TransSounder: A Hybrid TransUNet-TransFuse Architectural Framework for Semantic Segmentation of Radar Sounder Data

SUBMISSION DATE / POSTED DATE

26-10-2021 / 01-11-2021

CITATION

Ghosh, Raktim; Bovolo, Francesca (2021): TransSounder: A Hybrid TransUNet-TransFuse Architectural Framework for Semantic Segmentation of Radar Sounder Data. TechRxiv. Preprint. https://doi.org/10.36227/techrxiv.16870633.v1

DOI

10.36227/techrxiv.16870633.v1

TransSounder: A Hybrid TransUNet-TransFuse Architectural Framework for Semantic Segmentation of Radar Sounder Data

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

***Abstract*—Radar Sounders (RSs) are nadir-looking sensors operating in high frequency (HF) or very high frequency (VHF) bands that profile subsurface targets to retrieve miscellaneous scientific information. Due to complex electromagnetic interaction between back-scattered returns, the interpretation of RS data is challenging. The investigations of ice-sheet subsurface structures require automatic techniques to account for both the sequential spatial distribution of subsurface targets and relevant statistical properties embedded in RS signals. Automatic techniques exists for characterizing these targets either related to probabilistic inference model or convolutional neural network (CNN) deep learning methods. Unfortunately, CNN-based methods capture local spatial context and merely model the long-range sequential context. In contrast to CNN, the Transformer-based models are reliable architectures for capturing long-range sequence-to-sequence global spatial contextual prior. Motivated by the aforementioned fact, we propose a novel Transformer-based semantic segmentation architecture named TransSounder to effectively encode the sequential structures of the RS signals. The TransSounder was constructed on a hybrid TransUNet-TransFuse architectural framework to systematically augment the modules from TransUNet and TransFuse architectures. Experimental results obtained on Multi-channel Coherent Radar Depth Sounder (MCoRDS) dataset confirm the robustness and capability of Transformers to accurately characterize the different subsurface targets.**

***Index Terms*—Semantic Segmentation, Transformers, TransUNet, TransFuse, MCoRDS, Radar Sounder, Sequence-To-Sequence Model**

I. Introduction

EARTH climate change is one of the most pivotal research topics in the domain of environmental monitoring [1]. An accelerated loss of the polar ice has been observed in recent decades [2]. The temperature rise significantly affects the dynamics of temporal changes within the deep ice layers. Therefore, an investigation of the ice-sheets subsurface structures is of high necessity. However, due to the inaccessibility of the sub-glacial environment, retrieving and characterizing environmental parameters of ice sheet subsurface structures and other targets becomes challenging [1]. In this context, nadir-looking airborne RS sensors and ground-based radar depth sounder (RDS) are non-intrusive instruments capable of providing important scientific information about the ice sheets.

RSs are active sensors designed to transmit linearly modulated electromagnetic (EM) pulse echoes and receive reflected echoes from the subsurface interfaces [3]. Generally, these interfaces are formed due to dielectric discontinuities between distinct targets with varying EM properties. The central frequency of these sensors ranges from high frequency (HF) to very high frequency (VHF) of the EM spectrum [1]. RS instruments are designed to retrieve amplitude and depth information of the subsurface ice sheets up to several kilometers beneath the surface. The transmitted EM signals encounter varying geometric and dielectric properties of subsurface targets and suffer from attenuation loss [4]. The amplitude returns are used to generate radargrams after noise reduction and platform instabilities correction [5].

Over the past two decades, the investigation of subsurface structures of ice sheets was carried out by visual interpretation of the radargrams generated by miscellaneous airborne and ground-based radar sounder (RS) instruments. However, manual investigations are time-consuming and not efficient for large-scale modelling. To address these limitations, automatic techniques are proposed. They are either associated with the probabilistic inference model or recently adopted deep learning-based segmentation framework. [6] constructed a statistical signal processing approach to automatically characterize radargrams with distinct targets and [7] demonstrated the capability of Support Vector Machine (SVM) to classify subsurface targets in the radargrams. [8], [9] utilized the probabilistic inference models for detecting and estimating the ice layers from radar sounder data. Although these methods are lightweight and computationally efficient, the requirement of prior probabilistic feature extraction becomes inefficient for large-scale modelling. Recently, the CNN-based approaches are incorporated to segment the radargrams for estimating the thickness of ice layers [10], [11], [12]. Although CNN-based approaches exhibit classic localized spatial representational power, they cannot coherently resolute the long-range sequential dependency for capturing the global spatial context. In contrast to local contexts, computing the response at a position in a sequence by considering all the other positional elements over the entire sequence with the weighted average is a contextualization of global features with sequential structures in the embedding space [13]. The sequential structures in RS signals coupled with more than one target (ice, bedrock, noise, etc.) are often too complex to model as extracting

This work was supported by the Italian Space Agency through the "Attivit'a Scientifiche per JUICE fase C-D" under Contract Agenzia Spaziale Italian-Istituto Nazionale di Astrofisica (ASI-INAF) and Contract 2018-25-HH.0.

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 Center of Digital Society, Fondazionae Bruno Kesseler, 38123 Trento, Italy (e-mail: bovolo@fbk.eu)

the global contextual feature from the signals is difficult. In order to resolve the issues of modelling global spatial contexts in radargrams, it is necessary to incorporate tools that can intrinsically capture the long-range spatial sequential context in RS signals.

In contrast to CNN-based methods, Transformers have transpired as alternative architectures solely relying on self-attention mechanism for modelling global long-range sequential context, thus dispensing entirely the convolution operation [14]. [15] theoretically established the universal approximation property of Transformers on sequence-to-sequence functions. The success of Transformers have been witnessed in the domain of Natural Language Processing (NLP) [14], [16] and demonstrated in the domain of image recognition [17], [18]. Very recently, [19], [20] developed TransUNet and TransFuse model respectively in the domain of medical image segmentation. The TransUNet architecture incorporates a joint CNN-Transformer module as an encoder network and merits the UNet like upsampling framework in the decoder part [19]. On the other hand, TransFuse utilizes solely a Transformer-based module as an encoder and uses a BiFusion module in the decoder part [20]. A brief description of the architecture is given in Section II.

In the case of radar sounder signals, the backscattering response from different subsurface targets in the radargram signals does not explicitly exhibit concrete distinguishable linear homogeneous features through the whole sequence in radargrams. However, RS signals embed a systematic sequential context with inherent properties such as exponential decay of signals from top to bottom and mixed back-scattering responses from the targets with varying dielectric properties. Therefore, to discriminate and capture the long-range dependency of these back-scattering responses from distinct targets, the utilization of Transformers as an encoder in the segmentation model could be a crucial step forward. In the field of semantic segmentation of radargrams, the Transformer-based architectures have not been explored so far. This is for the first time we address the problem of pixel-wise segmentation of radargrams by incorporating the Transformers as an encoder in deep learning architecture.

In this paper, we design a novel approach for Transformers-based semantic segmentation of radar sounder signals. Our contribution is as follows:-

1) We propose a hybrid TransUNet-TransFuse architectural framework named TransSounder to jointly augment their modules.
2) We incorporate a systematic augmentation of modules with attention mechanisms in both encoder and decoder architecture to accurately capture the global and local context in the down-sampling and also in the up-sampling operation.
3) We compare the TransSounder architecture with the state-of-the-art TransUNet [19] and TransFuse [20] architecture for pixel-wise segmentation of radargrams.

