Image & Video Undestanding - Project Report

Davide Gattolin (Mat. 885797) *Ca' Foscari University of Venice*

1 Introduction

The problem & previous literature

The cryosphere is one of the major factors influencing the climatic conditions on the planet. Ice sheets and glaciers influence the global energy budget, ocean currents, and sea level. In recent years, global warming has been the leading cause of the rapid reduction in the mass of ice. The melting of glaciers is one of the factors causing a rise in the level of the ocean, which influences the threatened ecosystems and human habitations all over the planet. The reduction in the area covered with ice is also responsible for the rise in the absorption of solar energy, causing an increase in global warming creating a self-feeding loop. The importance of these analysis is further discussed in [1] in which the importance of studying grounding lines is highlighed.

1.1 Problem Statement

In order to be able to make accurate predictions regarding glacier/ice sheet behavior in the future, it has become important to understand the processes underlying grounding line motion. This becomes particularly significant with respect to our understanding of ice/ocean interactions and their implications in understanding climate scenarios regarding sea level rise projections.

In order to detect the grounding line a series of satellites have been adopted to capture DInSAR* images of the glaciers in the Antarctica. This type of images capture the vertical shift in the icesheet due to tides, which presents peculiar patterns in the grounding lines areas; since DInSAR uses coherent microwave radar signals, data acquisition is essentially unaffected

*Differential INterferometry Synthetic Aperture Radar

by cloud cover, fog, and moderate precipitation, allowing for all-weather, day-and-night surface deformation monitoring. Up to this moment most of the grounding lines detection work was done by hand on the phase component of the DInSAR images.

This goal of this project is to develop a ML model capable of detecting grounding lines using the input set of DInSAR images, as well as the manually created mask, for the purpose of automating such a task which can be described as a semantic segmentation one.

1.2 Literature Overview

In this section are discussed some of the papers that inspiered this project and which summarize better the work done on the topic of GL segmentation.

1.2.1 Mohajerani et al. (2021): First CNN-Based Grounding Line Detector

The first deep learning model, proposed by Mohajerani et al. (2021) [2], automates GL identification using DInSAR data. It consists of a 40-layer CNN using an asymmetrical encoder-decoder network and parallel atrous convolutional operations achieved through an Atrous Spatial Pyramid Pooling (ASPP) module. It enables the network to detect multi-scale spatial information on tidal flexure patterns. The CNN uses a weighted binary cross-entropy loss function, which assigns greater weightage to the false negatives given the great imbalance between the two classes. The network estimated a mean positional accuracy of about 232m̃ against manually interpreted GLLocs, which can be as variable as human interpreters.

1.2.2 Ross et al. (2024): Extension to X-Band Data

Ross et al. (2024) [3] extended the approach of Mohajerani et al. to X-band COSMO-SkyMed data, which offer higher spatial resolution and denser interferometric fringes than C-band SAR. Their model achieved improved accuracy, with a mean GL difference of approximately 183m̃ relative to manual delineations.

1.3 Current Challenges and Limitations

Despite the remarkable progress made possible by deep learning, several challenges remain unsolved. This is the challenge of neural networks and any other analytical approach in tackling high dynamic and noisy conditions such as those that occur in the Amundsen Sea Embayment, with its extremely high rates of ice flow and strong tidal controls, resulting in highly dynamic and, mostly, non-stationary interferometry results, which often require visual validation from the analysts. Automatic approaches also tend to over-smooth the grounding lines along highly geometrically sinuous areas. Then there is the generalization for different areas, a task which is currently ongoing in research. Temporally trained models on different acquisitions in the same region often outperform models trained from geographically different areas, therefore suggesting imperfect transferability across different ice sheet conditions.

The main contributions of this project aim to be:

