IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium | 979-8-3503-2010-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/IGARSS52108.2023.10281822

ECHOVIT: VISION TRANSFORMERS USING FAST-AND-SLOW TIME EMBEDDINGS

Oluwanisola Ibikunle[1], Debvrat Varshney[2], Jilu Li[1], Maryam Rahnemoonfar[3], John Paden[1]

[1] Center for Remote Sensing and Integrated Systems (CReSIS), University of Kansas, Lawrence, KS, USA
[2] Department of Information Systems, University of Maryland Baltimore County, Baltimore, MD, USA
[3] Department of Computer Science and Engineering, Lehigh University, Bethlehem, PA, USA

ABSTRACT

This paper details the preliminary efforts of applying the deep learning transformer architecture to automatically track annual layer stratigraphy in echogram images obtained from mapping near-surface ice layers using airborne radars. Following the success of the transformer architecture in the natural language processing and computer vision communities, we explore a variant termed Echogram Vision Transformer (EchoViT) on the radar echogram layer tracking (RELT) problem. The proposed approach divides the echogram images into patches using different schemes inspired by tokenization methods in natural language processing. We then apply a soft-attention mechanism to model interdependencies between the patches, capturing spatiotemporal stratigraphic information. Experiments conducted on the CREED dataset demonstrate the superiority of transformer-based architectures over existing convolutional-based architectures. Furthermore, the EchoViT fast-time and EchoViT slow-time patchifying schemes achieved precise tracking of the layers with submeter MAE of 3.39 and 3.55, respectively, while the use of cropped patches led to suboptimal results.

***Index Terms*—** deep learning, transformer, soft-attention, ViT, EchoViT

1. INTRODUCTION

Climate change is a critical global issue that has far-reaching implications for the Earth's environment and ecosystems. Increased mass loss from polar ice sheets in recent years due to global warming has led to the quest for methods that can accurately measure the annual accumulation rate to better predict future sea-level rise [1, 2]. The Snow Radar [3], an airborne radar developed at the Center of Remote Sensing and Integrated Systems (CReSIS), has been flown on several missions over the Greenland Ice Sheet (GrIS) to map shallow and near-surface snow. Echogram images[3, 4] created from the data collected reveal the spatiotemporal accumulation patterns that can be tracked to estimate annual snow accumulation [4]. However, manually tracking these layers in the echogram images is both laborious and ineffective considering the volume of data that has been created. As such, there is an urgent need to develop robust, scalable, automatic, and accurate layer tracking algorithms.

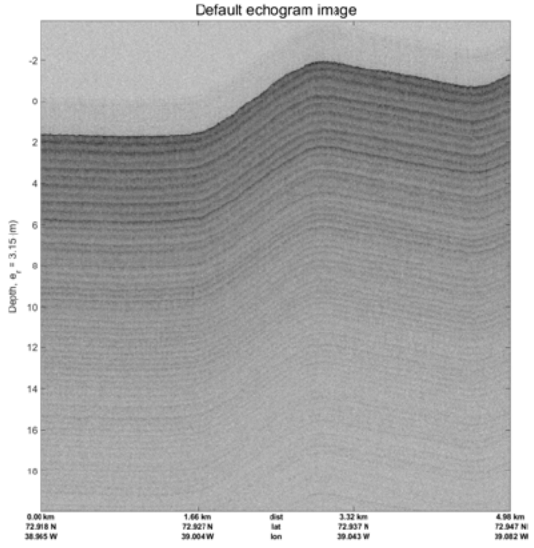


Fig. 1. Snow Radar echogram

Figure 1 is an echogram image created from Snow Radar data collected around the summit of Greenland in 2012. The echogram image represents the data matrix obtained by taking the logarithm of the power-detected, coherently and incoherently averaged received backscatter from the nadir elevation angle. The horizontal axis corresponds to the direction of flight, referred to as the "slow-time" axis, capturing the spatial variation of accumulation patterns in the columns. The vertical axis represents the fast-time axis, which indicates the radar propagation time to each detected snow layer. It is called "fast" relative to the slower aircraft speed in the flight direction because the transmitted signal travels at a speed comparable to the speed of light. This axis can be utilized to infer the depth of each layer and the interannual accumulation rate.