To briefly highlight the proposed TransSounder architectural framework, the successive convolution operations extract the high dimensional feature spaces from the radargram training samples majorly embedded with the local spatial context. These high dimensional feature spaces are fed into the Transformers to establish a global spatial contextual relationships between the previously extracted local spatial context. After these downstream tasks in the encoder architecture, the local and global contextual features are combined in the decoder architecture to recover the final low dimensional feature maps with the respective number of classes. The results indicate that the TranSounder is a robust architectural framework in the context of segmenting the radar sounder signals.

In terms of modelling the sequential structures in the RS signals, we first observed that it is reliable to tokenize the high dimensional feature spaces extracted after successive convolution operations as the Transformers explicitly model the global spatial context. The convolution operations preserve the local spatial contextual details in the RS signals at the initial stage. However, in terms of multi-hop dependency modelling, such as when the messages have to be passed through back-and-forth between distant positions, the convolution operations lack the exchanging of long-range information processing [13]. To model long-range sequential relations between the high dimensional local spatial contexts embedded at different positions in RS signals, the utilization of joint CNN-Transformer encoder is important. Here, Transformers intrinsically embed the long-range spatial contextual modelling between local spatial contexts extracted by CNN and preserve the hierarchical details at different depth of the network. Hence, the TransSounder architecture incorporates a joint CNN-Transformer encoder with a similar experimental setup that is in parallel with the encoder part of the TransUNet architecture [19]. The output from the Transformers branch is upsampled successively with the Transpose Convolution operations to match the dimensions with the corresponding CNN branches at different depths. We utilize the BiFusion module [20] in the decoder architecture to perform the feature-level fusion between the high dimensional downsampled features from the CNN branches and the upsampled features from the Transformers block. A complete mathematical treatment of TransSounder architecture has been depicted in Section III.

The rest of the paper is organized as follows. We elucidate the related works in Section II. A detailed methodological framework is depicted in Section III. The experimental results are reported in Section IV. In this section, we highlight the description of the dataset, construct the experimental setup and discuss the segmentation results. Finally, we draw conclusions from our research work.

II. Related works

In this section, we highlight the hierarchical development of the automatic methods of classification and segmentation of RS data. We first depict the probabilistic methods for characterizing the radagrams in terms of targets and additionally we highlight a few pieces of literature on detecting and estimating the ice layers. After that, we elucidate the CNN-based approaches to segment the radargrams, and we briefly highlight the methods in which the ice layers are detected using CNN-based semantic segmentation. Next, we elaborate on the concepts of TransUNet and TransFuse architectures.

A. Probabilistic Method

The studies such as [6], [7] presented automatic techniques for characterizing subsurface targets in radargrams. [6] demonstrated the capability of statistical signal processing approaches to automatically characterize and generate feature maps from radargrams. Based on the amplitude fluctuations of received signals, miscellaneous probability density functions (pdfs) were empirically fitted with the histograms to model radargrams. The experimental results demonstrated that K-distribution and Rayleigh distribution successfully characterize the subsurface targets and noise respectively. However, the study was limited to extracting homogeneous and linear features from radargrams whereas the subsurface features in a radargram can be highly heterogeneous. [7] demonstrated the significance of machine learning-based approaches in addressing the spatial heterogeneity of the subsurface targets by processing the probabilistic features with the Support Vector Machine (SVM) algorithm. Although the SVM is computationally efficient, a number of prior probabilistic feature extractions for distinct classes have to be considered which may be prone to errors in terms of generalization capabilities.

In terms of detecting and estimating the thickness of ice layers, significant research activities related to automatic techniques have been carried out. [21] proposed an active contour-based and edge-based (edge detection and thresholding) solution for locating ice sheets, and the interface between ice layers and bedrock. [22] developed a level-set method to detect ice layers. [8] proposed a probabilistic inference task to estimate bedrock and surface layer boundaries. This was done by utilizing the Markov-Chain Monte Carlo simulation method to sample from the joint distribution over all possible layers. [9] suggested a probabilistic graphical model based on inference task for 3D ice layer extractions within the framework of computer vision. They incorporated the concept of generating seed surfaces, and thereby constraint-based refining of those generated surfaces via discrete energy minimization technique. Although the probabilistic models are computationally light-weight, the prior handcrafted feature extraction or pdfs consideration for distinct target classes are often difficult in terms of large-scale modelling or generalization capabilities.

B. CNN-based Approaches

A few research work focused on developing methods for estimating the thickness of the ice sheets by deep learning-based segmentation with CNN-based approach. Very recently, [10] developed a joint Triple task CNN architecture, and multi GAP Recurrent Neural Network (RNN) architecture to address the problem of deep tiered segmentation of internal ice layers. By incorporating the 2D-CNN task, they addressed three problems simultaneously: i) detecting the location of top layers, ii) roughly approximating the thickness of ice layers, iii) quantifying the number of visible layers in the echogram. Later, they incorporated the RNN operation for pixel-level refinement of boundary positions to account for the differences across the layers. Also, [11] incorporated a CNN-based Capsule Network by utilizing the SegCaps network architecture [23], to segment radargrams in terms of layers, bedrock, noise, and free space. In contrast to CNN architecture, the capsule network stores the information about spatial orientation at the neuron level which is useful for the segmentation of radargrams. [12] carried out a comparative analysis between different FCN-based architecture for tracking deep ice layers and subsequently estimating the thickness. Although CNN-based approaches accurately capture the local spatial context, they lack modelling the long-range sequence-to-sequence spatial contextual features. The RS signals intrinsically depict the sequential structures of backscattered responses in the radargrams. Therefore, it is of paramount importance to incorporate tools that can resolute the long-range sequential context.

C. Transformer-based Segmentation Method

In contrast to CNN-based architectures, the Transformers are reliable architectures for modelling long-range global contextual features by incorporating the Multi-head Self Attention mechanism (MSA) coupled with Multi-layered Perceptron blocks (MLP) [14]. Recently, TransUNet and TransFuse architectures have been proposed by incorporating Transformers as an encoder for medical image segmentation [19], [20].

*1) **TransUNet**:* Although Transformers are reliable models for capturing long-range dependencies in the sequence-to-sequence prediction paradigm, they lack the ability to extract the local features whereas convolution operations reliably do. Therefore, to effectively utilize the combined capability of Transformers and convolutions, [19] developed a TransUNet architecture by incorporating the Transformers in the Encoder side. The TransUNet model is an encoder-decoder architecture based on the framework of UNet [24]. In the encoder part of TransUNet, the convolution operations extract the deep features. They are then tokenized by linear embedding operations coupled with positional encoding tensors. These tokens are fed into the number of sequential Transformer blocks. Therefore, the encoder architecture captures the local spatial details during convolutions, and then the global context is captured by Transformers. After capturing the global context from the tokens, the resulting Tensors are reshaped to match the size of tensors from the CNN block. In the decoder architecture, the reshaped output from Transformers is concatenated with the CNN modules at distinct spatial dimensions to augment the local as well as global spatial contexts. To match the spatial dimensions of the input tensors, the Cascaded Upsampling operations (CUP) are carried out until the final dimensions are retrieved. A detailed architectural description can be found in [19].

*2) **TransFuse**:* [20] developed a TransFuse architecture by utilizing shallow Convolutional Neural Network (CNN) and Transformers based encoder in parallel. The CNN branch encodes the local spatial details by successively increasing the receptive field, and gradually capture the global context. On the other hand, the Transformer branch captures the global context by incorporating the global self-attention mechanism. Later, the BiFusion module fuses the features extracted from CNN and Transformer branches concurrently with a distinct

depth of tensors. The intuition behind coupling them is that the CNN captures the local spatial details, and Transformer models the long-range global semantic content from the input. Therefore, the BiFusion module plays a pivotal role in capturing the global and local contextual information from the input data. The parallel structure is computationally efficient in model sizes and inference speed. In the decoder part, the progressive upsampling (PUP) method is adopted from the SETR method [25], to recover the spatial dimension concurrently. For further details about the methodology and corresponding architecture of TransFuse method, please refer to [20].