- **Analysis of patch size effects on segmentation performance.** We investigate the impact of different patch sizes (128×128 and 256×256) in comparison with the 521×512 patches adopted in previous work. To this end, two new datasets were generated by accounting for patch overlap and the presence of empty patches (without grounding lines), in order to build datasets that are balanced, diverse, and representative of the underlying spatial variability.
- **Design of a loss function robust to class imbalance.** Owing to the extreme imbalance in the foreground and background in the grounding line mask, we introduce a compound loss function composed of Binary Cross-Entropy and Focal Tversky Loss. This allows the false positive and false negative weights to be directly controlled and adjusted. The compound weights of the two loss terms are identified through parameter ablation on a validation set, different combination of parameters for the FTL where tested too.
- **Evaluation of advanced U-Net-based architectures.** In addition to the standard U-Net used in the reference Mohajerani et al. work, we evaluate two enhanced architectures:
  - **U-Net**, used as a baseline to provide a fair comparison with prior literature.
  - **Attention U-Net**, which integrates attention gates in the skip connections.
  - **U-Net++ with deep supervision**, which introduces nested and densely connected skip pathways to reduce the semantic gap between encoder and decoder feature maps. The use of deep supervision further improves gradient propagation and encourages discriminative feature learning at multiple spatial scales, leading to better localization of fine structures and increased robustness across varying patch sizes.

2 Data and Methods

The dataset & the architectures

2.1 Dataset

Starting from the DInSAR detections of the Sentinel-1 satellite taken from the Getz Ice Shelf area, two datasets were created to investigate the impact of patch size on network performance: the first one contains 128×128 patches, while the second one contains 256×256 patches.

For each DInSAR scene, the complex signal was decomposed into phase and magnitude, which were used as the two input channels of the network, while the corresponding grounding lines were rasterized into binary masks in the same spatial reference system. Grounding line geometries were burned into the raster grid using a pixel-level rasterization procedure, ensuring perfect spatial alignment between the input images and the segmentation masks. During rasterization, two strategies are available: the standard Bresenham-based rasterization, which marks only pixels whose centers are intersected by the geometry, and the `all_touched` option, which labels all pixels intersected by the grounding line. In this work, the latter was adopted to reduce label sparsity and avoid discontinuities in the ground truth masks.

Patch extraction was performed through a stochastic sampling strategy designed to ensure both spatial diversity and sufficient coverage of grounding line pixels. Only patches containing at least one grounding line pixel were selected for the main training set, while an additional set of patches without grounding lines was generated to model background variability. To avoid excessive spatial redundancy, a coverage map was maintained for each scene and used to penalize repeatedly sampled regions, enforcing a minimum percentage of previously unseen pixels within each extracted patch.

For each patch size, patches were sampled until a predefined upper limit was reached, yielding two size-specific datasets stored in separate directories. Each sample consists of a two-channel DInSAR patch (phase and magnitude) and the corresponding binary grounding-line mask. Before dividing the sets into training, testing and validation to each were added some patches without any grounding line for a more precise training procedure.

2.2 Network Architectures, Loss Function and Training Setup

In these experiments, only CNNs based U-Nets were used because of their applicability for geo physical object segmentation, when there are thin structures present along with potential class imbalance.

Network Architectures. The three architectures evaluated are the following:

U-Net: employed as a baseline model, allowing it to be compared directly with that in Mohajerani et al. It is composed of an encoder-decoder structure with skip connections that convey spatial information at a resolution that passes from the contracting part of the model to the expanding part.

Attention U-Net: adds to the standard U-Net structure attention gates at skip connections. The purpose of these attention gates is to assign weights to spatial features based on their relevance to the class of interest. The relevance of background regions to the class of interest is suppressed using this mechanism. In ground line segmentation tasks, the foreground region is very sparse and lies in highly variable ice-ocean texture patterns.

U-Net++ with deep supervision: feature fusion techniques are again upgraded with the addition of nested skip connections. Deep supervision is another technique in which additional prediction heads are added to the decoder network. In this method, outputs are generated from different levels of the decoder network. With this approach, predictions are directly predictive. With this method, thin structures in images are localized.

Loss Function. Given the goal of predicting a line we expect a severe class imbalance, with pixels belonging to the grounding line occupying only a small fraction of each patch. To address this, a composite loss function was employed:

$$\mathcal{L} = \lambda \, \mathcal{L}_{BCE} + (1 - \lambda) \, \mathcal{L}_{FocalTversky}, \qquad (1)$$

where the Binary Cross-Entropy (BCE) part has a stabilizing effect during training at the pixel level, and the Focal Tversky Loss (first proposed in [4]) encourages the FNs vs FPs trade-off in a controlled way.

The weighting coefficient λ was selected via ablation on a validation set to optimize segmentation performance under class imbalance.

Training Parameters All models were trained on patches of size either 128×128 or 256×256, with two-channel inputs corresponding to DInSAR phase and magnitude. The output was a single-channel binary segmentation mask resulting from the rasterization described previously. The optimization was conducted with Adam, and convergence was achieved with early stopping with regards to validation loss for all neural networks.