To track these layers, numerous deep learning algorithms [5, 6, 7, 8] have been developed for their ability to learn from

Center for Remote Sensing and Integrated Systems (CReSIS) is the new name of the Center for Remote Sensing and Ice Sheets (CReSIS).

This work is supported by NSF IIS-1838236 and ACI-1443054.

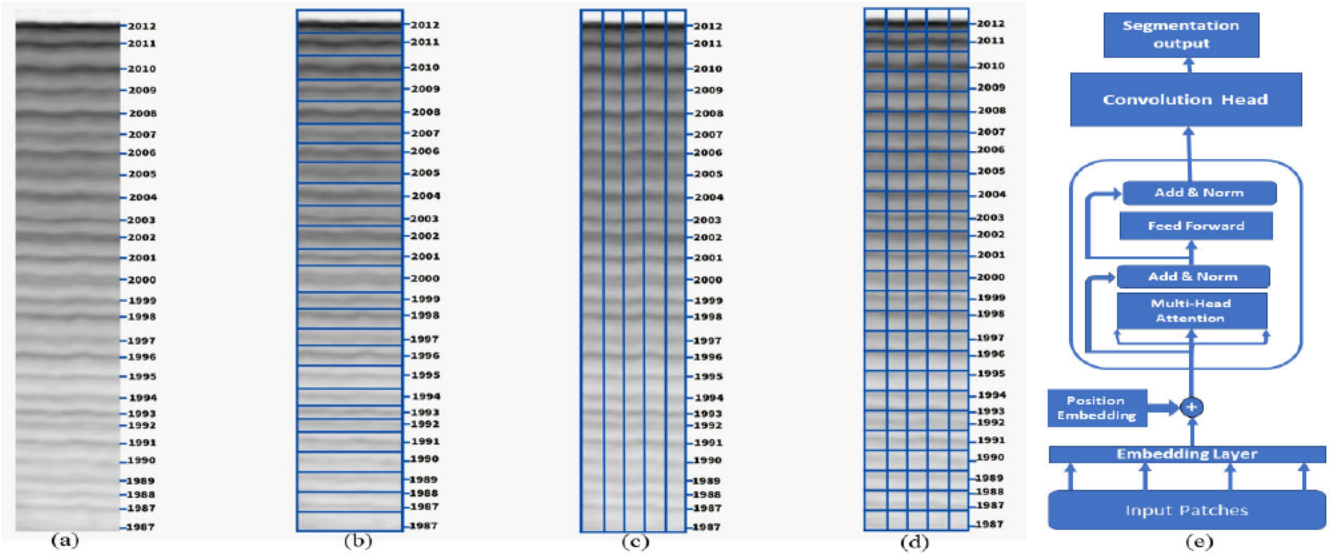


Fig. 2. (a) Surface-flattened echogram (b.) Slow-time patch (c.) Fast-time patch (d.) Cropped patch (e.) Pixel-wise EchoViT architecture

data and create models with broad applicability. The remarkable performance of transformer-based models in challenging language understanding [9] and computer vision tasks [10] has sparked interest in exploring this architecture for the radar echogram layer tracking (RELT) problem. In this work, we investigate a variant called the Echogram Vision Transformer (EchoViT).

The outstanding performance of Vision Transformers can be attributed to their soft-attention mechanism, which captures intricate relationships between input sequential data. Unlike conventional convolutional neural networks, transformers can model global dependencies in input tokens right from the initial network layer. However, the inherent quadratic complexity with respect to the length of the input sequence poses a challenge, especially for high-resolution geospatial echogram images. To overcome this challenge for optical images, various intelligent ”patchifying” schemes have been proposed, including cropped patches, hierarchical pyramidal schemes, and local windows.