Transformers capture the global spatial contextual features, however, it often lacks the ability to model the local spatial contextual details. Therefore a joint CNN-Transformer encoder is a reliable solution. Although, TransUNet architecture incorporates a CNN-Transformer encoder [19], the simple concatenation operation of high level and low level features in the decoder network may not preserve the hierarchical structure of the RS signals. In terms of RS signals, it can be important to incorporate attention mechanisms in the decoder architecture to retain the precise spatial information by preserving the sequential structures throughout the range (spatial attention) and exclude the redundant channels in the high-dimensional feature space (channel attention). On the other hand, TransFuse incorporates a channel and spatial attention based BiFusion module which is reliable to preserve local and global spatial information effectively [20]. However, TransFuse do not incorporate a CNN-Transformer encoder thus dispensing to contextualize the local spatial details hierarchically at the initial stage of the encoder network. To combine the encoder block of TransUNet architecture and the decoder block of TransFuse architecture effectively, the aforementioned limitations can be resolved in terms of segmenting the radargrams with higher accuracy.

III. Proposed Trans-Sounder Network

A. Problem Formulation

Let us denote a radargram as a 2-D matrix with traces in the along track or azimuth direction denoted as $[1, ..., n_T]$, and the samples in the depth or range direction denoted as $[1, ..., n_S]$. The backscattered information contained in a radargram R is:

$$R = \{R(i, j) | i \in X = [1, ..., n_T], j \in Y = [1, ..., n_S]\} \quad (1)$$

The primary goal of this research work is to classify each $R(i, j)$ pixel in the radargram into a distinct class by incorporating a supervised semantic segmentation architecture. We broadly categorize the target structures of radargrams into 3 classes: layers, bedrock, and noise. We propose a hybrid architectural framework named TransSounder for pixel-wise segmentation of radargrams by incorporating Transformer-based segmentation models. Here, we jointly augment the modules from TransUNet and TransFuse architecture to accomplish the aforementioned objective.

Let N be the number of training samples build of pairs of radargram patches and the corresponding labels. Thus the training set is denoted as $\{(X_1, L_1), (X_2, L_2), ..., (X_N, L_N)\}$ where $X = \{X_1, X_2, ...X_N\}$ are the radargram patches and $L = \{L_1, L_2, ..., L_N\}$ the corresponding labels. Let a training patch have a dimension $X_i \in \mathbb{R}^{H \times W}$ where the spatial structure is of $H \times W$. Therefore the spatial dimension of a label L_i is $H \times W$. A detailed mathematical treatment of TransSounder architecture is elucidated below.

B. TransSounder

As shown in Figure 1, the TransSounder architecture consists of a hybrid CNN-Transformer block as an encoder. The CNN branches capture the local spatial details successively from the training samples with the spatial dimensions of the tensors convolved from $[H, W]$ to $[\frac{H}{2^\lambda}, \frac{W}{2^\lambda}]$. λ depends on H and W (please refer Section IV-B for discussion on how to choose it). The high dimensional tensors at the depth of $[\frac{H}{2^\lambda}, \frac{W}{2^\lambda}]$ embed the local spatial contextual features at the last layer of CNN blocks. The tensor at the last layer is split into 1×1 tensors sequentially. These sequential tensors are tokenized to model global contextual relationships between them later in the Transformers branch. The tokenization technique encodes each subset of 1×1 tensors with the patch embedding as well as positional encoding operations. The embedded tokens are fed into the Transformers block with the depth of 8 as depicted in Figure 1. After extracting tensors associated with the global contextual features from the Transformers block, the reshaping is done to match the spatial dimensions with the tensors at the last layer of CNN blocks (spatial dimension of $[\frac{H}{2^\lambda}, \frac{W}{2^\lambda}]$). Successive Transpose Convolution operations are performed over these reshaped tensors to match the spatial dimension with the tensors at different depths of CNN blocks. Reshaped tensors from the Transformers and the tensors from CNN blocks are jointly fed as an input to the BiFusion modules $\{BF_1, BF_2, ..., BF_\lambda\}$. The output from BF_1 (spatial dimension of $[\frac{H}{2^\lambda}, \frac{W}{2^\lambda}]$) is then upsampled to match the spatial dimension with the output from BF_2, and subsequently, the concatenation operation is performed between these two outputs. Here, we utilize Bilinear Upsampling operation [26], however, other operations such as bicubic [27], trilinear upsampling [28] can be used without loss of generality. Later, a convolution operation is carried out on the concatenated tensor and the subsequent process is repeated until the final dimension of the output tensor $[H, W]$ is recovered at the decoder architecture. Detailed step-by-step mathematical descriptions are depicted below.

*1) **Hybrid CNN-Transformer Block as an Encoder**:* The concept has been adopted from [19]. We perform the successive Double-Convolution operation (convolving twice with window size of $w \times w$) on $\{X_1, X_2, ..., X_N\}$ to reduce the dimension of tensors upto $[\frac{H}{2^\lambda} \times \frac{W}{2^\lambda}]$ by following the sequence as $[\frac{H}{2^j} \times \frac{W}{2^j}]$ where $j \in \{0, 1, 2, ..., \lambda\}$. The CNN operation captures successively the local spatial contexts with the high the dimensional feature spaces.

Let us denote the set of tensors with high-dimensional convolved feature spaces extracted from N radargram training samples after each Double-Convolution operation as $\{Y_1^j, Y_2^j, ..., Y_N^j\}$ with the dimension of $[\frac{H}{2^j} \times \frac{W}{2^j}]$ where $j \in$

