The manuscript presents a new dataset of multi-mission SAR images of Antarctic ice shelf fronts, complete with manual annotations of the calving front. Five deep learning baseline models were trained on this dataset and evaluated on a separate test set. The dataset represents a valuable resource for developing and benchmarking deep learning models for calving front extraction, and the reviewer appreciates the considerable effort the authors have invested in its creation. To further strengthen the work, several aspects would benefit from clarification and refinement.

General Comments

First, it would be helpful to more clearly define the intended objective of a model on this dataset. For example, is an MDE of 0 m considered the ideal target, or is there a more realistic performance benchmark? Providing this context would improve the interpretability of the results. In addition, showing model performance on entire scenes (rather than patches) would give a more complete picture of real-world applicability.

Line 308: Did you perform hyperparameter tuning for the learning rate? Please clarify. If not, please consider doing so, as this can strongly influence model performance and may bias the comparison. The learning rate interacts closely with the architecture, optimizer, and parameter initialization, and different models (e.g., CNNs vs. Transformers, or pretrained vs. randomly initialized networks) typically require different learning rate regimes to converge properly. A learning rate that is too high may prevent convergence or lead to unstable training for some architectures, while a rate that is too low may result in underfitting or unnecessarily slow convergence. As a result, using a single fixed learning rate across all models can disadvantage certain architectures and lead to misleading performance differences that are not attributable to the model design itself but rather to suboptimal training settings.

Line 310: „We selected 150 epochs to ensure stable fine-tuning of model heads“ - Training length does not imply stability. Do you mean that all models converged within 150 epochs? Furthermore, the statement about fine-tuning only the heads requires clarification. In standard practice, only foundation models are kept frozen with head-only fine-tuning because their internal representations are general-purpose. Other pretrained models, however, are not designed to perform well on entirely new downstream tasks without updating the full network. If in your experiments only the heads of these non-foundation models were fine-tuned, the reported results may not reflect their true potential. To obtain a fair and meaningful comparison, these models should be fully fine-tuned (updating all layers) except for the foundation model, which can remain frozen except for the head.

Line 441–442: „Additionally, the buffer used to clip the scene around the current coastline lies close to the image border and follows the calving of iceberg A-68, which may have further influenced model performance.“ – Please clarify how areas outside the buffer were handled by the models (e.g., treated as white or black). Additionally, why was clipping done in a way that some patches contain very little actual content? Would it be possible to use full patches when they lie partially on the buffer edge, rather than excluding areas outside the buffer?

Line 521–522: „Even though, this did not result in lower prediction accuracies for specific regions because they weren’t represented within the training set.“ – I strongly recommend performing a quantitative analysis separating test scenes that are spatially covered by the training set from those that are not, in order to rigorously assess and demonstrate the model’s generalization performance.

Specific Comments

Figure 1 and Related Text:

- Several relevant works on Antarctic calving front delineation are currently missing and should be incorporated to provide a more comprehensive overview:
    - https://ieeexplore.ieee.org/document/11296938
    - https://www.mdpi.com/2072-4292/15/21/5168
    - https://tc.copernicus.org/articles/17/3485/2023/
    - https://tc.copernicus.org/preprints/tc-2023-52/
    - https://essd.copernicus.org/articles/14/4287/2022/ (baselines)
    - https://arxiv.org/abs/2501.05281 (foundation model)
    - 10.1109/IGARSS52108.2023.10281828
    - https://ieeexplore.ieee.org/document/10903382
    - https://arxiv.org/abs/2512.11560
    - https://arxiv.org/abs/2601.21663
- Some references appear to be mislabeled:
    - Zhang et al. (2021) is labeled as AutoTerm (likely Zhang et al. 2023).
    - Loebel (2021) should be Loebel (2022).
    - "Transformer" should be extended to include Transformer–CNN hybrids.
    - Gourmelon et al. (2025) should be specified as Gourmelon et al. (2025, TYRION).
- The definition of "edge-based methods" would benefit from clarification. It is currently unclear whether this includes methods trained on binary calving fronts (or additionally on binary calving fronts) or if a different definition is intended.

Line 57: „future SAR satellite mission" - This claim is not fully supported, as the training set appears to include the same satellites as the test set. Demonstrating generalization to a new mission would require at least one unseen satellite in the test set.

Table 1: coverage of Gourmelon et al. 2022 seems to be incorrect. Moreover, please define "benchmark dataset" more clearly (e.g., requiring preprocessed imagery and manual labels), otherwise additional datasets may need to be included.

Line 123: „in polar regions" - Figure 1 specifies Antarctic; this should be made consistent.

Line 158–160: State 353 m instead of 221 m, or clarify that this refers to the multi-annotator study of CaFFe. Additionally, it is 238 m on the CaFFe benchmark with additional prior knowledge and 75 m for the comparison to the multi-annotator study; please ensure this is clearly explained. „Transformer architecture" → this appears to be a hybrid Transformer–CNN architecture.

Line 168–170: „The test set was defined after expert evaluation of the images. It includes a representative selection of typical ice-shelf boundaries with varying levels of complexity to ensure comprehensive coverage of the problem space acquired by different sensors." - These sentences would fit better after the description of the training and validation sets.

Line 171–172: „the test set is based on 38 scene subsets especially focusing on the front" - How much calving front is present in the training and validation sets if they are not front-focused? Please clarify.

Line 241: „multi-expert annotation protocol" - This is a strong aspect of the dataset. However, please clarify for the reader that annotations are not independent and therefore not suitable for significance testing or uncertainty estimation.