However, unlike electro-optical images, echograms naturally exhibit patchification of temporal accumulation patterns along the fast-time axis and spatial patterns across the slow-time axis. In this study, we explore the application of the transformer's soft-attention mechanism in combination with different patchifying schemes to create a variant of the vision transformer called Echogram Vision Transformer (EchoViT) specifically for the radar echogram layer tracking problem. Our hypothesis is that leveraging the effectiveness of transformers and employing the appropriate patchifying scheme will enhance global feature extraction from echogram images and improve the accuracy of EchoViT in tracing accumulation layers.

We conducted experiments on the test set of the CREED dataset (details in Section 3) to evaluate our approach. These experiments indicate that two of the patchifying schemes, EchoViT fast-time (FT) and EchoViT slow-time (ST) patches, outperform existing deep learning architectures such as DeepLabv3, UNet, and FCN on the RELT problem.

2. METHODOLOGY

Given an enhanced echogram image, the goal of the deep learning network is to identify and track snow layer pixels in the along-track axis. We approach this problem as a binary segmentation task and propose the EchoViT architecture, which incorporates a pairwise self-attention mechanism to capture spatial correlations between echogram patches. EchoViT is an encoder-only transformer architecture that features a specifically designed binary segmentation output layer tailored to the input's patchifying scheme.

The input to the Encoder layer **E** is the patched echogram pixels mixed with corresponding learnable positional embedding Z_{pos} as shown in Figure 2(b)-2(d). Concretely, given a grayscale input echogram $\mathbf{G} \in \mathbb{R}^{N_t \times N_x}$ with $\mathbf{G}(m,n) = \{g \in \mathbb{R} \mid 0 \leq g \leq 1,\ m = [1, ..., N_t],\ n = [1, ..., N_x]\}$ where N_t is the number of fast-time bins and N_x is the number of slow-time bins. We explore three patching schemes:

1. Fast time patch $P_{ft} \in \mathbb{R}^{N_t \times 1}$ to give a patch sequence $Z_{ft} = \mathcal{L}\{[P_{ft_1}, P_{ft_2}, ..., P_{ft_{Nx}}]\} + Z_{pos_{ft}}$
2. Slow time patch $P_{st} \in \mathbb{R}^{1 \times N_x}$ to give a patch sequence $Z_{st} = \mathcal{L}\{[P_{st_1}; P_{st_2}; ...; P_{ft_{Nt}}]\} + Z_{pos_{st}}$
3. Cropped patch $P_{cr} \in \mathbb{R}^{b \times b}$ to give a patch sequence $Z_{cr} = \mathcal{L}\{[P_{cr_1}, P_{cr_2}, ..., P_{cr_N}]\} + Z_{pos_{cr}}$

where b is the cropped patch size, $N = \frac{N_t * N_x}{b^2}$. $\mathcal{L}\{.\}$ is the linear embedding operation and Z_{pos} is the corresponding positional embedding.

The EchoViT's encoder is similar to the classic ViT [10]. The input tokens are layer-normalized, then passed to the multihead self-attention layer whose intermediate output is added to a residual copy of the input before being passed to another layer-normalizing module. A position-wise feed-forward network is subsequently used to further enhance the representation of each position in the sequence, to output a comprehensive and contextualized encoding of the input sequence.

The encoder's output maintains the same dimension as the input tokens, except for the cropped patch scheme. The cropped patch output is reshaped to match the desired $\mathbb{R}^{N_t \times N_x}$ dimension for pixel-wise prediction. For simplicity, all three schemes use a 1×1 filter convolutional layer with sigmoid activation function as the output module.

