by keeping the spatial and channel dimensions invariant. Therefore, the set of all augmented x_j in $\mathcal{B}_\alpha$ can be denoted as

$$\mathcal{B}'_\alpha = \{x'_j \in \mathcal{T}'_{\text{shf}} | K(\alpha - 1) + 1 \leq j \leq K\alpha, \\ \forall \alpha \in \{1, \ldots, \lceil N/K \rceil\}\}. \tag{3}$$

Let us denote the shuffled sequence on $\mathcal{B}_\alpha$ as

$$\mathcal{B}^{\text{rnd}}_\alpha = \{\Gamma_j \in \mathcal{B}_\alpha | K(\alpha - 1) + 1 \leq j \leq K\alpha, \\ \forall \alpha \in \{1, \ldots, \lceil N/K \rceil\}\}. \tag{4}$$

Each x_j in $\mathcal{B}_\alpha$ has a triplet associative mapping with the 1) Self ($x_j \in \mathcal{B}_\alpha$), 2) Augmented ($x'_j \in \mathcal{B}'_\alpha$), and 3) Random ($\Gamma_j \in \mathcal{B}^{\text{rnd}}_\alpha$) during training. Γ_j is selected randomly from the batch $\mathcal{B}^{\text{rnd}}_\alpha$ to understand the contrast between x_j and Γ_j. The prepared dictionary training triplets for the batch $\mathcal{B}_\alpha$ are

$$\mathcal{D}_{tr} = \{x_{K(\alpha-1)+1} : (x_{K(\alpha-1)+1}, x'_{K(\alpha-1)+1} \\ \Gamma_{K(\alpha-1)+1}), x_{K(\alpha-1)+2} : (x_{K(\alpha-1)+2}, x'_{K(\alpha-1)+2} \\ \Gamma_{K(\alpha-1)+2}), \ldots, x_{K\alpha} : (x_{K\alpha}, x'_{K\alpha}, \Gamma_{K\alpha}) \\ \forall \alpha \in \{1, \ldots, \lceil N/K \rceil\}\}. \tag{5}$$

Mathematically, for every x_j, Self: x_j and Augmented: x'_j are the positive descriptors, and Random: Γ_j is a negative descriptor for the triplet associative mapping. Here, a degree of randomness can affect optimization when choosing the positive and negative descriptors from the mini-batches associated with the random shuffling. In order to bootstrap the aforementioned randomness, the Gaussian blur is incorporated in the input radargram training sample x_j to generate x'_j as positive descriptors for the learning framework. By utilizing Gaussian blur as positive descriptor, the original structure of the input radargram training sample is retained while contrasting between the positive and negative descriptors. These associative mappings between the triplets in $\mathcal{D}_{\text{tr}}$ allow the URS architecture to understand the discriminative pixel embeddings in input radargrams by contrasting the attractive and repulsive interactions in congruence to correlation strengths between the elements in the $\mathcal{D}_{\text{tr}}$ triplets.

After preparing these triplets in $\mathcal{D}_{\text{tr}}$ for contrastive learning, a two-stream learning framework (see Fig. 2) has been constructed to create the discriminative pixel embedding with dense intermediate visual representations. Let $\mathcal{N} : \mathbb{R}^{\text{CHW}} \rightarrow \mathbb{R}^{C'H'W'}$ be a frozen backbone that maps every element x_j in $\mathcal{B}_\alpha$, and x'_j in B'_α to higher dimensional feature spaces with rich discriminative semantic contents for the segmentation. Let us denote these higher dimensional tensors generated by $\mathcal{N}$ are: $y_j = \mathcal{N}(x_j), y'_j = \mathcal{N}(x'_j)$, with the dimension $C' \times H' \times W'$ ($H' < H$, $W' < W$, and $C' >> C$). As the backbone $\mathcal{N}$ is frozen, it is necessary to drive the learning gradient with a lightweight network to amplify the embedded high-frequency and low-frequency components of the intermediate dense feature tensors y_j and y'_j and keep the number of parameters and the computational time low. Further, the amplified correlation patterns are utilized for cluster compactifications to assign pseudo labels. Therefore, similar to [25], we utilize a segmentation head to map the tensors to a lower dimensional pixel embedding. We denote the segmentation head as $\mathcal{S} : \mathbb{R}^{C'H'W'} \rightarrow \mathbb{R}^{C''H'W'}$, where $C'' < C'$ and $\mathcal{S}$ is built with the small convolutional block. Let us denote $z_j = \mathcal{S}(y_j)$ and $z'_j = \mathcal{S}(y'_j)$ as the tensors generated from $\mathcal{S}$ by inputting y_j and y'_j, respectively. The corresponding elements y_j, y'_j, z_j, and z'_j are utilized to estimate the feature correspondence tensors and segmentation correspondence tensors, respectively at the later stage of the architecture.

C. Expansive Networks

After incorporating the segmentation head $\mathcal{S}$ on a two-stream network, z_j is fed to an expansive network to reconstruct the tensors with a similar spatial dimension as the input radargram tensors x_j. Since the frozen backbone performs a significant resolution reduction (approximately by a factor of 8), the spatial reconstruction operation leverages to amplify the local spatial details embedded in the radargram. The expansive network is denoted by $\mathcal{E} : \mathbb{R}^{C''H'W'} \rightarrow \mathbb{R}^{\text{CHW}}$. The corresponding mathematical equations can be denoted as $r_j = \mathcal{E}(z_j)$. The expansive network incorporates successive convolution and Transpose convolution operations, respectively. These aforementioned operations are similar to the symmetric expanding path as used in a decoder of UNET architecture. We tabulate the parametric settings of successive convolutions and Transpose convolutions operations of expansive networks in Table I.

D. Feature Correspondence Tensors

In self-supervised settings, the intermediate dense features often embed rich semantic information for downstream tasks, such as object localization, semantic segmentation, edge detection, etc. The intermediate dense feature tensors can be extracted from activation maps of deep convolutional layers of a network, or queries, keys, and values matrices from intermediate layers of self-supervised ViT-based architecture. The feature correspondence tensors estimate the correlation volume as similar to [40].