Architectures (layer depths and heads)

- **U-Net** (`in_ch=2`, `out_ch=1`): 4 encoder stages, each with two `Conv(3x3)+BN+ReLU`: channels = {4, 8, 16, 32}; max-pooling 2x2 between stages. Bottleneck: ASPP with dilation rates {1,6,12,18} and a pooled branch, projected back to 32 channels. Decoder: 4 up-conv (2x2) stages with skip concatenation and two `Conv(3x3)+BN+ReLU` per stage: channels = {16, 8, 4, 2}. Final head: `Conv(1x1)` $\rightarrow$ 1 logit map.
- **Attention U-Net** (`in_ch=2`, `out_ch=1`): Same encoder channel progression as U-Net with 2-layer conv blocks. Bottleneck: ASPP with dilation rates {1,6,12,18}. Decoder: up-conv stages with attention gates on each skip (gating features from decoder, skip features from encoder); channels mirror U-Net (16, 8, 4, 2). Final head: `Conv(1x1)` $\rightarrow$ 1 logit map.
- **U-Net++** (`in_ch=2`, `out_ch=1`, `base_ch=32`, deep supervision): Encoder depths: features {32, 64, 128, 256, 512} with two `Conv(3x3)+BN+ReLU` per node and 2x2 pooling. Dense decoder grid with nested skip connections: nodes $x_{0,1}$, $x_{0,2}$, $x_{0,3}$, $x_{0,4}$ aggregate progressively upsampled features. Deep supervision heads: four `Conv(1x1)` on $\{x_{0,1}, x_{0,2}, x_{0,3}, x_{0,4}\}$, each yielding a logit map (weights optionally applied in loss).

Training setup

- Optimizer: Adam, learning rate 1×10^{-4}, weight decay 1×10^{-5}.
- Batch size: 25; with splits 70/15/15 (train/val/test) on concatenated GL+GL-less datasets (GL-less images capped at 1000 samples).
- Epochs: up to 500; early stopping patience: 12; gradient clipping: 1.0.
- Loss: Focal Tversky (FTL) with (α, β, γ) swept on a validation set to do ablation; optional BCE mixing with weight `bce_weight`; deep-supervision weighting for U-Net++ optionally applied.

Validation sweeps Validation sweeps were conducted in a grid-search fashion over the discrete set of architectural and loss-related configurations.

- FTL parameters (α, β, γ) combinations: $(0.7, 0.3, 1.33)$, $(0.8, 0.2, 1.33)$, $(0.75, 0.25, 1.33)$, $(0.85, 0.15, 1.33)$.
- BCE mix weights: 0.20, 0.25, 0.30, 0.35.
- Deep-supervision weights (U-Net++ only): `None`, $(0.1, 0.2, 0.3, 0.4)$, $(0.05, 0.1, 0.2, 0.65)$.
- Sweep protocol: 8 epochs per config, Adam (1×10^{-4}, weight decay 1×10^{-5}), batch size 25; best config selected by validation loss.

3 Results

Evaluation methods & results discussion

In this section are discussed the results of the different NNs used in this project.

3.1 Metrics

In order to evaluate and compare results we need to choose a set of metrics to evaluate and understand the performances of the NNs. In the following table are illustrated the metrics with their strenghts and limitations.

Precision Precision measures the reliability of positive predictions and is defined as:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

It quantifies how many of the predicted positive pixels actually belong to the target class. A low precision indicates the presence of many false positives, typically caused by over-segmentation or boundary leakage.

In semantic segmentation, high precision corresponds to conservative predictions that avoid including back-

ground pixels inside the object region. However, optimizing only precision may lead to under-segmentation and fragmented objects.

Recall Recall measures the ability of the model to detect all relevant pixels:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

It represents the fraction of ground-truth pixels that are correctly identified. Low recall indicates that the model fails to capture parts of the target object, resulting in false negatives.

High recall is desirable when missing parts of the object is costly, but it often comes at the expense of increased false positives.

Dice Coefficient The Dice Similarity Coefficient (DSC) evaluates the overlap between prediction and ground truth:

$$Dice = \frac{2TP}{2TP + FP + FN} \tag{4}$$

Dice provides a balanced measure between precision and recall and is widely used in medical image segmentation. However, it is insensitive to spatial distribution: two segmentations with similar overlap may have very different boundary accuracy.