3. DATASET AND EXPERIMENTAL SETUP

3.1. Dataset

The EchoViT models were trained on the Climate-change Radar Enhanced Echogram Dataset (CREED) which consists of 11307 enhanced echogram images for training and 1302 images for validation. The test set is divided into L1, L2, and L3 regions based on echogram image quality degradation, with L1 having the highest quality. For the final evaluation, we tested the models on 128 echogram images in the L1 segment.

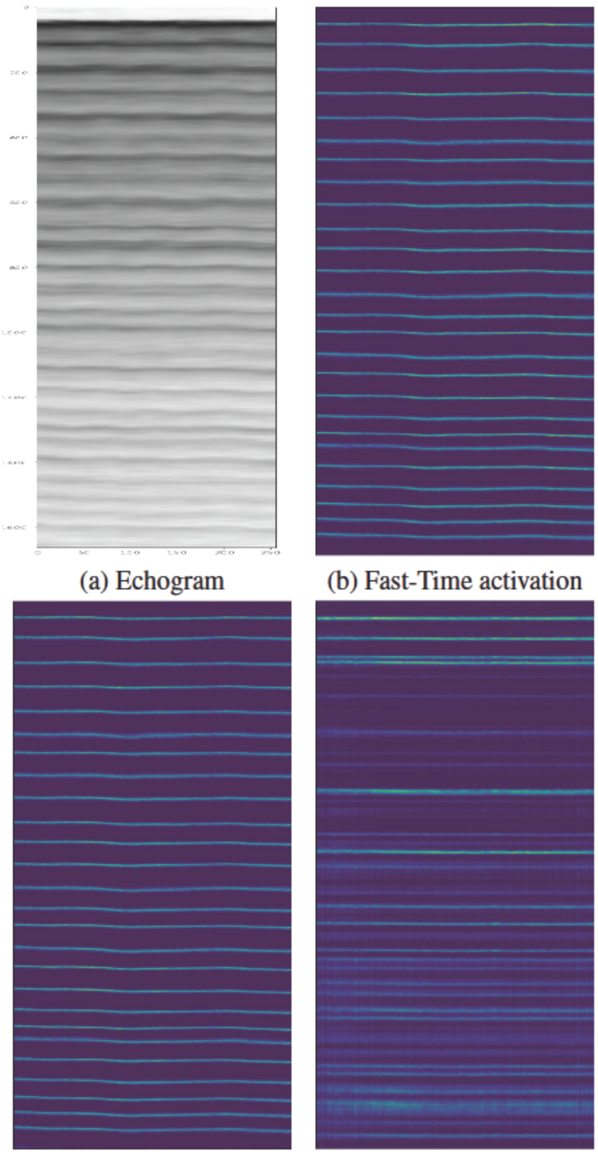
3.2. Experimental setup

Each echogram image has a fixed dimension $N_t = 1664$ and $N_x = 256$. We avoided reshaping the echograms as this could potentially distort the depth information in the vertical axis.

The training hyperparameters used include $T = 10$ attention heads and attention layers. We trained each model for 200 epochs using a batch size of 8. The training was performed on a Core i9 machine with an RTX A5000 GPU

Models	MAE
UNet	8.43
FCN	6.23
DeepLabv3	5.98
AttentionUNet	4.03
ResNet50	3.84
EchoViT ST Embed (ours)	3.55
EchoViT FT Embed (ours)	**3.39**

Table 1. Mean Absolute Error (MAE) in terms of number of pixels for each model on the L1 test set.



(a) Echogram

(b) Fast-Time activation

(c) Slow-Time activation

(d) Cropped activation

Fig. 3. Visualization of the EchoViT final activation maps for the three patching schemes

4. RESULT AND DISCUSSION

In this study, we investigated the performance of the transformer architecture to radar echogram images using different patching schemes. To evaluate the model's performance, the sigmoid outputs were first thresholded to generate a binary (layer or no-layer pixel) output. These binary outputs are subsequently post-processed to uniquely identify and track each accumulation layer.