Algorithm 1: Unsupervised Feature Learning Framework for Semantic Segmentation of Radar Sounder Data.

```
Require: $N$ number of training samples $\mathcal{T} = \{X_1, \ldots, X_N\}$
Ensure: Randomly Shuffle $\{X_1, X_2, \ldots, X_N\}$ to generate
  a shuffled sequence denoted as $\mathcal{T}_{shf} = \{x_1, \ldots, x_N\}$ for
  every batches $\mathcal{B}_\alpha \in \mathcal{T}_{shf}$
  for $i \leftarrow 1$ to $\mathcal{I}$ do
    [$\mathcal{B}_1 : \{x_1, x_2, \ldots, x_K\}, \mathcal{B}_2 : \{x_{K+1}, x_2, \ldots, x_{2K}\}, \ldots, \mathcal{B}_{N/K} : \{x_{N-K+1}, x_{N-K}, \ldots, x_N\}$]
    Gaussian Blur on
      $\mathcal{B}_\alpha = \{x_{K(\alpha-1)+1}, x_{K(\alpha-1)+2}, \ldots, x_{K\alpha}\}$ to create
      $\mathcal{B}'_\alpha = \{x'_{K(\alpha-1)+1}, x'_{K(\alpha-1)+2}, \ldots, x'_{K\alpha}\}$
    Create dictionary triplets for each $x_j$ in $\mathcal{B}_\alpha$ to form
      $\mathcal{D}_{tr} = \{x_{K(\alpha-1)+1} : (x_{K(\alpha-1)+1}, x'_{K(\alpha-1)+1}, \Gamma_{K(\alpha-1)+1}), x_{K(\alpha-1)+2} : (x_{K(\alpha-1)+2}, x'_{K(\alpha-1)+2}, \Gamma_{K(\alpha-1)+2}), \ldots, x_{K\alpha} : (x_{K\alpha}, x'_{K\alpha}, \Gamma_{K\alpha})\}$.

    for $j \leftarrow 1$ to $K$ do
      $y_j = \mathcal{N}(x_j)$
      $y'_j = \mathcal{N}(x'_j)$
      $z_j = \mathcal{S}(y_j)$
      $z'_j = \mathcal{S}(y'_j)$
      $r_j = \mathcal{E}(z_j)$
    end for
    Estimate Feature Correspondence Tensor $F_{\text{hwmn}}$ using $y_j$
      and $y'_j$
    Estimate Segmentation Correspondence Tensor $S_{\text{hwmn}}$
      using $z_j$ and $z'_j$
    Estimate Contrastive Correlation Loss - $\mathcal{L}_1$
    Estimate Spatial Similarity Loss - $\mathcal{L}_2$
    Estimate Cluster Compactification Loss - $\mathcal{L}_3$
    Utilize the Losses to optimize the network parameters
  end for
```

Here, the correlation volume is constructed by estimating a 4-D tensor while considering the individual correlations among all the pairs between the high-dimensional latent space generated from y_j and y'_j, respectively. We denote the correlation volume as $\mathcal{F}_{\text{corr}}$. Let us denote a 1×1 subset extracted from y_j as Y_{chw} and from y'_j as Y_{cmn} with the channel dimension as c, and individual spatial location as (h, w) and (m, n), respectively. Therefore, the corresponding equations of individual correlations of these tensors can be depicted as

$$F_{\text{hwmn}} = \sum_c \frac{Y_{\text{chw}}}{|Y_{hw}|} \frac{Y'_{\text{cmn}}}{|Y'_{mn}|}. \tag{6}$$

Equation (6) estimates the dot product or cosine similarity between Y_{chw} and Y'_{cmn} to generate feature correlation tensors at each spatial position of y_j with respect to the other spatial position of y'_j. In special cases, the similarity is measured between the two regions of the same tensors ($Y_{\text{chw}} = Y'_{\text{cmn}}$). The tensor F_{hwmn} corresponds to the generalization of higher order class activation maps [41] as a singular component of $\mathcal{F}_{\text{corr}}$. In order to reduce the computational complexity of 4-D $\mathcal{F}_{\text{corr}}$, a stratified grid sampling has been incorporated to reduce the spatial dimension of the intermediate feature tensors (y_j, y'_j) from $H' \times W'$ to $\mathcal{G}_1 \times \mathcal{G}_2$. The elements in $\mathcal{F}_{\text{corr}}$ are correlated with the cooccurrence of true labels.

E. Segmentation Correspondence Tensors

Similar to the $\mathcal{F}_{\text{corr}}$, estimations of the correlation volume are performed on intermediate feature tensors z_j, and z'_j generated by the segmentation head $\mathcal{S}$. We represent the subset of 1×1 tensor in z_j as Z_{chw}, and in z'_j as Z_{cmn}, respectively, with the channel dimension as c, and individual spatial location as (h, w), and (m, n), respectively. Mathematically, the correlation between Z_{chw} and Z_{cmn} in the segmentation correspondence tensor ($\mathcal{S}_{\text{corr}}$) can be written as

$$S_{\text{hwmn}} = \sum_c \frac{Z_{\text{chw}}}{|Z_{hw}|} \frac{Z'_{\text{cmn}}}{|Z'_{mn}|}. \tag{7}$$

As similar to the F_{hwmn}, S_{hwmn} embed the correlation values. Similar to the behavior of the attractive–repulsive duality, the contrastive polarity is established in terms of the correlation strengths embedded at different spatial positions of z_j and z'_j extracted from input radargrams. In other words, the positively similar couples will create positive pressure to drag the optimization toward the highest correlation values and vice-versa.

F. Distillation of Feature and Segmentation Correspondences

By utilizing the correlation values of the discriminative pixel embedding, a contrastive framework of attraction–repulsion can be established by utilizing the correlation strength at different spatial positions. The dot-product pushes the entries together if there is a significant amount of positive couplings and pulls apart the entries if the couplings are significantly negative. By utilizing F_{hwmn} and S_{hwmn}, a correlation loss function can be established as

$$\mathcal{L}_a(x_j, x'_j, b) = \sum_{\text{hwmn}} -(F_{\text{hwmn}} - b) S_{\text{hwmn}}. \tag{8}$$

Here, the strength of dense feature correspondences at different spatial positions is established between F_{hwmn} and S_{hwmn} while creating the dual attractive–repulsive pair. The b parameter plays a crucial role in preventing the collapse of the gradient flow to establish the equilibrium between the dual attractive–repulsive combinations. In other words, the $F_{\text{hwmn}} - b$ determines the sign of positivity and negativity, and the values S_{hwmn} further drive the optimization toward antialignment or alignment depending upon the sign of $F_{\text{hwmn}} - b$. To balance the optimization further, the spatial centering (SC) operation is incorporated

$$F^{\text{SC}}_{\text{hwmn}} = F_{\text{hwmn}} - (1/H'W') \sum_{m'n'} F_{hwm'n'}. \tag{9}$$

To decrease the colinearity between the weakly correlated patterns concentrated in the feature and segmentation correspondence tensors, zero-clamping has been utilized for establishing