As a consequence, Dice alone is insufficient for evaluating fine-grained contour quality.

Boundary F-score (BFScore) The Boundary F-score evaluates the alignment between predicted and ground-truth contours. It is computed by matching boundary pixels of the prediction with boundary pixels of the ground truth within a predefined spatial tolerance.

Given boundary precision P_b and boundary recall R_b, BFScore is defined as:

$$BF = \frac{2P_bR_b}{P_b + R_b} \tag{5}$$

BFScore directly evaluates contour quality and is particularly suited for thin structures such as grounding lines, where boundary localization is more informative than region overlap. Unlike Dice and IoU, BFScore focuses exclusively on edge alignment, making it complementary to region-based metrics.

The metric is sensitive to the tolerance parameter, which determines the allowable distance for matching boundary pixels.

Intersection over Union with Spatial Tolerance The Intersection over Union (IoU) measures the overlap between predicted and ground truth regions:

$$IoU = \frac{TP}{TP + FP + FN} \tag{6}$$

In grounding line segmentation, small spatial misalignments between prediction and ground truth may occur due to geolocation uncertainties and rasterization effects. To account for this, a tolerance-based IoU is computed by dilating the ground-truth mask with a spatial tolerance of two pixels before computing the overlap.

This relaxed formulation allows minor boundary shifts to be considered correct detections, providing a more realistic evaluation of localization performance in remote sensing segmentation tasks. However, excessive tolerance may partially mask fine boundary inaccuracies.

Hausdorff Distance 95% (HD95) The Hausdorff Distance measures the maximum deviation between two sets of points. Given a ground-truth contour G and a predicted contour P, it is defined as:

$$HD(G, P) = \max\left(\sup_{g\in G}\inf_{p\in P}\|g - p\|, \sup_{p\in P}\inf_{g\in G}\|p - g\|\right) \tag{7}$$

Due to its sensitivity to outliers, the 95th percentile version is commonly used:

$$HD_{95} = percentile_{95}(D(G, P) \cup D(P, G)) \tag{8}$$

HD95 measures the geometric discrepancy between contours and is particularly sensitive to boundary errors, making it complementary to region-based metrics such as Dice. It is widely used in medical image segmentation where boundary accuracy is critical.

Discussion While region based metrics such as Dice and Recall evaluate pixels agreement, they fail to capture boundary precision. Conversely, HD95 focuses on geometric accuracy but ignores region consistency. For this reason, both overlap and boundary based metrics are used to obtain a complete assessment of the performance of the model.

Threshold ablation Model outputs are sigmoid probabilities; a decision threshold τ is needed to binarize predictions. We perform a validation sweep over $\tau \in [0.1, 0.9]$ (17 values, linspace) and pick the τ that maximizes mean Dice on the validation set. This chosen τ is then used on the test set for all reported metrics. Calibrating τ is crucial because small shifts in threshold can markedly change boundary aligned metrics and recall/precision balance, especially in segmentation problems with high class imbalance; using the τ optimized through parameters ablation on the validation set prevents arbitrary bias and yields more stable, comparable test metrics.

Performance on the datasets In the following table are shown in detail the perfomances for each of the proposed architectures at different patch sizes.

Model	Dice	Precision	Recall	BFScore	IoU_{tol2}	HD95	Threshold
UNet 128	0.19	0.17	0.46	0.40	0.35	41.37	0.90
UNet 256	0.19	0.20	0.44	0.39	0.33	58.25	0.85
AttUNet 128	0.17	0.13	**0.51**	0.37	0.34	37.53	0.90
AttUNet 256	0.17	0.17	0.48	0.33	0.30	55.41	0.85
UNet++ 128	0.29	0.32	0.45	0.58	0.52	**22.37**	0.90
UNet++ 256	**0.37**	**0.43**	0.47	**0.65**	**0.57**	26.38	0.90

Table 2: Architectures performances on 128×128 and 256×256 images

To help with interpretation Table 2 is plotted in a barplot of Fig.1 and Fig.2.