The mean absolute error of each layer in the 128 echograms in the L1 test set was calculated by comparing each layer to the corresponding manually annotated ground truth. Notably, when using the cropped patch, as depicted in Figure 3, the model exhibited poor performance and failed to accurately identify the snow layers in a distinctive manner. The cropped patching scheme, although trained with similar hyper-parameters and model architecture, falls short when compared to the other schemes. This is likely because the rectangular cropping of the echograms distorts the naturally occurring spatial and temporal patterns captured by the echogram images.

5. CONCLUSION

The EchoViT (Echogram Vision Transformer) model takes its inspiration from the classic ViT designed to investigate different "patchifying" schemes for remotely-sensed echogram images. To generate the desired dense pixel-wise classification output, we designed a simple fully convolutional prediction module to process the output of the encoder. Our experiments reveal that fast-time and slow-time patching schemes correctly model the input-output relationship of the echograms and the internal layers by correctly tracking the snow accumulation layers in the echograms.

More so, the EchoViT establishes a new state-of-the-art performance of 3.3 overall mean absolute error (MAE) on the L1 test segment of the CREED dataset which is equivalent to a sterling submeter ($\sim$ 14cm) tracking error. This surpasses the previous benchmarks by top convolutional-based models such as UNet and FCN. This demonstrates the viability of Transformer architectures to solve the radar echogram layer tracking (RELT) problem prompting the need to explore transformer-based semi-supervised and unsupervised models to take advantage of the ever-growing large repertoire of unlabelled remotely-sensed data.

6. REFERENCES

[1] Kurt M Cuffey and Shawn J Marshall, "Substantial contribution to sea-level rise during the last interglacial from the greenland ice sheet," *Nature*, vol. 404, no. 6778, pp. 591–594, 2000.

[2] Andrew Shepherd and Duncan Wingham, "Recent sea-level contributions of the antarctic and greenland ice sheets," *science*, vol. 315, no. 5818, pp. 1529–1532, 2007.

[3] Fernando Rodriguez-Morales, Sivaprasad Gogineni, Carlton J Leuschen, John D Paden, Jilu Li, Cameron C Lewis, Benjamin Panzer, Daniel Gomez-Garcia Alvestegui, Aqsa Patel, Kyle Byers, et al., "Advanced multifrequency radar instrumentation for polar research," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 5, pp. 2824–2842, 2013.

[4] Lora S Koenig, Alvaro Ivanoff, Patrick M Alexander, Joseph A MacGregor, Xavier Fettweis, Ben Panzer, John D Paden, Richard R Forster, Indrani Das, Joesph R McConnell, et al., "Annual greenland accumulation rates (2009–2012) from airborne snow radar," *The Cryosphere*, vol. 10, no. 4, pp. 1739–1752, 2016.

[5] Masoud Yari, Oluwanisola Ibikunle, Debvrat Varshney, Tashnim Chowdhury, Argho Sarkar, John Paden, Jilu Li, and Maryam Rahnemoonfar, "Airborne snow radar data simulation with deep learning and physics-driven methods," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 12035–12047, 2021.

[6] Raktim Ghosh and Francesca Bovolo, "Transsounder: A hybrid transunet-transfuse architectural framework for semantic segmentation of radar sounder data," *IEEE Transactions on Geoscience and Remote Sensing*, 2022.

[7] Maryam Rahnemoonfar, Masoud Yari, John Paden, Lora Koenig, and Oluwanisola Ibikunle, "Deep multi-scale learning for automatic tracking of internal layers of ice in radar data," *Journal of Glaciology*, vol. 67, no. 261, pp. 39–48, 2021.

[8] Debvrat Varshney, Maryam Rahnemoonfar, Masoud Yari, John Paden, Oluwanisola Ibikunle, and Jilu Li, "Deep learning on airborne radar echograms for tracing snow accumulation layers of the greenland ice sheet," *Remote Sensing*, vol. 13, no. 14, pp. 2707, 2021.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.