TABLE I
ARCHITECTURAL SETTINGS FOR EXPANSIVE NETWORK OF PROPOSED METHOD

CNN Upsampling	**In. Ch**	**Out. Ch**	**In. Dim**	**Out. Dim**	**Kernel Size**
Transpose Convolutions	192	96	40×40	80×80	2×2
Double Convolutions	96	96	80×80	80×80	3×3
Transpose Convolutions	96	48	80×80	160×160	2×2
Double Convolutions	48	48	160×160	160×160	3×3
Transpose Convolutions	48	24	160×160	320×320	2×2
Double Convolutions	24	24	320×320	320×320	3×3
Single Convolution	24	1	320×320	320×320	1×1

orthogonality. The equation can be depicted as

$$\mathcal{L}_a(x_j, x_j', b) = -\sum_{\text{hwmn}} (F^{\text{SC}}_{\text{hwmn}} - b)\max(S_{\text{hwmn}}, 0). \tag{10}$$

G. Loss Functions

1) Contrastive Correlation Loss: The contrastive correlation loss drives the optimization between every triplet x_j, x_j', and Γ_j with three types of correlations: 1) self-correlation (between x_j and x_j), 2) correlation with the augmented view (between x_j and x_j'), and 3) correlation with the random samples (between x_j and Γ_j). Therefore, the linear combination between different correlations can be depicted as

$$\begin{aligned}\mathcal{L}_1 = \lambda_s \mathcal{L}_s(x_j, x_j, b_s) + \lambda_a \mathcal{L}_a(x_j, x_j', b_a) \\ + \lambda_r \mathcal{L}_r(x_j, \Gamma_j, b_r)\end{aligned} \tag{11}$$

where the weights $\lambda_s, \lambda_a,$ and λ_r are assigned to balance the dual attractive–repulsive interactions between different correlation strengths. Further, $b_s, b_a,$ and b_r direct the change of sign depending upon the correlation strength of different feature correspondence tensors. While estimating the contrastive correlation loss, in order to reduce the computation complexity of F_{hwmn} and S_{hwmn}, a stratified grid sampling is utilized to reduce the spatial size of the feature tensors to a lower dimensional space. For multiplication between the feature correspondence tensor and segmentation correspondence tensor, the randomly selected coordinates of F_{hwmn} and S_{hwmn} are kept similar during self-correlation.

2) Spatial Similarity Loss: The spatial similarity loss measures the difference between the pixelwise similarity between the input tensor x_j, and the reconstructed tensors r_j. A mean-absolute error between x_j and r_j, can be depicted as

$$\mathcal{L}_2 = ||x_j - r_j||. \tag{12}$$

It was observed during the learning process that if we combine the spatial similarity loss with the contrastive correlation loss, the learning optimization becomes convergent toward decomposing the distinct segments more effectively, especially for the radar sounder data.

If we combine the contrastive correlation loss ($\mathcal{L}_1$) and spatial similarity loss ($\mathcal{L}_2$), the mathematical expression of the joint loss function can be defined as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_1 + \mathcal{L}_2. \tag{13}$$

The intuition behind summing $\mathcal{L}_1$ and $\mathcal{L}_2$ is that $\mathcal{L}_1$ uses the correlation values of the discriminative pixel embeddings to establish a framework for attraction–repulsion strategies based on the correlation strength between every triplet x_j, x_j', and Γ_j in various spatial positions of the high-dimensional feature tensors. The objective of $\mathcal{L}_1$ is to establish a discriminative class separability in unsupervised learning. However, $\mathcal{L}_1$ does not consider the resolution reductions when extracting rich discriminative semantic contents from the pretrained frozen visual featurizer. To address the issue of resolution reductions during optimization, $\mathcal{L}_2$ is taken into account. While $\mathcal{L}_2$ addresses the resolution reductions, it does not inherently leverage the contrastive optimization strategy to distinguish between different ontologies. As a result, $\mathcal{L}_1$ and $\mathcal{L}_2$ complement each other to enable efficient optimization strategy in purely unsupervised settings.

3) Cluster Compactification Loss: During training, the cluster compactification loss is estimated by utilizing the Einstein summation between the randomly initialized weight matrix w_{nc}, with Z_{chw} derived from $\mathcal{S}$. Mathematically, the product can be written as $K_{\text{nhw}} = \sum_c w_{nc} Z_{\text{chw}}$. To create the pseudo labels for the prediction module, one-hot encoding is performed after taking the argmax on the dimension of the specified clusters n on K_{nhw}, depending on the number of classes to be detected in the final prediction. Let us denote the updated tensor as K'_{nhw}. The cluster loss is then computed by following:

$$\mathcal{L}_3 = -\sum_{h,w} K'_{\text{nhw}} K_{\text{nhw}}. \tag{14}$$

The weights of w_{nc} are updated iteratively with respect to pseudo labels created in K'_{nhw}. Note that, the gradient optimization of $\mathcal{L}_3$ is performed separately from the loss $\mathcal{L}_1$ and $\mathcal{L}_2$.

H. Prediction Module

In the prediction module (bottom left corner in Fig. 2), three consecutive blocks are utilized for the final inference of the unsupervised network. At first, the test samples are fed to the self-supervised ViTs-based frozen visual backbone. After that, the tensors are fed through the learned segmentation network. Finally, these learned intermediate feature spaces are upsampled to the required channel and spatial dimension and decomposed to the number of segments with respect to the learned weights.

III. Experimental Results

In this section, we report the results and associated comparative analysis between both unsupervised and supervised deep learning architectures for semantic segmentation of the radargrams. We first elucidate upon the description of the dataset. Next, we construct the experimental setup with associated resources. Finally, the segmentation results with quantitative and qualitative analysis are interpreted with some discussions.

A. Dataset

Experiments are conducted on data from multichannel coherent radar depth sounder (MCoRDS) owned by the Centre of Remote Sensing of Ice Sheets (CReSIS) unit. Two bandwidths, i.e., 9.5 and 30 MHz are utilized to acquire the dataset by different sensors with a central frequency of 193.5 MHz. The instrument was on-boarded on DC-8 aircraft. The acquisition took place over several regions of Antarctica with an altitude of 7000 m. The campaign was conducted on November 2010, of which 8 radargrams (bandwidth 9.5 MHz) were utilized for our experimental setup. The coordinates of acquisition ranges between ($-86°00'$N to $-15°67'$E) to ($-86°02'$N to $29°45'$E). A total of 27 350 traces are acquired over a distance of about 400 km. For improving the range resolution, pulse compression techniques are used to suppress the sidelobe level [42]. To improve the along-track resolution and suppress the contributions of the clutter further, SAR focusing techniques with the multilooking approach are utilized with 1 look in the cross-track direction and 11 looks in the along track direction. Therefore, the final range resolution and along track resolution is 13.6 and 25 m, respectively. To characterize the radargram, three classes are chosen: 1) ice layers (blue), 2) bedrock (green), and 3) noise (yellow). The ambiguous pixels (blue in the reference map) are not accounted for measuring the final prediction accuracy, however, these pixels are labeled from the trained network in supervised settings. In supervised settings, the deep networks utilize the location of ambiguous pixels to capture the semantic contexts with learnable parameters of filters, however, these locations are discarded while estimating the final loss. In unsupervised settings, we do not have access to the label information in any location including ambiguous pixels, therefore the losses were computed by taking into account the overall spatial extents of the discriminative contexts embedded in the latent space. Please see [42] for further details about the dataset.