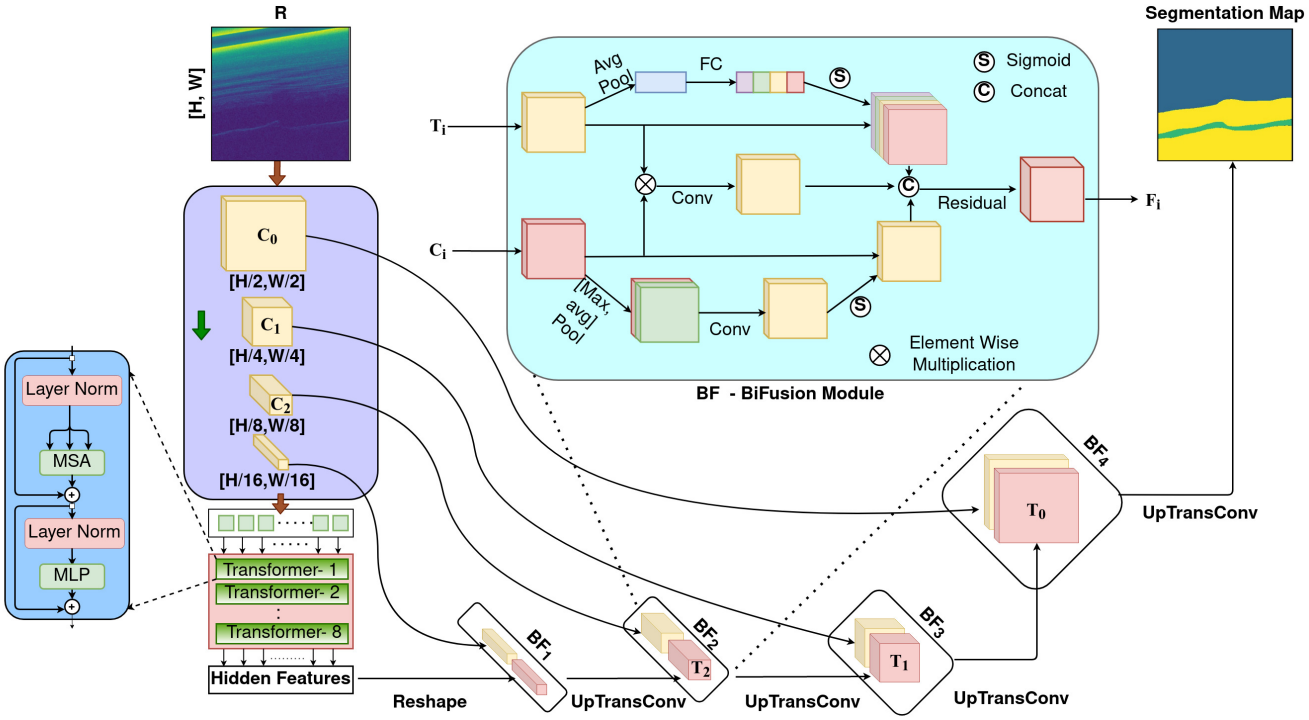


Figure 1: Schematic Layout of TransSounder Architecture

$\{0, 1, 2, ..., \lambda\}$. In the sequentialization operation, these high-dimensional deep features at $j = \lambda$ are tokenized to feed into the Transformers. The detailed dimensionality of this operation is depicted below.

*2) **Sequentialization of Convolved Patches**:* The tokenization operation is performed by selecting each individual Y_i^λ from the set of deep features after the last Double-Convolution operation at $j = \lambda$ as $\{Y_1^\lambda, Y_2^\lambda, ..., Y_N^\lambda\}$. Let us denote this sequence $\{Y_1^\lambda, Y_2^\lambda, ..., Y_N^\lambda\}$ as $\{Cn_1, Cn_2, ..., Cn_N\}$ for consistency in notations at later stage. Each Cn_i is divided into the sequence of flattened 2D patches where $\{Cn_i^j \in \mathbb{R}^{P^2 \times c}$ where $j \in \{1, 2, ..., m\}\}$. Each patch is of size $P \times P$ and $m = \frac{H/2^\lambda \times W/2^\lambda}{P^2}$. This tokenization operation is performed to feed into the Transformers for establishing the global contextual relationships between the high-dimensional subset of the tensors extracted from Cn_i.

*3) **Patch Embedding and Positional Encoding Operation**:* After the sequentialization of each Cn_i, the tensorized versions of these patches are projected onto a d-dimensional embedding space with trainable parameters. This operation is performed to reduce the dimensionality of the tensors. Also, the positional encoding has been added with the patch embedding operator to retain the precise localization information at different depth of the network. The equation can be depicted as follows:

$$z_0 = [Cn_i^1.E; Cn_i^2.E; ...; Cn_i^m.E] + E_{pos} \quad (2)$$

where $E \in \mathbb{R}^{P^2.c \times d}$ is the projection onto the patch embedding, and $E_{pos} \in \mathbb{R}^{m \times d}$ denotes the positional encoding.

*4) **Transformer Block**:* After carrying out the patch embedding and positional encoding operations over the tokenized patches denoted as Cn_i, these embedded tokens are fed into the Transformers. An operational sequences of Transformers have been depicted in Eq. (3a-3c).

The Transformer encoders consist of L-layers of subsequent Multihead Self-Attention (MSA) and Multi-Layer Perceptron (MLP) blocks [14]. Inside MLP blocks, the Gaussian Error Linear Unit (GELU) non-linearity is utilized [29]. The equations in 3 depict the successive operational sequence in the Transformer block.

$$z_l' = f(g(z_{l-1})) + z_{l-1} \quad (3a)$$

$$z_l = r(g(z_l')) + z_l' \quad (3b)$$

$$SA(z_i) = softmax(\frac{q_i k^T}{\sqrt{d_h}})v \quad (3c)$$

where f denotes the MSA operator, g indicates the layer normalization (LN) operator, r indicates the MLP operator, z_l indicates the l^{th} layer. Eq. (3c) depicts the self attention mechanism with the concept of queries (q), keys (k) and values (v). The [q,k,v] $= z \times W_{qkv}$, where $W_{qkv} \in \mathbb{R}^{d \times 3d_h}$ is the projection matrix and z_i, q_i are the i^{th} row of z and q. For more details about Transformers, please refer to [14].

*5) **Transpose Convolution on the Output from Transformer**:* After feeding the tokenized patches into the Transformers block with successive MSA and MLP operations (see Eq. (3a-3c), the output tensors from these Transformers are reshaped to match the similar spatial dimension at the last layer of the CNN branch. Let us define the sequence of output tensors from the last Transformers branch as $\{T_1, T_2, ..., T_N\}$. The spatial dimension of these tensors at the last layer of Transformer branch are $[\frac{H}{2^\lambda} \times \frac{W}{2^\lambda}]$.

Here, T_i $(i \in 1, 2, ..., N)$ embeds the long-range sequential contexts between the tokens representing the high dimensional local spatial contexts extracted from the CNN layers. Let us also define the Transpose Convolution operations over the output from the Transformer branch as $\{T_1^j, T_2^j, ..., T_N^j\}$ where $j \in \{0, 1, 2, ..., \lambda\}$. The sequential increment of spatial dimensions for successive Transpose Convolution operations are $[\frac{H}{2^{\lambda-j}} \times \frac{W}{2^{\lambda-j}}]$ and also from the CNN block, the corresponding hidden features were previously denoted as $\{Y_1^j, Y_2^j, ..., Y_N^j\}$ with the dimension of $[\frac{H}{2^j} \times \frac{W}{2^j}]$ where $j \in \{0, 1, 2, ..., \lambda\}$.

These concurrent hidden sequences of tensors from the CNN branches ($\{Y_1^j, Y_2^j, ..., Y_N^j\}$) and also the upsampled tensors from the Transformers branches $\{T_1^j, T_2^j, ..., T_N^j\}$ with the similar spatial dimensions at different stages are jointly fed into the BiFusion module for fused feature representations with the global and local spatial contexts in the RS signals. Here, the tensors from j^{th} CNN branch is spatially similar with the $(\lambda - j)^{th}$ upsampled tensors from the transformers branch. A detailed mathematical treatment inside the BiFusion module has been elucidated below.

*6) **BiFusion Module**:* The BiFusion module fuses the encoded deep local spatial contextual features from the CNNs with the modelled global spatial contextual features from the Transformers blocks. The module consists of 3 building blocks: a Spatial Attention (SA) operation, a Channel Attention (CA) operation, and a multi-modal fusion operation. The fused feature representations are depicted below in Eq. (4a-4d).

$$T_u^{'} = ChannelAttn(T_1^j, T_2^j, ..., T_N^j) \tag{4a}$$

$$C_u^{'} = SpatialAttn(Y_1^{\lambda-j}, Y_2^{\lambda-j}, ..., Y_N^{\lambda-j}) \tag{4b}$$

$$B_u^{'} = Conv(T_u^{'} W_u^1 \odot C_u^{'} W_u^2) \tag{4c}$$

$$F_u^{'} = Residual[T_u^{'}, C_u^{'}, B_u^{'}] \tag{4d}$$

where $\{T_1^j, T_2^j, ..., T_N^j\}$ denotes the reshaped global contextual features from Transformer block at the j^{th} branch, $\{Y_1^j, Y_2^j, ..., Y_N^j\}$ indicates the high dimensional tensors from j^{th} CNN block, $|\odot|$ represents the element-wise dot product, Conv denotes the 3×3 convolution layer.