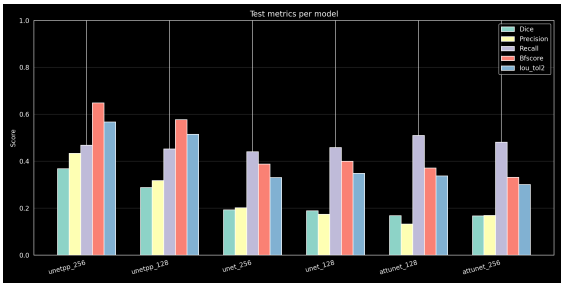

Figure 1: Barplot of the performances for the tested NN

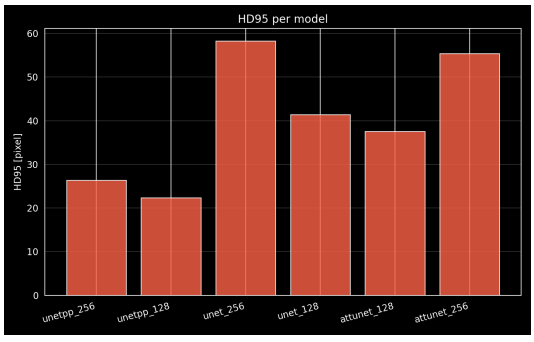

Figure 2: Barplot of the HD95 performances for the tested NN

The metrics clearly show how the most dense model, UNet++, is the best performing one for both overlap and boundary based metrics. We can also observe that Att128 has the best recall but, by observing its precision, it is possible to deduce that it is due to over segmentaiton, not good performances. The vanilla U-Net performs worse on both the 128×128 and 256×256 images showing the impact of attention gates and more advanced connections like in UNet++ on fine structures that may have long distance relationship in the images and on heavily imbalanced classes discrimination.

Results visualization In the following figures are shown two examples of how the different networks work at the two different image sizes.

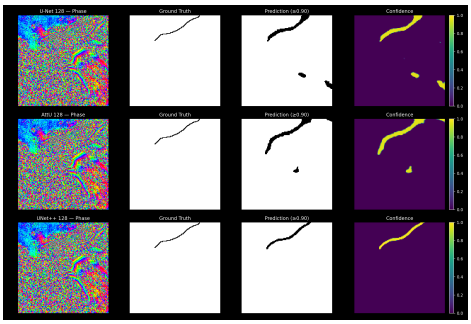

Figure 3: NNs trained on 128×128 images

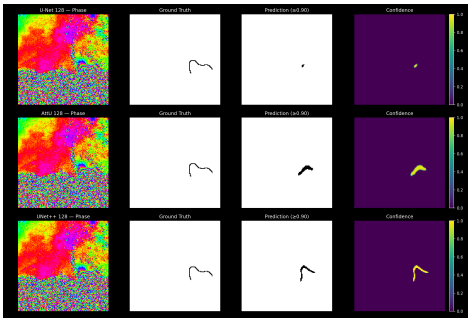

Figure 4: NNs trained on 128×128 images

From figures 3 and 4 it can be ovbserved how the standard U-Net architecture tends to have allucinations on which the grounding line is and also struggles with smaller grounding line that requires greater sensitivity to smaller patterns. The U-Net++ shows a clear improvement both in the shape and allucinations aspects.

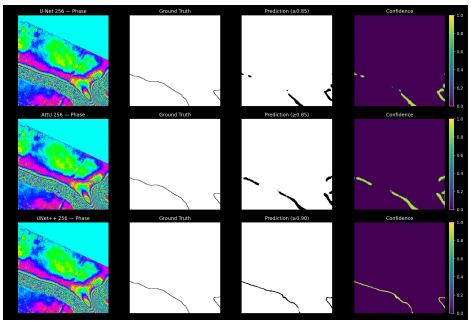

Figure 5: NNs trained on 256 × 256 images

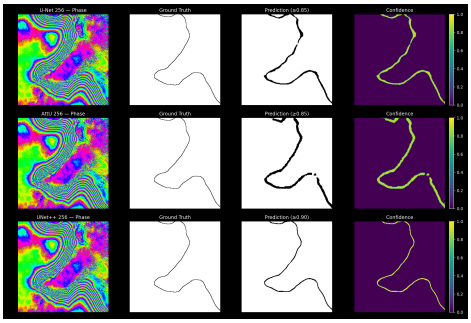

Figure 6: NNs trained on 256 × 256 images

Figures 5 and 6 show the same differences that can be observed between the networks on the 128 × 128 images but with an overall slight improvement.

Training and validation loss analysis In order to evaluate more fairly the performances of the proposed neural networks, it is necessary to analyze their training and validation loss trends, which represent a reliable indicator of possible overfitting. Figure 7 reports the evolution of the training and validation loss for all architectures and input resolutions considered.