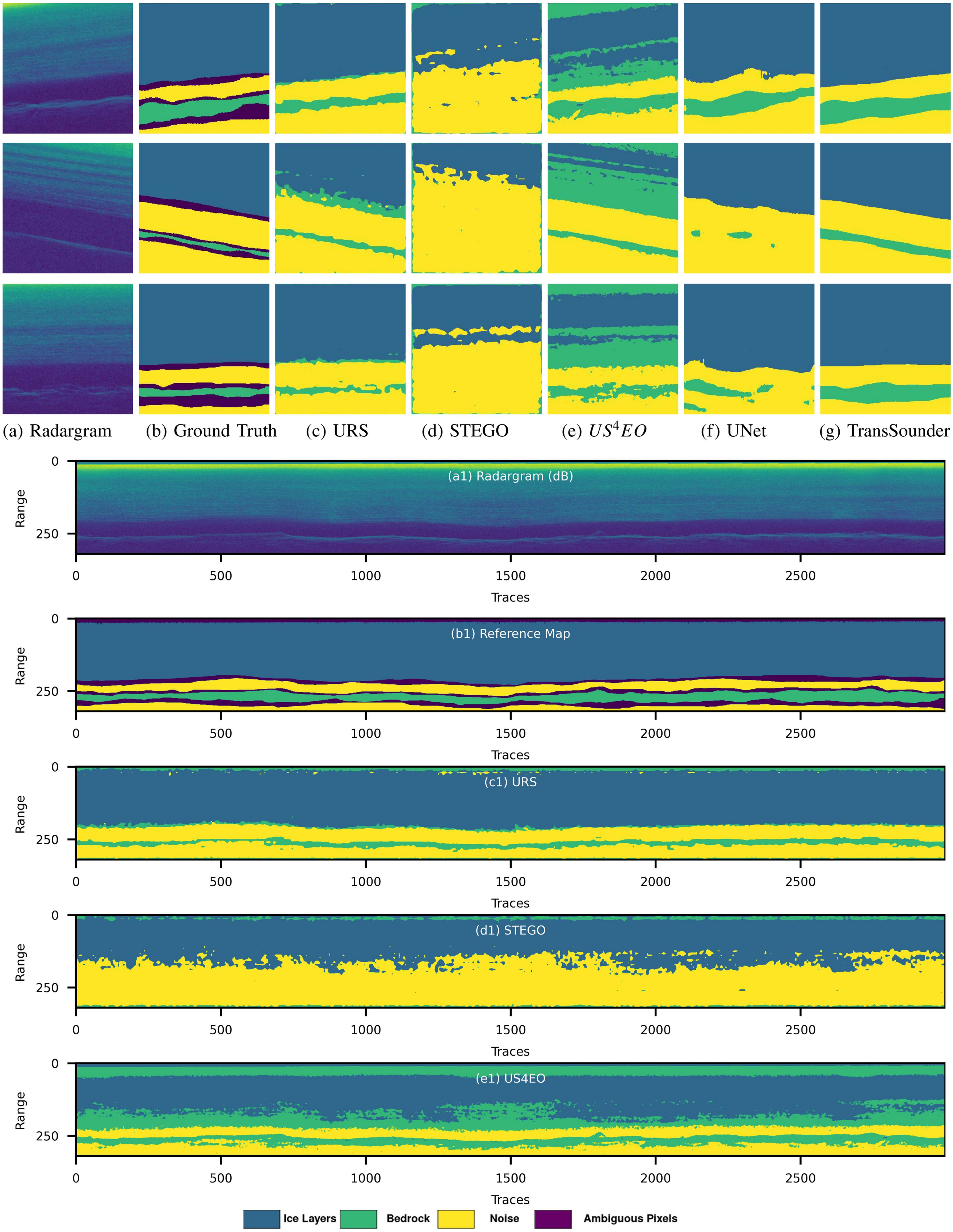
B. Experimental Setup

In terms of the experimental setup, 1200 randomly selected samples are employed for training [20]. No label is utilized for training the network or tuning hyperparameter in the unsupervised segmentation architectures. The test set is kept similar to [20]. The corresponding channel and spatial dimension for the training and testing sets are 1, and 320 × 320, respectively. In the case of associative mapping, Gaussian blur is incorporated as an augmented view generator ($\mathcal{A}$) for every training sample. As a pretrained visual featurizer ($\mathcal{N}$), DINO architecture is adopted as a frozen backbone which is a self-supervised ViT-based teacher–student network [43]. The $\mathcal{N}$ is incorporated over each x_j, x'_j to generate the discriminative pixel-embedding tensors with the dimension 384 × 40 × 40 (y_j, and y'_j). To amplify the correlation patterns embedded in the feature tensors, the segmentation head ($\mathcal{S}$) is enforced to reduce the channel dimension of the feature tensors (y_j, and y'_j) from $C' = 384$ to $C'' = 192$ (z_j, and z'_j). Hereafter, the expansive network $\mathcal{E}$ is utilized on the intermediate feature tensors y_j to reconstruct the tensors with spatial dimensions 320 × 320. $\mathcal{E}$ is a network with successive convolutional networks coupled with the intermediate transpose convolution operations to upsample (rate is 2, as shown in Fig. 2) the tensors from 40 × 40 to 320 × 320. The window size of the CNNs in $\mathcal{E}$ is kept as 3 × 3. Further, the LeakyReLU activation function is utilized. The values of $\lambda_s, \lambda_a,$ and λ_r [in (11)], are $0.67, 0.25,$ and 0.63, respectively [25]. For b_s, b_a, and b_r [in (11)], the values are $0.08, 0.02,$ and 0.46, respectively [25]. In order to reduce the computation of the feature correspondence tensors and segmentation correspondence tensors, a random grid sampling with spatial dimension $\mathcal{G}_1 \times \mathcal{G}_2 = 21 \times 21$ is utilized to compute F_{hwmn} and S_{hwmn} for estimating the contrastive correlation loss. AdamW optimizer is used with a learning rate of $1e - 5$ while keeping the epoch at 100 for the unsupervised networks. The batch size is kept as 16. We utilize the PyTorch deep learning framework to implement all the networks on the NVIDIA Tesla V100 GPU. Several assessment metrics, such as precision, recall, F-1 score, MIoU, and overall accuracy (OA) were used to assess the quantitative analysis between different supervised and unsupervised segmentation methods.