The CA module is incorporated as Squeeze-and-Excitation (SE) block primarily proposed in [30]. The SE block captures the global information from the Transformer block. On the other hand, the SA module is borrowed from Convolutional Block Attention Module (CBAM) architecture [31]. The goal of the CBAM block is to capture intrinsically significant local spatial context from CNN blocks and ignore other regions. The cross-relationships between these blocks are established by performing the element-wise dot product operation between $T_u^{'}$ and $C_u^{'}$ tensors. After these aforementioned operations, the $T_u^{'}$, $C_u^{'}$, $B_u^{'}$ are passed through the residual blocks. Finally, the fused features $F_u^{'}$ jointly capture global and local features. For further details about BiFusion module, please refer to [20].

*7) **Concatenation Operation on the Output from BiFusion Module**:* After extracting the outputs from the BiFusion module by utilzing Eq.(3a-3c), the Bilinear Upsampling (BUP) operation is then incorporated on F_u tensor to match the similar dimension of F_{u+1} tensor. Let us denote this operation as $BUP(F_u)$. This $BUP(F_u)$ tensor is then concatenated with F_{u+1} tensor. Here, this concatenation operation accumulates the fused global and local spatial contextual representations at different scales. Hereafter, a convolution operation is performed on this concatenated tensor. This process is repeated until the final dimension $[H, W]$ is recovered. The mathematical formulation can be done as follows:

$$F_{u+1}^{'} = Conv(Concat(BUP(F_u), F_{u+1})) \tag{5}$$

The methodological framework, and corresponding operation in the decoder architecture is in parallel with the [20].

*8) **Loss Function**:* We use the binary cross-entropy (BCE) loss function to train the model. The cross entropy is a measure of the statistical distance between two probability distribution functions for a set of events or a distinct random variable [32]. The BCE loss is widely used for classification as well as pixel-wise segmentation task. Without loss of generality, different types of loss function other than BCE can be utilized for similar experimental setup.

We incorporate a triple headed BCE loss function as similar to [20]. The first head is derived from the BiFusion block 1 (BF_1), where the tensor at the depth $[H/2^4, W/2^4]$ ($\lambda = 4$) is upsampled to $[H, W]$ using bi-linear interpolation method (denoted as $\mathcal{L}_{BF_1}$ in Eq. 7). On the other hand, the hidden features from the Transformers block is upsampled with the similar size of different depth of CNN layers. Here, from the Transpose Convolution operation at the size of $[H/2, W/2]$, is upsampled to the size $[H, W]$ using bi-linear interpolation (denoted $\mathcal{L}_T$ in Eq. 7). The last head is coming from the successive operations of BiFusion modules at BF_4 (loss is denoted as $\mathcal{L}_{BF_4}$ in Eq. 7). The loss is then estimated with respect to the corresponding labels ($L = \{L_1, L_2, ..., L_N\}$) of the individual radargram training samples. BCE loss with a single head (BF_1) can be depicted as:

$$\mathcal{L}_{BF_1} = \mathcal{L}(L, head(BF_1)) = \sum_{i=1}^{H \times W} \sum_{c=1}^{C} -x_{ic} \log p_{ic} \tag{6}$$

where L is the set of labels of individual radargrams with size $H \times W$, C is the number of classes ($C = 3$ in our cases), p_{ic} indicates the predicted probability of i^{th} pixel on radargrams on class c of $head(BF_1)$, x_{ic} denotes the ground truth probability on i^{th} pixel of L.

In our TransSounder architecture, a linear combination of the triple-headed BCE loss is incorporated by utilizing all the aforementioned 3 heads:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{BF_4} + \beta \mathcal{L}_{BF_1} + \gamma \mathcal{L}_T \tag{7}$$

where α, β, and γ in Equation 7, are the Hyperparameters to be tuned according to the gradient flow of the networks.

IV. Experimental Results

In this section, we report the results by applying TransUNet, TransFuse, UNet, and TransSounder architectures for segmenting the radargrams. In the succeeding subsections, we first present a description of the dataset used for training and testing. Next, we highlight the experimental setup and associated resources. At last, we elucidate the segmentation results and we also depict the qualitative and quantitative interpretations of these results.

A. Description of Data Sets

We test our deep learning architectures on Multi-Channel Coherent Radar Depth Sounder (MCoRDS) which is owned by Centre of Remote Sensing of Ice Sheets (CReSIS) unit. The dataset was acquired by different sensors operated with different bandwidths, i.e., 9.5 MHz and 30 MHz. The central frequency of the instrument is 193.5 MHz. The instrument on-boarded on DC-8 jet aircraft, was flown with an altitude of about 7000 m. The acquisition took place over several regions of Antarctica. The first dataset campaign using MCoRDS, was conducted on November 2010 over the central Antarctica. 8 radargrams acquired on November 2010 were generated sequentially from the acquisitions on bandwidth 9.5 MHz. The latitude of these acquisitions varies from (−86°00′N to −15°67′E) to (−86°02′N to 29°45′E) over a distance of about 400 Km (total 27350 traces). The range resolution in ice and along track resolution corresponds to 13.6 m and 25 m respectively.

B. Experimental Setup

For the radargrams, a set of labelled patches is available. $N = 1600$ samples were used for training and 267 samples were used for testing. During training, two specific data augmentation techniques were used: i) grid distortion and ii) elastic deformation so that the sequential structures of the signals are invariant. The spatial dimension of each radargram training sample is $H \times W = 400 \times 400$. For the double convolution operation on CNN branches, the window size is kept as $w \times w = 3 \times 3$. In general, the window size can be experimented with different spatial structures. We set $\lambda = 4$ to donwsample the tensors from 400×400 to 25×25 over the successive CNN branches on the encoder network. The λ can vary from 1 to 7 in our case. If we choose a very small λ value, we may not be able to extract the deep high dimensional feature spaces, and on the other hand if we set it very high, the computational cost and the number of hyperparameters will increase. Therefore, a value of $\lambda = 4$ represt an trade-off between the extremes. All the networks are implemented in PyTorch, and experiments are conducted on two NVIDIA RTX 2080Ti GPUs. We performed parallel computing to effectively deploy the underlying networks with batch sizes of 16. We increased the training iterations from 30 to 100 for every architectures. The total number of training iterations for Transformer-based architectures (TransSounder, TransUNet, TransFuse) was kept as 50 due to the early convergence rate. While carrying out the experiments with 50 iterations for UNet architecture, we observed that networks didn't converge to a local minima. Therefore, we fixed the number of iterations for the UNet architecture as 100. ADAMW optimizer with learning rate 1e-5 are chosen. For the TransSounder architecture, the hyperparameters on the loss function in Eq. (7) is chosen as: $\alpha = 0.5$, $\beta = 0.3$, and $\gamma = 0.2$. Further, we utilize LeakyReLU activation function. Several evaluation metrics are used to evaluate the performance of the different methods such as Overall Accuracy (OA), F1 Score, Precision, Recall, and Kappa Score. OA is measured by dividing the total number of correctly classified pixels by the total number of pixels in the whole radargram.