Line 274–275: „Atrous Dilated Convolutions which enlarge the convolutions“ - Consider rephrasing to “which increase the receptive field of the convolutional filters.”

Line 283: Please add Gourmelon et al. 2025a, as this work evaluates the foundation model SAM.

Line 285: „input each Shelf-Bench SAR image into the RGB channels“ - Please clarify this sentence, particularly for readers unfamiliar with modality mismatches.

Table 3: Why is the inference time for DINOv3 higher than for ViT? A brief explanation would be helpful.

Line 301: „clip method“ - Please provide the exact percentile values used.

Line 319: „where the authors found this combination of loss functions improved performance“ - In this paper, focal loss does not appear to have been used; please verify.

Line 320: „class boundaries“ → class areas.

Line 320: „useful for learning the often complex and irregular shape“ - To the best of my knowledge, the standard Dice loss is not particularly suited for this, so please clarify or provide evidence for this claim, and indicate whether any modifications or weighting were applied to better handle such structures.

Line 321: „addresses class imbalance“ - Dice loss also already addresses class imbalance; please clarify the distinction.

Heading 4.3: consider renaming to „Baseline Results and Discussion“.

Table 4: „balances false positive and false negatives“ - More precisely, balances precision and recall (false positive rate and false negative rate).

Line 346: „Only DINOv3“ - Figure 4 appears to show similar behavior for ViT, U-Net, and DeepLabV3; please check consistency.

Line 347: „U-Net delineated the front closest to the ground truth“ - Figure 4 suggests FPN may perform similarly or better; please clarify whether this refers to only Figure 4 or to the entire test set results.

Line 350: „sea ice and fast ice challenges all models leading to slight discrepancies“ - Please elaborate on this interpretation. Would models not classify fast ice as part of the shelf?

Line 355: „reproducible segmentation performance“ - This would typically require multiple training runs of the same model; please clarify.

Line 359: „boundary ambiguity“ - As the boundary is defined between ice and ocean, ambiguity applies to both classes; consider clarifying.

Line 363: „the best performing model on ice recall (ViT) finds most of the relevant ice instances.“ – The differences in recall between the models appear relatively small. Therefore, this statement may be somewhat overstated, as “all relevant ice instances” would correspond to a recall of 1.0. Consider rephrasing this more cautiously to reflect the relatively close performance.

Line 365–366: „The lower precision for ice indicates that the Shelf-Bench dataset includes a sufficient number of difficult ‘Ice’ samples to challenge model performance.“ - Relative differences in precision between classes can not support this conclusion.

Line 382–384: „Satellite scene boundary artifacts were removed…“ - Please provide more detail on these processing steps.

Line 392: „The ViT delineates the coastline to be further inland than the ground truth and other models, highlighted by the higher MDE (Figure 5a).“ - Figures 5b, 5c, and 5e appear to show a different trend; please check consistency.

Line 396: „produce more stable front delineations“ - Please clarify whether this refers to specific panels (5c, 5e) or a general observation.

Figure 5: The figure resolution appears somewhat low. Is this due to journal formatting, or could a higher-resolution version be provided to allow zooming in on details? Additionally, the SAR images look more grainy or lower in resolution compared to Figure 4; please clarify the reason for this. The use of different colors for each model may be unnecessary, as the column already identifies the model. Using a single, easily visible color for all models (e.g., avoiding yellow against bright ice) could improve clarity.

Line 408: „a more robust performance indicator“ - Please clarify what this refers to.

Line 450: „median“ - For the most challenging cases, the mean MDE is reported; please ensure consistency.

Line 459–460: „The baseline models behave differently on Shelf-Bench, indicating that the dataset contains complementary challenges that no single architecture captures equally well“ - Please also acknowledge that the stochastic nature of neural network training has an influence on the baselines performance as well.

Line 461: „the dataset spans multiple spatial scales“ - This is somewhat vague; consider specifying that objects of interest occur at different scales.

Line 461–462: „… it also includes sufficiently clear fronts to evaluate pixel-level delineation accuracy.“ - This statement is unclear and would benefit from clarification.

Line 467–469: „The lower performance… suggests…“ - Please elaborate on how this conclusion is supported by the results.

Line 499–500: : „Increasing the tile size and introducing overlap tiling with centred predictions would provide broader spatial context and help reduce such ambiguities.“ - Please cite prior work (e.g., on the CaFFe dataset) that has explored this.

Line 515: „most realistic“ - Given that the median MDE is lower than the mean, easier scenes may be more prevalent; consider revising.

Line 525: „temporal“ → spatial?

Line 526: „temporal repetition“ - Please clarify (e.g., within the training set?).

Line 545: „targeted synthetic label generation“ - Please explain what is meant by this.

Line 574: „integrating temporal information“ - Please consider citing Dreier et al. 2025 (https://arxiv.org/abs/2512.11560), which explores this for the CaFFe dataset.

Data availability section: Consider moving descriptive elements to the dataset section and keeping only the access link here.

Line 598: „locally z-score normalized“ - Please clarify what “locally” refers to and which statistics were used for normalization.

Table A3: Why are there missing entries? Please clarify.

Technical Comments

Table 2: introduce the abbreviations IMP, GRD, AMI, ASAR, EW. Please also use “multi-looked” consistently (with or without hyphen). What does the number after “/” in the repeat cycle indicate?

Line 205: introduce the abbreviation IM.

Line 410: Missing closing parenthesis.