C. Segmentation Results

In this section, we report the quantitative (see Table II) and qualitative (see Fig. 3) evaluation of three unsupervised segmentation architectures: 1) the proposed URS, 2) STEGO [25], and 3) the unsupervised single-scene semantic segmentation for EO (US^{4}EO) architecture [35], and two supervised semantic segmentation architectures: 1) UNET [9], and 2) TransSounder [20]. We carried out the ablation study on our proposed URS architecture by utilizing the different combinations of loss functions ($\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_1 + \mathcal{L}_2$) along with the training strategies based on different positive descriptors during contrastive training (gb - gaussian blur and knn).

In terms of quantitative assessment for the unsupervised architectures, the proposed URS architecture achieved the highest MIoU in comparison with the other unsupervised segmentation architectures. URS improved OA of 17% and MIoU of 23.47% compared to the STEGO architecture. Further, URS outperformed US^{4}EO with an increased OA of 26.68% and MIoU of 12.08%. Therefore, URS outperformed both STEGO and US^{4}EO (see Table II). Further, the ablation study showed that the hybrid unsupervised learning optimization with contrastive correlation loss ($\mathcal{L}_1$) along with the spatial similarity loss ($\mathcal{L}_2$), achieved the highest rate of accuracy (see Table II). Since negative descriptors were not establishing strong discriminative embeddings in contrastive framework of STEGO, as expected the learning was not stable while incorporating the positive and negative samples of radargrams due to their spatial similarity.



(a) Radargram (b) Ground Truth (c) URS (d) STEGO (e) US^4EO (f) UNet (g) TransSounder

Fig. 3. Original Radargram (a, a1), Reference Map (GT) (b, b1), and associated prediction maps are highlighted in this figure with (c, c1) - URS, (d, d1) - STEGO, (e, e1) - US4EO, (f) UNet, and (g) TransSounder. Ambiguous pixels are not considered for accuracy. (a) Radargram. (b) Ground Truth. (c) URS. (d) STEGO. (e) US^{4}EO. (f) UNet. (g) TransSounder.

TABLE II
ACCURACY ASSESSMENT AND ABLATION STUDY OF MISCELLANEOUS UNSUPERVISED AND SUPERVISED ARCHITECTURES

Types	Algorithms	Precision	Recall	F1-Score	MIoU	OA
Unsupervised	**proposed URS-gb** ($\mathcal{L}_1 + \mathcal{L}_2$)	**0.8286**	0.7518	**0.7723**	**0.6884**	**0.8574**
Ablations	**URS-gb** ($\mathcal{L}_1$)	0.6795	0.6431	0.6324	0.5621	0.7872
	URS-gb ($\mathcal{L}_2$)	0.7261	0.7255	0.6076	0.5377	0.5332
	URS-knn ($\mathcal{L}_1 + \mathcal{L}_2$)	0.6678	0.6705	0.6588	0.5886	0.7774
Unsupervised	**STEGO**	0.6583	0.6598	0.6065	0.5268	0.7328
	US4EO	0.8021	**0.7577**	0.6932	0.6142	0.6768
Supervised	**UNET**	0.8376	0.8011	0.8140	0.7413	0.9275
	UNET-vit	0.8673	0.9372	0.8812	0.8055	0.8882
	TransSounder	0.9907	0.9859	0.9883	0.9771	0.9934

knn denotes the K-nearest neighbours and gb denotes the gaussian blur, respectively.
The bold values signifies the highest metrics achieved by the unsupervised segmentation methods.

TABLE III
CONFUSION MATRIX: URS ARCHITECTURE

	Unlabel	Layers	Bedrock	Noise	P-Acc
Unlabel	**1.000**	0.0	0.0	0.0	1.000
Layers	0.0	**0.9989**	0.0007	0.0003	0.9989
Bedrock	0.0000	0.7244	**0.2609**	0.0146	0.2609
Noise	0.0000	0.1909	0.0616	**0.7473**	0.7473
U-Acc	1	0.8000	0.5201	0.9942	0.8574

The bold values signifies the classwise classification accuracy achieved by the unsupervised methods.

TABLE V
CONFUSION MATRIX: US^{4}EO

	Unlabel	Layers	Bedrock	Noise	P-Acc
Unlabel	**1.000**	0.0	0.0	0.0	1.000
Layers	0.0	**0.9989**	0.0010	0.0000	0.9989
Bedrock	0.0	0.8251	**0.0971**	0.0776	0.0971
Noise	0.0	0.0288	0.0361	**0.9350**	0.9350
U-Acc	1.000	0.5246	0.7943	0.8894	0.6768

The bold values signifies the classwise classification accuracy achieved by the unsupervised methods.

TABLE IV
CONFUSION MATRIX: STEGO ARCHITECTURE

	Unlabel	Layers	Bedrock	Noise	P-Acc
Unlabel	**1.000**	0.0	0.0	0.0	1.000
Layers	0.0	**0.9985**	0.0014	0.0000	0.9985
Bedrock	0.0	0.7764	**0.1663**	0.0571	0.1663
Noise	0.0	0.4444	0.0810	**0.4745**	0.4745
U-Acc	1.0	0.6276	0.0061	0.9996	0.7328

The bold values signifies the classwise classification accuracy achieved by the unsupervised methods.

The overall contrastive framework was weakened due to the spatial homogeneity. In contrast to KNN, when considering Gaussian blur as positive descriptors (during training) for URS and STEGO, the MIoU improved by 14.49% and 6.28%, respectively. In addition, the parameters in US^{4}EO ($\approx 23M$), are much higher than in URS ($\approx 0.6M$). Even though supervised architectures achieved the highest OA (TransSounder with 0.9934), the OA of URS is significant when compared with end-to-end fully supervised settings with a significant amount of parameters ($\approx$ 125 M for TransSounder and $\approx$ 31 M for UNet). While assessing the confusion matrix, URS, STEGO, and US^{4}EO achieved the similar semantic segmentation accuracy for ice layers (see Tables III–V). In US^{4}EO, a significant amount of ice layers samples were misclassified as bedrock ($\approx$ 0.8251 from Table V). Due to this, the OA of US^{4}EO got affected significantly. However, US^{4}EO achieved the highest classification rate of 0.9350 (see Table V) for the noise class. On the other hand, URS detected the bedrock with highest OA of 0.2609 (see Table III). The low accuracy of the bedrock is mostly motivated by the distribution of the number of samples per class being strongly unbalanced (the number of samples in bedrock classes is much less than in the ice layers and noise classes). Thus, the unsupervised learning became skewed toward discriminating the ice layers and noise more accurately than the bedrock class. Further, the self-supervised ViT-based frozen backbone incorporated a spatial resolution reduction (ratio of $320/8$), that affected the learning mechanisms to extract rich semantic contexts for the bedrock classes. In terms of computational training time, the proposed URS, STEGO, and US^{4}EO took 1.67 h, 1.43 h, and 2.12 h, respectively. On the other hand, the supervised architectures, such as TransSounder and UNET took 7.33 h and 3.43 h, respectively during training.