Table I: Acurracy Assesment

Algorithms	Precision	Recall	F1-Score	Kappa	OA
TransFuse	0.9902	**0.9862**	0.9882	0.9907	0.9943
TransUNet	0.9839	0.9801	0.9819	0.9891	0.9934
UNET	0.8391	0.8271	0.8324	0.9004	0.9393
TransSounder	**0.9913**	**0.9862**	**0.9887**	**0.9910**	**0.9945**

Table II: Confusion Matrix: TransFuse Method

	Unlabel	Layers	Bedrock	Noise	P-Acc
Unlabel	**1.000**	0.0	0.0	0.0	1.000
Layers	0.0	**0.9957**	0.0001	0.0041	0.9957
Bedrock	0.0008	0.0	**0.9560**	0.0430	0.9560
Noise	0.0001	0.0041	0.0022	**0.9934**	0.9934
U-Acc	0.9991	0.9976	0.9760	0.9883	0.9943

C. Segmentation Results

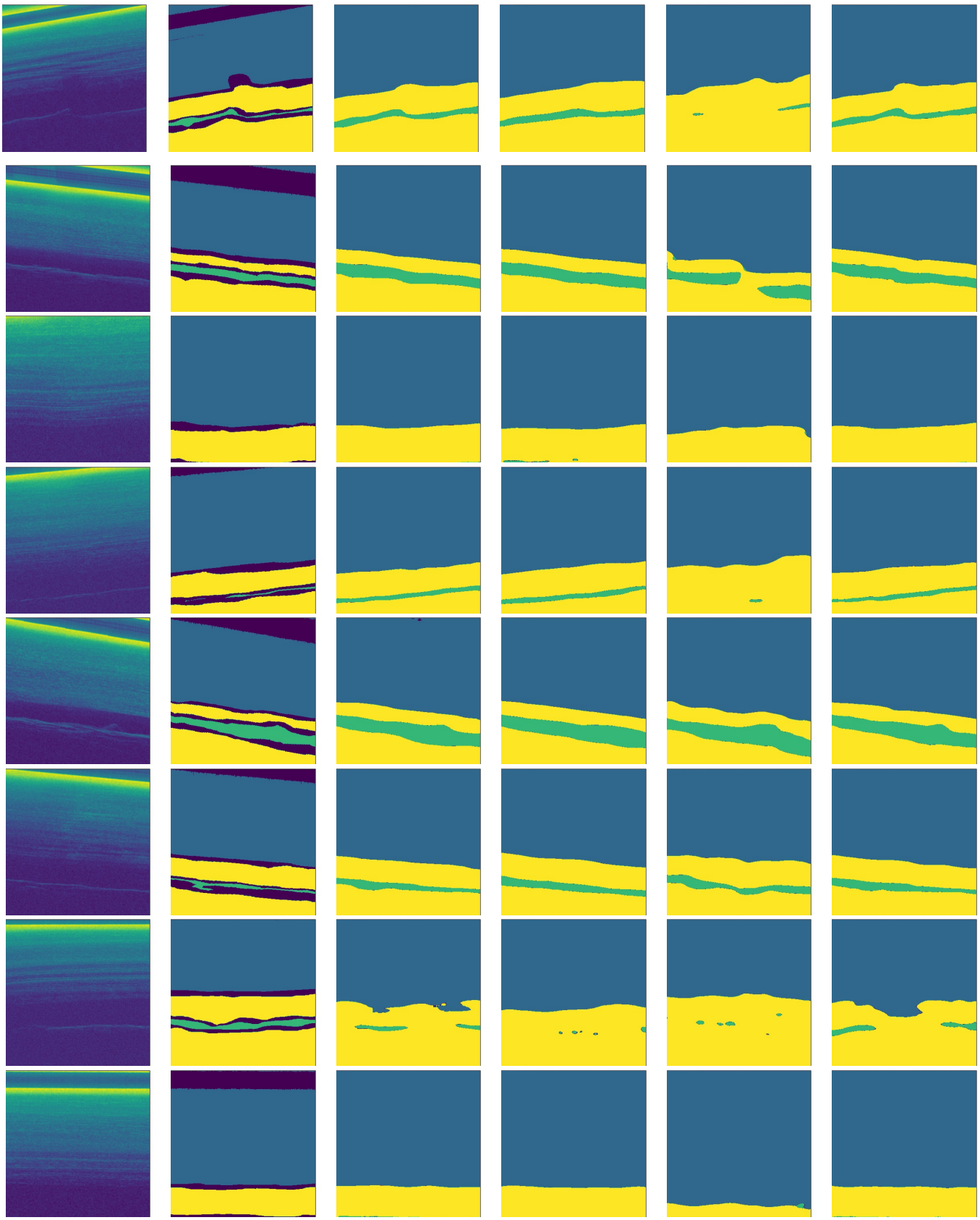
We tabulate these accuracy assessments for all the four experiments in Table I. In this table, the highest accuracy

Table III: Confusion Matrix: TransUNet Method

	Unlabel	Layers	Bedrock	Noise	P-Acc
Unlabel	**1.000**	0.0	0.0	0.0	1.000
Layers	0.0	**0.9960**	0.000	0.0038	0.9960
Bedrock	0.0029	0.0	**0.9325**	0.0644	0.9325
Noise	0.0004	0.0026	0.0048	**0.9919**	0.9916
U-Acc	0.9981	0.9984	0.9522	0.9864	0.9934