From the figure, all models show a stable decrease of the training loss, indicating that optimization proceeds correctly. However, relevant differences emerge when comparing the gap between training and validation curves, which provides insight into generalization behavior.

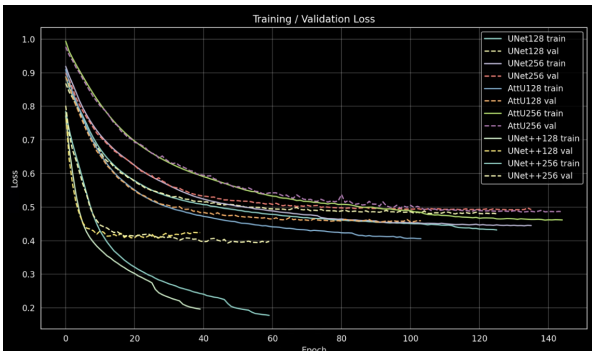

Figure 7: Training and validation loss for all architectures and input resolutions considered

The UNet++ architectures exhibit the most pronounced separation between training and validation loss. While the training loss decreases rapidly and reaches the lowest values among all models, the validation loss remains significantly higher and plateaus earlier. This widening gap is a clear indicator of overfitting, suggesting that the increased model capacity and dense skip connections of UNet++ favor memorization of the training data when not sufficiently regularized. This may be a result of the more complex architecture based on a higher number of parameters.

The Attention U-Net models show a more balanced behavior. Although a gap between training and validation curves is still present, it is less pronounced than in UNet++, indicating improved generalization. The attention mechanism appears to mitigate overfitting by focusing the learning process on relevant spatial features rather than purely increasing representational capacity.

The vanilla U-Net models display the smallest gap between training and validation losses. While their absolute performance is inferior, the closer alignment of the curves suggests better regularization and less tendency to overfit, albeit at the cost of reduced expressive power.

Overall, the loss analysis highlights a clear trade-off between model complexity and generalization:

- UNet++ achieves the best training performance but shows the strongest signs of overfitting;
- Attention U-Net provides a good compromise between expressiveness and generalization;
- vanilla U-Net generalizes more conservatively but underperforms in absolute accuracy.

These observations are consistent with the quantitative metrics reported in Table 2 and explain why UNet++ achieves superior test performance despite exhibiting stronger overfitting behavior during training. The overfitting may be a result of a poor choice of patches for the creation of the dataset given the relatively small amount of grounding lines.

4 Future work

Given the literature and the work done for this project it is possible to highlight the most urging issues to tackle in order to improve the generalization performances for the grounding line detection task.

- Use a better and more diverse dataset:
  - Use the Ross et al. dataset in the X-band that showed to lead to better performances;
  - Use the Anderson et al. which is in the C-band but has an highers resolution of 50×50 meters per pixel;
  - Re-design the dataset creation algorithm in order to reduce the repetition of information and do data augmentation in order to mitigate the effect of accepting less overlaps.
  - Use detections from different areas of the Antarctic. As highlighted in the discussed paper the NNs showed better performances on the region of training in contrast to others.
- Test for coherence relevance, researchers suggest that most of the information relevant for the grounding line detection is given from the phase component.
- Label smoothing, this new approach may contrast the over-estimation of the grounding line caused by the necessity of adopting the `all_touched` approach during the rasterization process.
- Use transformers architectures since they thrive with long term relationships such as the ones presented in the grounding lines the images.

5 Bibliography

References

[1] E. Rignot, "Grounding zone dynamics and ice sheet stability," *Nature Geoscience*, vol. 9, pp. 447–451, 2016.

[2] Y. Mohajerani, S. Jeong, B. Scheuchl, I. Velicogna, E. Rignot, and P. Milillo, "Automatic delineation of glacier grounding lines in differential interferometric synthetic aperture radar data using deep learning," *The Cryosphere*, vol. 15, pp. 471–485, 2021.

[3] N. Ross, P. Milillo, and L. Dini, "Automated grounding line delineation using deep learning and phase gradient-based approaches on cosmo-skymed dinsar data," *Remote Sensing of Environment*, vol. 262, p. 112 497, 2021.

[4] N. Abraham and N. M. Khan, "A novel focal tversky loss function with improved attention u-net for lesion segmentation," *IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 683–687, 2019.