In terms of qualitative assessment, several observations can be made while comparing the prediction maps (see Fig. 3). The proposed URS is more effective in discriminating the classwise contexts against the STEGO and US^{4}EO architecture [see Fig. 3(c) and (c1)]. URS preserved the overall sequentiality (top to bottom) of the distinct subsurface targets in the radargrams. The ablation study on the proposed URS showed that the probability amplitudes (embedded in the intermediate features) are more contextually richer (see Fig. 4) to accurately delineate the different classes while combining the contrastive correlation loss with the spatial similarity loss. In STEGO, the bedrock class was not detected in most of the along track locations and was heavily misclassified as noise [Fig. 3(d) and (d1)]. STEGO was unable to identify the boundaries between the ice layers and noise in the radargrams. For both US^{4}EO and STEGO, it was observed that the higher the rate of attenuation w.r.t the depth (near the

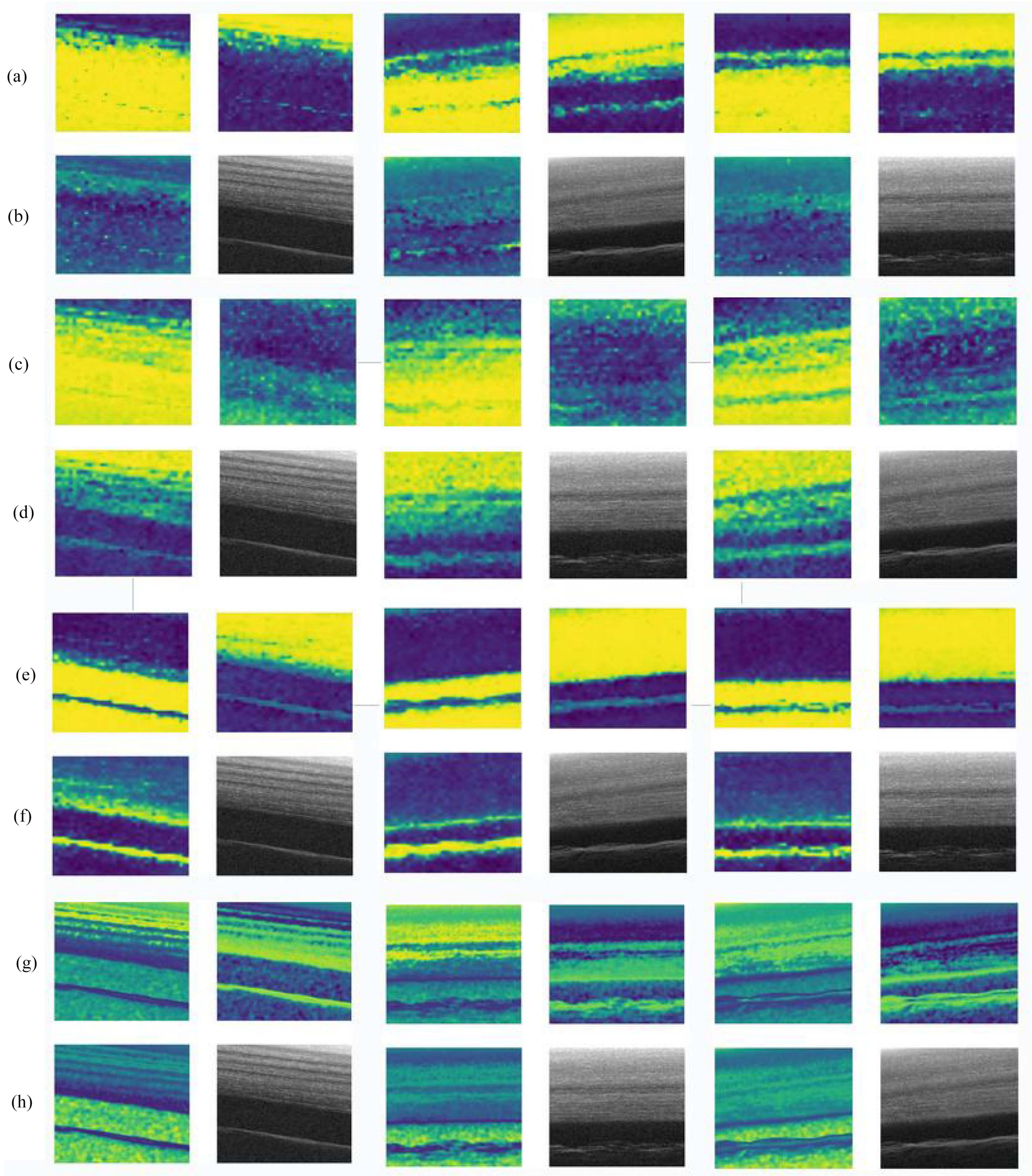


Fig. 4. Plots of intermediate feature maps derived from the pseudo-label generations for cluster compactifications during training. (a) and (b) represent the ablation study on proposed URS with Gaussian blur as positive descriptor and only with Contrastive Correlation Loss - denoted as URS-gb $(\mathcal{L}_1)$, (c) and (d) represent the ablation study on the proposed URS method with Gaussian blur as positive descriptor and only with the Spatial Similarity Loss - denoted as URS-gb $(\mathcal{L}_1)$, (e) and (f) represent the ablation study on the proposed URS method with Gaussian blur as positive descriptor and combining the Contrastive Correlation Loss and Spatial Similarity Loss - denoted as URS-gb $(\mathcal{L}_1 + \mathcal{L}_2)$. (e) and (f) represent the intermediate feature tensors of US^{4}EO methods.

fuzzy boundaries between ice layers and noise), higher the rate of misclassification between ice layers and noise [see Fig. 3(d), (d1), (e), and (e1)]. In contrast, the proposed URS method is less affected by the attenuation w.r.t the depth. On the other hand, US^{4}EO lost the sequentiality between the bedrock and ice layers in the final predictions as shown in Fig. 3(e) and (e1). The rate of misclassification near fuzzy boundaries (between ice layers and noise) is often higher in US^{4}EO, where a great chunk of ice layer pixels were misclassified as bedrock. This is in agreement with the empirical observations demonstrated by[44], where they showed that ViTs are robust against HF noise in contrast to CNNs. Further, the pixels belonging to the ice layers nearer to the surface (with high intensity regions below the free space) are misclassified as bedrock in US^{4}EO [Fig. 3(e) and (e1)]. Qualitatively, URS accurately modeled the local and global spatial details associated with the ice layers, bedrock, and noise, while preserving the sequentiality along the range directions.