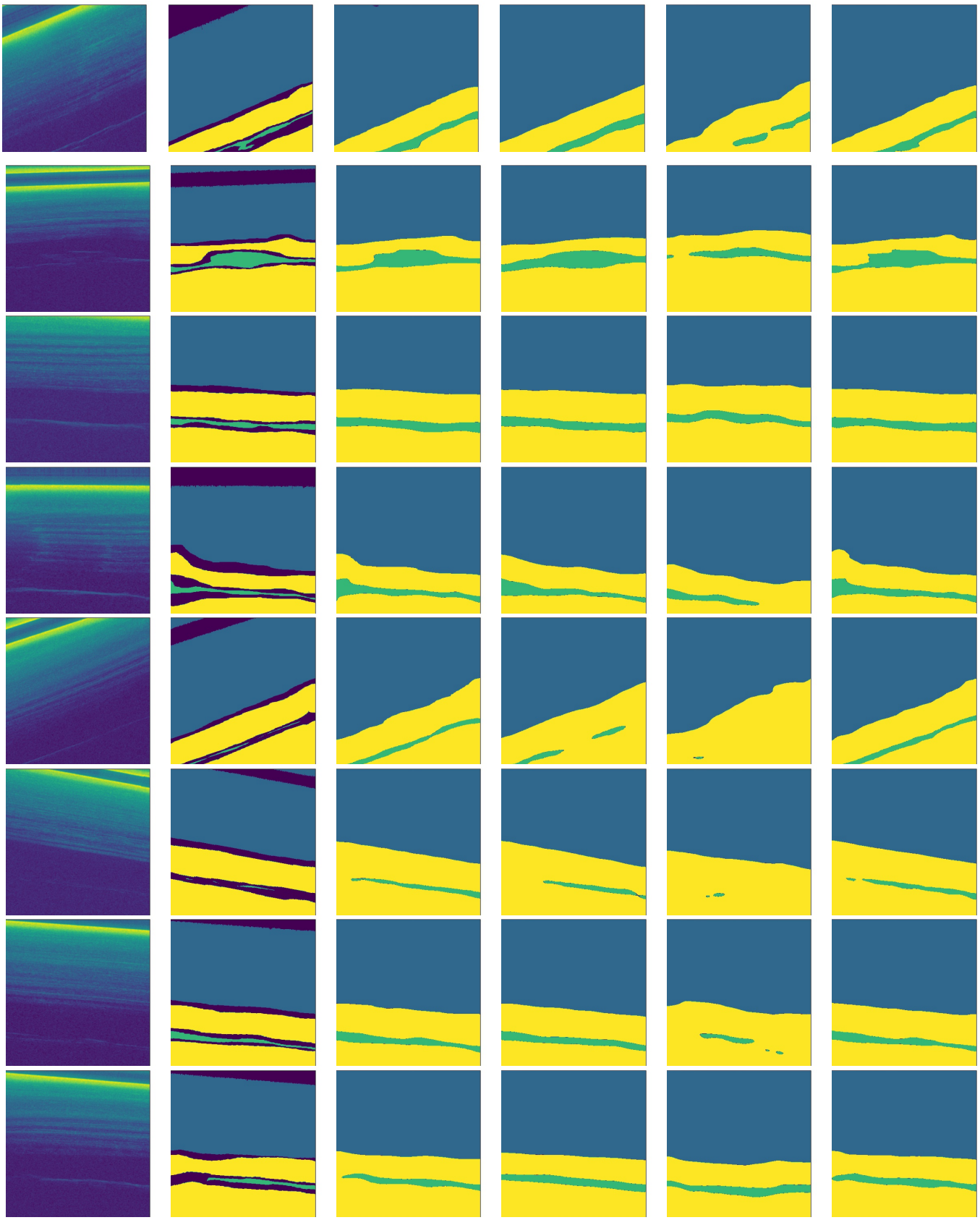
Table IV: Confusion Matrix: UNet Method

	Unlabel	Layers	Bedrock	Noise	P-Acc
Unlabel	**1.000**	0.0	0.0	0.0	1.000
Layers	0.0002	**0.9689**	0.0001	0.0303	0.9689
Bedrock	0.0067	0.0035	**0.4269**	0.5627	0.4269
Noise	0.0016	0.0041	0.0044	**0.9125**	0.9125
U-Acc	0.9940	0.9762	0.4958	0.8903	**0.9393**



(a) Radargram (b) Ground Truth (c) TransFuse (d) TransUNet (e) UNet (g) TransSounder

Figure 2: The Original Radargram (a), Ground Truth (b), and associated prediction maps are highlighted in this figure (from left to right)



(a) Radargram (b) Ground Truth (c) TransFuse (d) TransUNet (e) UNet (g) TransSounder

Figure 3: The Original Radargram (a), Ground Truth (b), and associated prediction maps are highlighted in this figure (from left to right)

Table V: Confusion Matrix: TransSounder

	Unlabel	Layers	Bedrock	Noise	P-Acc
Unlabel	**1.000**	0.0	0.0	0.0	1.000
Layers	0.0001	**0.9956**	0.0001	0.0041	0.9956
Bedrock	0.0019	0.0000	**0.9554**	0.0425	0.9554
Noise	0.0004	0.0037	0.0017	**0.9940**	0.9940
U-Acc	0.9980	0.9978	0.9811	0.9884	**0.9945**

values achieved by different architectures appear in bold. TransSounder achieved the highest OA value with 0.9945. Here, the attention mechanism in the joint CNN-Transformer encoder along with the attention mechanisms in BiFusion module helped TransSounder architecture to model the global and local spatial contexts more precisely. Although TransFuse incorporates attention mechanisms in both encoder and decoder architecture, it doesn't extract the high dimensional feature space at the initial stage, thus modelling a precise global context by utilizing the high dimensional local contexts becomes difficult throughout the downstream tasks in the encoder network. The performance of TransFuse and TransSounder architecture are closer in terms of assessment metrics (see Table I). Overall, the quantitative performance of TransSounder is better than other architectures. In terms of qualitative (Figures 2, and 3), and quantitative (Table I) assessment, the performance of the UNet was worst in comparison to other Transformer-based architectures. On several test sets, UNet was unable to predict the bedrock classes across the width of radargrams, thereby created fragmented patches in the final segmentation maps. Also, the pixels belonging to the bedrock class were misclassified as noise in UNet architecture.

In addition, we also tabulate the normalized confusion matrices for different architectures in Table II-V. The User's Accuracy and Producer's Accuracy are defined as U-Acc and P-Acc respectively. Several groups of observations can be made from these confusion matrices. The Transformer-based architectures (TransUNet, TransFuse, and TransSounder) achieved the highest rate of classification accuracy on the layers, bedrock, and noise class in comparison to the UNet model. Further, TransUNet had the highest classification accuracy for layers class with a rate of 0.9960 from Table III. For the noise class, TransSounder architecture outperformed other architectures. On the other hand, among the Transformer-based architectures, TransUNet received the highest error rate on the bedrock class with 0.0644 as highlighted in Table III. Here, the TransUNet architecture might be suffering from modeling and precise localization of the bedrock classes due to a lack of attention mechanism in the decoder network. The rate of this misclassification error on the bedrock was significantly higher for the UNet architecture (0.5627 from Table IV.

The poor performance of the UNet occurred due to its architectural constraints. The intrinsic locality of convolution operations does not explicitly model the global sequential dependencies of the localized feature spaces. Further, in the encoder part of UNet architecture, the concurrent CNN layers intrinsically capture the local spatial contextual prior, thereby missing the global spatial contexts. On the other hand, Transformer-based architectures model the sequence-to-sequence global spatial contexts. Additionally, Transformers retain the precise localization information while performing the downstream tasks in the encoder networks due to its positional encoding operators. Visual interpretations revealed from Figure 2 and 3 that sequential misalignment occurred as UNet doesn't have an explicit positional encoding operator as opposed to the Transformer-based architectures. Therefore, according to our observations, UNet architecture is susceptible to errors while capturing the long-range sequential contexts in an experimental framework.

In contrast to UNet, Transformer-based architectures were successfully able to preserve the sequentiality of the RS signals by capturing the global contexts with the positional encoding operators as depicted in Figures 2 and 3). The combination of MSA and MLP block in Transformers played a crucial role to model a long-range sequential information processing between the extracted tensors from the innermost CNN branch that embedded the high dimensional local spatial contexts. In other words, a global contextual relationship is established by Transformers between the tensors embedded local contexts over the spatial domain of RS signals. This paradigm is similar to the multi-hop representation in which the back-and-forth rapid information processing over the long range is a crucial operation. Consequently, the predicted outputs from these architectures were not affected by the misalignment problems between distinct classes. The back-scattering response between the free space and layers depicts a textural similarity in RS signals. We ignored inputting the ambiguous pixels into the encoder network. These ambiguous pixels from the free space of the radargrams were classified as layers by utilizing TransUNet, TransFuse, TransSounder, and UNet. Deciphering the correct mathematical descriptions behind these experimental results is beyond the scope of this paper.

V. Conclusion

In this paper, we construct a novel Transformer-based architectural framework named TransSounder by utilizing the distinct modules from recently developed TransUNet and TransFuse architecture. For the first time, we explore the potential of Transformer-based semantic segmentation architectures in the pixel-wise classification of radargrams. The Transformers model the long-range sequence-to-sequence global contextual prior that is important for radargrams as RS signals inherently depict sequential structures in the signal. However, Transformers lack the ability to model the local spatial contexts, whereas CNN can model the high-dimensional local spatial context more accurately. Thus, the TransSounder architecture employs a joint CNN-Transformer module as an encoder and utilizes the BiFusion module in the decoder architecture. Further, we perform a comparative analysis of the performance of TransSounder, TransUNet, TransFuse, and UNet.