In summary, the proposed URS method with self-supervised ViT-based contrastive framework demonstrated the potential of unsupervised semantic segmentation of radargrams. The stepwise reconstruction of the intermediate feature tensors with expansive network in URS boosted the overall performance of the architecture for capturing the discriminative semantic context more accurately and the coupling of contrastive correlation loss with the spatial similarity loss steered the gradient flow more effectively in URS. As expected, utilizing self-supervised ViTs as an encoder and expansive network as a proxy decoder demonstrated to be an efficient framework for the unsupervised feature learning in radargrams. Both URS and US^{4}EO captured the local spatial contexts more accurately than the supervised UNET architecture, thereby exhibiting the potential of the unsupervised feature learning framework with the contrastive learning-based approach.

IV. Conclusion

In this article, we constructed a self-supervised vision transformer-based feature learning framework for the unsupervised semantic segmentation of radar sounder data for the first time. We explored the ViTs-based unsupervised segmentation methods in the domain of RS signal segmentation. The proposed unsupervised deep feature learning framework is a lightweight computationally efficient method in contrast to the supervised ViTs for modeling the semantic contexts for the downstream semantic segmentation tasks. The experiments confirmed that the proposed unsupervised framework embeds discriminative pixel embeddings with rich semantic contexts without using any training from scratch over the unsupervised framework. This interpretation is aligned with the framework of NLP, where pretrained Transformers with large samples are fine-tuned with the smaller targets in the downstream tasks. Further, the proposed URS yielded rich semantic information in the intermediate dense features extracted from the RS signal. Experimental results on MCoRDS data confirm the capability of this lightweight unsupervised segmentation methods in the radar sounder data and show statistical significance while comparing with the heavily parameterized supervised segmentation approaches, such as TransSounder and UNET. URS shows the potential to be explored for other EO sensors, and be embedded as a lightweight unsupervised deep learning framework in the context of AI-powered planetary robots for in situ explorations, where a-priori geological knowledge is limited. In future work, we will explore the physics-driven deep networks to improve the accuracy for unsupervised semantic segmentation of radar sounder signal segmentation along with tracing the individual ice layers and measuring the ice thickness. Further, the possibility to design large radar sounders labeled dataset will be explored.

References

[1] S. Thakur and L. Bruzzone, “An approach to the generation and analysis of databases of simulated radar sounder data for performance prediction and target interpretation,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 10, pp. 8269–8287, Oct. 2021.

[2] L. Carrer, C. Gerekos, F. Bovolo, and L. Bruzzone, “Distributed radar sounder: A novel concept for subsurface investigations using sensors in formation flight,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 9791–9809, Dec. 2019.

[3] A.-M. Ilisei, A. Ferro, and L. Bruzzone, “A technique for the automatic estimation of ice thickness and bedrock properties from radar sounder data acquired at antarctica,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2012, pp. 4457–4460.

[4] A. Shepherd and H. J. Ivins, “A reconciled estimate of ice-sheet mass balance,” *Science*, vol. 338, no. 6111, pp. 1183–1189, 2012.

[5] M. A. Cooper, T. M. Jordan, M. J. Siegert, and J. L. Bamber, “Surface expression of basal and englacial features, properties, and processes of the Greenland ice sheet,” *Geophysical Res. Lett.*, vol. 46, no. 2, pp. 783–793, 2019.

[6] M. L. Goldberg, D. M. Schroeder, D. Castelletti, E. Mantelli, N. Ross, and M. J. Siegert, “Automated detection and characterization of antarctic basal units using radar sounding data: Demonstration in institute ice stream, West Antarctica,” *Ann. Glaciol.*, vol. 61, no. 81, pp. 242–248, 2020.

[7] R. Drews et al., “Layer disturbances and the radio-echo free zone in ice sheets,” *Cryosphere*, vol. 3, no. 2, pp. 195–203, 2009.

[8] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803, doi: 10.1109/CVPR.2018.00813.

[9] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, May 2015, pp. 234–241.

[10] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, “Feature pyramid networks for object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 936–944, doi: 10.1109/CVPR.2017.106.

[11] 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, doi: 10.1109/CVPR.2015.7298965.

[12] V. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet: A deep convolutional encoder-decoder architecture for image segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440, doi: 10.1109/CVPR.2015.7298965.

[13] Y. Wang, M. Xu, J. D. Paden, L. S. Koenig, G. C. Fox, and D. J. Crandall, “Deep tiered image segmentation for detecting internal ice layers in radar imagery,” in *Proc. IEEE Int. Conf. Multimedia Expo*, 2021, pp. 1–6.

[14] Y. Cai, J. Ma, H. Li, and S. Hu, “Automatic classification of ice sheet subsurface targets in radar sounder data based on the capsule network,” in *Proc. 8th Int. Conf. Comput. Pattern Recognit.*, 2019, pp. 199–204, doi: 10.1145/3373509.3373585.

[15] D. Varshney, M. Rahnemoonfar, M. Yari, and J. Paden, “Deep ice layer tracking and thickness estimation using fully convolutional networks,” in *Proc. IEEE Int. Conf. Big Data*, 2020, pp. 3943–3952.

[16] M. Liu-Schiaffini, G. Ng, C. Grima, and D. Young, “Ice thickness from deep learning and conditional random fields: Application to ice-penetrating radar data with radiometric validation,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5119014.

[17] E. Donini, F. Bovolo, and L. Bruzzone, “A deep learning architecture for semantic segmentation of radar sounder data,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4506514.

[18] M. Chendeb El Rai, J. H. Giraldo, M. Al-Saad, M. Darweech, and T. Bouwmans, “SemiSegSAR: A semi-supervised segmentation algorithm for ship SAR images,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 4510205.

[19] E. Lee, S. Jeong, J. Kim, and K. Sohn, “Semantic equalization learning for semi-supervised SAR building segmentation,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 4511505.

[20] R. Ghosh and F. Bovolo, “TransSounder: A hybrid transunet-transfuse architectural framework for semantic segmentation of radar sounder data,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4510013.

[21] J. Chen et al., “TransUNet: Transformers make strong encoders for medical image segmentation,” 2021, *arXiv:2102.04306*.

[22] Y. Zhang, H. Liu, and Q. Hu, “TransFuse: Fusing transformers and CNNs for medical image segmentation,” *Med. Image Comput. Comput. Assisted Intervention – MICCAI 2021*, M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, Ed., Cham: Springer International Publishing, 2021, pp. 14–24.

[23] R. Ghosh and F. Bovolo, “An FFT-based CNN-transformer encoder for semantic segmentation of radar sounder signal,” in *Image and Signal Processing for Remote Sensing XXVIII*, vol. 12267, Bellingham, WA, USA: SPIE, 2022, Art. no. 122670R, doi: 10.1117/12.2636693.

[24] O. Ibikunle, D. Varshney, J. Li, M. Rahnemoonfar, and J. Paden, "ECHOVIT: Vision transformers using fast-and-slow time embeddings," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2023, pp. 8162–8165.

[25] M. Hamilton, Z. Zhang, B. Hariharan, N. Snavely, and W. T. Freeman, "Unsupervised semantic segmentation by distilling feature correspondences," in *Proc. Int. Conf. Learn. Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=SaKO6z6Hl0c

[26] L. Melas-Kyriazi, C. Rupprecht, I. Laina, and A. Vedaldi, "Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8354–8365.

[27] W. V. Gansbeke, S. Vandenhende, S. Georgoulis, and L. V. Gool, "Unsupervised semantic segmentation by contrasting object mask proposals," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10032–10042, doi: 10.1109/ICCV48922.2021.00990.

[28] X. Ji, J. F. Henriques, and A. Vedaldi, "Invariant information distillation for unsupervised image segmentation and clustering," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9864–9873, doi: 10.1109/ICCV.2019.00996.

[29] M. Chen, T. Artières, and L. Denoyer, "Unsupervised object segmentation by redrawing," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, Art. no. 1140.

[30] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proc. Comput. Vis. – Eur. Conf. Comput. Vis.*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Ed., Cham: Springer International Publishing, 2018, pp. 139–156 2018, *arXiv:1807.05520*.

[31] X. Zhan, J. Xie, Z. Liu, Y. Ong, and C. C. Loy, "Online deep clustering for unsupervised representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6687–6696, doi: 10.1109/CVPR42600.2020.0067210.1109/CVPR42600.2020.00672.

[32] S. E. Chazan, S. Gannot, and J. Goldberger, "Deep clustering based on a mixture of autoencoders," in *Proc. IEEE 29th Int. Workshop Mach. Learn. Signal Process.*, 2019, pp. 1–6, doi: 10.1109/MLSP.2019.8918720.

[33] 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, 2021, doi: 10.1109/TPAMI.2019.2962683.

[34] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, Art. no. 149.

[35] S. Saha, M. Shahzad, L. Mou, Q. Song, and X. X. Zhu, "Unsupervised single-scene semantic segmentation for earth observation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5228011.

[36] E. Donini, M. Amico, L. Bruzzone, and F. Bovolo, "Unsupervised semantic segmentation of radar sounder data using contrastive learning," in *Image and Signal Processing for Remote Sensing XXVIII*, vol. 12267. Bellingham, WA, USA: SPIE, 2022, pp. 171–180.

[37] M. Pérez-García, J. Paoletti Haut, and J. López, "Novel spectral loss function for unsupervised hyperspectral image segmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 5506505.

[38] L. Mou, P. Ghamisi, and X. X. Zhu, "Unsupervised spectral–spatial feature learning via deep residual Conv–Deconv network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 391–406, Jan. 2018.

[39] M. Caron et al., "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9630–9640.

[40] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," *Comput. Vis. ECCV 2020: 16th Eur. Conf.*, Glasgow, UK, Aug. 2328, 2020, pp. 402–419, doi: 10.1007/978-3-030-58536-5_24.

[41] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929, doi: 10.1109/CVPR.2016.319.

[42] A.-M. Ilisei and L. Bruzzone, "A system for the automatic classification of ice sheet subsurface targets in radar sounder data," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3260–3277, Jun. 2015.

[43] M. Caron et al., "Emerging properties in self-supervised vision transforme RS," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9630–9640, doi: 10.1109/ICCV48922.2021.00951.

[44] N. Park and S. Kim, "How do vision transformers work?" in *Proc. Int. Conf. Learn. Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=D78Go4hVcxO

Raktim Ghosh (Student Member, IEEE) received the B.E. degree in mining engineering from the Indian Institute of Engineering Science and Technology, Shibpur, Howrah city, West Bengal, India, and the M.Sc. degree in geo-information science and EO with a specialization in Geoinformatics under the joint education programme between the Indian Institute of Remote Sensing, ISRO, Dehradun, Uttarakhand, India, and the Faculty of Geo-information Science and EO, University of Twente, Enschede, The Netherlands. He is currently working toward the Ph.D. degree as a joint member of the Remote Sensing Laboratory at the Department of Information and Communication Technologies, the University of Trento and the Remote Sensing for Digital Earth Unit, Fondazione Bruno Kessler, Trento, Italy.

His research interests include the automatic analysis of radar sounders data for investigating the subsurface features.

Francesca Bovolo (Senior Member, IEEE) received the laurea (B.S.) degree, the laurea specialistica (M.S.) degree (summa cum laude) in telecommunication engineering, and the Ph.D. degree in communication and information technologies from the University of Trento, Trento, Italy, in 2001, 2003, and 2006, respectively.

She was a Research Fellow with the University of Trento, until 2013. She is currently the Founder and the Head of Remote Sensing for Digital Earth Unit, Fondazione Bruno Kessler, Trento, Italy, and a member of the Remote Sensing Laboratory, Trento, Italy. She is one of the coinvestigators of the Radar for Icy Moon Exploration instrument of the European Space Agency Jupiter Icy Moons Explorer and a member of the science study team of the EnVision mission to Venus. Her research interests include remote-sensing image processing, multitemporal remote sensing image analysis, change detection in multispectral, hyperspectral, and synthetic aperture radar images, and very high-resolution images, time-series analysis, content-based time-series retrieval, domain adaptation, light detection and ranging, and radar sounders. She conducts research on these topics within the context of several national and international projects.

Dr. Bovolo was a recipient of the First Place in the Student Prize Paper Competition of the 2006 IEEE International Geoscience and Remote Sensing Symposium (Denver, 2006). She has been an Associate Editor for IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EOS AND REMOTE SENSING since 2011 and the Guest Editor for the Special Issue on Analysis of Multitemporal Remote Sensing Data of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. She is a referee for several international journals. She is a member of the program and scientific committee of several international conferences and workshops. She was the Technical Chair of the Sixth International Workshop on the Analysis of Multitemporal Remote-Sensing Images (MultiTemp 2011, and 2019). She has been a Co-Chair of the SPIE International Conference on Signal and Image Processing for Remote Sensing since 2014. She was the Publication Chair of the International Geoscience and Remote Sensing Symposium in 2015.