Experimental results on the MCoRDS dataset confirm that Transformers have the potential to capture the long-range sequence-to-sequence global contexts more effectively than

CNN-based segmentation networks. Also, the TransSounder achieved the highest overall accuracy of 0.9945. Visual interpretation revealed that the TransSounder inherently preserves the spatial details more accurately due to the joint CNN-Transformer encoder network and the BiFusion modules incorporated in the decoder network. However, in this research work, the task is customized for a fixed sequence in which the target structure depicts a similar spatial pattern throughout the range of the RS signals. A future direction of this research work could be related to the transferability of the models to completely different sequential target structures associated with different bandwidths. In addition, we will explore the applicability of Transformer-based unsupervised segmentation methods for better generalizability.

Acknowledgment

This work was supported by the Italian Space Agency through the "Attivit'a Scientifiche per JUICE fase C-D" under Contract Agenzia Spaziale Italian - Istituto Nazionale di Astrofisica (ASI-INAF) and Contract 2018-25-HH.0.

References

[1] 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 *2012 IEEE International Geoscience and Remote Sensing Symposium*, 2012, pp. 4457–4460.

[2] A. Shepherd and H. J. Ivins, ..., "A reconciled estimate of ice-sheet mass balance," *Science*, vol. 338, no. 6111, pp. 1183–1189, 2012. [Online]. Available: https://science.sciencemag.org/content/338/6111/1183

[3] L. Carrer, C. Gerekos, F. Bovolo, and L. Bruzzone, "Distributed radar sounder: A novel concept for subsurface investigations using sensors in formation flight," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 12, pp. 9791–9809, 2019.

[4] V. V. Bogorodsky, C. R. Bentley, and P. E. Gudmandsen, "Radioglaciology," 1986.

[5] 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 Transactions on Geoscience and Remote Sensing*, pp. 1–19, 2021.

[6] A. Ferro and L. Bruzzone, "Analysis of radar sounder signals for the automatic detection and characterization of subsurface features," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 11, pp. 4333–4348, 2012.

[7] A.-M. Ilisei and L. Bruzzone, "A system for the automatic classification of ice sheet subsurface targets in radar sounder data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 6, pp. 3260–3277, 2015.

[8] S. Lee, J. Mitchell, D. J. Crandall, and G. C. Fox, "Estimating bedrock and surface layer boundaries and confidence intervals in ice sheet radar imagery using mcmc," in *2014 IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 111–115.

[9] M. Xu, D. J. Crandall, G. C. Fox, and J. D. Paden, "Automatic estimation of ice bottom surfaces from radar imagery," in *2017 IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 340–344.

[10] 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 *2021 IEEE International Conference on Multimedia and Expo (ICME)*, 2021, pp. 1–6.

[11] 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 *Proceedings of the 2019 8th International Conference on Computing and Pattern Recognition*, ser. ICCPR '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 199–204. [Online]. Available: https://doi.org/10.1145/3373509.3373585

[12] M. Rahnemoonfar, D. Varshney, M. Yari, and J. Paden, "Deep ice layer tracking and thickness estimation using fully convolutional networks," *CoRR*, vol. abs/2009.00191, 2020. [Online]. Available: https://arxiv.org/abs/2009.00191

[13] X. Wang, R. B. Girshick, A. Gupta, and K. He, "Non-local neural networks," *CoRR*, vol. abs/1711.07971, 2017. [Online]. Available: http://arxiv.org/abs/1711.07971

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[15] C. Yun, S. Bhojanapalli, A. S. Rawat, S. J. Reddi, and S. Kumar, "Are transformers universal approximators of sequence-to-sequence functions?" *CoRR*, vol. abs/1912.10077, 2019. [Online]. Available: http://arxiv.org/abs/1912.10077

[16] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, vol. abs/2010.11929, 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[18] Q. Yu, L. Xie, Y. Wang, Y. Zhou, E. K. Fishman, and A. L. Yuille, "Saliency transformation network: Incorporating multi-stage visual cues for pancreas segmentation," *CoRR*, vol. abs/1709.04518, 2017. [Online]. Available: http://arxiv.org/abs/1709.04518

[19] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *CoRR*, vol. abs/2102.04306, 2021. [Online]. Available: https://arxiv.org/abs/2102.04306

[20] Y. Zhang, H. Liu, and Q. Hu, "Transfuse: Fusing transformers and cnns for medical image segmentation," *CoRR*, vol. abs/2102.08005, 2021. [Online]. Available: https://arxiv.org/abs/2102.08005

[21] C. M. Gifford, G. Finyom, M. Jefferson, M. Reid, E. L. Akers, and A. Agah, "Automated polar ice thickness estimation from radar imagery," *IEEE Transactions on Image Processing*, vol. 19, no. 9, pp. 2456–2469, 2010.

[22] M. Rahnemoonfar, G. C. Fox, M. Yari, and J. Paden, "Automatic ice surface and bottom boundaries estimation in radar imagery based on level-set approach," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 9, pp. 5115–5122, 2017.

[23] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," 2016.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, May 2015.

[25] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. S. Torr, and L. Zhang, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," 2021.

[26] B. Lin, G. Yang, Q. Zhang, and G. Zhang, "Semantic segmentation network using local relationship upsampling for remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2021.

[27] S. Zheng, P. An, Y. Zuo, X. Zou, and J. Wang, "Depth map upsampling using segmentation and edge information," in *Image and Graphics*, Y.-J. Zhang, Ed. Cham: Springer International Publishing, 2015, pp. 116–126.

[28] A. Milioto, J. Behley, C. McCool, and C. Stachniss, "Lidar panoptic segmentation for autonomous driving," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 8505–8512.

[29] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," 2020.

[30] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *CoRR*, vol. abs/1709.01507, 2017. [Online]. Available: http://arxiv.org/abs/1709.01507

[31] S. Woo, J. Park, J. Lee, and I. S. Kweon, "CBAM: convolutional block attention module," *CoRR*, vol. abs/1807.06521, 2018. [Online]. Available: http://arxiv.org/abs/1807.06521

[32] M. Yi-de, L. Qing, and Q. Zhi-bai, "Automated image segmentation using improved pcnn model based on cross-entropy," in *Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004.*, 2004, pp. 743–746.

Raktim Ghosh received Bachelor of Engineering (BE) degree in Mining Engineering from the Indian Institute of Engineering Science and Technology, Shibpur and received Master of Science degree in Geo-information Science and Earth Observation with a specialisation in Geoinformatics under the joint education programme (JEP) between the Indian Institute of Remote Sensing, ISRO and Faculty of Geo-information Science and Earth Observation (ITC), University of Twente. Currently, he is pursuing the PhD 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 active research interests are related to the automatic analysis of radar sounders data for investigating the subsurface features.

Francesca Bovolo (S'05–M'07–SM'13) 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, and a member of the Remote Sensing Laboratory, Trento. She is one of the co-investigators of the Radar for Icy Moon Exploration instrument of the European Space Agency Jupiter Icy Moons Explorer and 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, and Light Detection and Ranging (LiDAR) and radar sounders. She conducts research on these topics within the context of several national and international projects. Dr. Bovolo is a member of the program and scientific committee of several international conferences and workshops. She 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 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 is the Publication Chair of the International Geoscience and Remote Sensing Symposium in 2015. She has been an Associate Editor of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING since 2011 and the Guest Editor of the Special Issue on Analysis of Multitemporal Remote Sensing Data of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. She is a referee for several international journals.