# Adapting Vision GNN for Patch-level Classification of Antarctic SAR Iceberg Imagery

Olivia Patterson, Faith Kalendek, Joe Gung, Thanh Nam Tran, Sree Nitya Kollu
University of Maryland, Baltimore County
1000 Hilltop Circle, Baltimore, Maryland 21250
opatter1@umbc.edu, faithk3@umbc.edu, jgung1@umbc.edu, ttran19@umbc.edu, skollu1@umbc.edu

## Abstract

*Increasing our understanding of iceberg and glacier processes such as calving, drifting, fragmentation, and melting, is essential to climate science, improved climate modeling, and iceberg dynamics. These processes provide indicators of global temperature shifts, sea level rise, and ecosystem changes. Synthetic Aperture Radar (SAR) has become a widely used tool for monitoring these phenomena due to its ability to capture high-resolution images regardless of weather or lighting conditions. Recent advances in deep learning have enabled automated analysis of SAR imagery, with Convolutional Neural Networks (CNNs) being commonly used for tasks such as detection and classification. However, CNNs are inherently limited by their grid-based structure, which may not fully capture the irregular spatial relationships present in SAR imagery. Vision Graph Neural Networks (ViG) provide an alternative approach by representing images as graphs of interconnected patches, allowing more flexible modeling of spatial dependencies. In this work, we investigate whether ViG can effectively perform patch-level classification on Antarctic SAR imagery and create a foundation for future works in understanding iceberg dynamics.*

*Link to code: https://github.com/olipat/SARViG*

## 1. Introduction

Improving our knowledge of iceberg and glacier processes is essential for enhancing climate modeling and cryosphere monitoring. Processes such as calving, drifting, fragmentation and melting all provide important information and indicators of global temperature change, sea level rise, freshwater management, and other environmental transformations like erosion [5]. Monitoring these processes and understanding their dynamics is pivotal in painting an overall picture for environmental trends.

Developed in the early 1950s, synthetic aperture radar (SAR) has become a primary remote sensing tool for iceberg and glacier monitoring, with its success attributed to its ability to capture high resolution imagery independent of weather and lighting conditions [1]. As increasingly large volumes of SAR data are collected, machine learning approaches have been explored for automated analysis tasks such as detection, segmentation, classification, and tracking [6]. Convolutional Neural Networks (CNNs) are a popular choice among researchers for these tasks. Khaleghian et al. (2021) clearly demonstrate that CNNs applied to SAR sea-ice classification can extract features from small patches and achieve accurate class separation [7]. However, in 2022, the Vision Graph Neural Network (ViG) was introduced as a graph-based alternative for visual representation learning. This new neural network models images as graphs of patch nodes and performs feature propagation through graph convolution and feed-forward modules, offering a flexible alternative to grid-based CNNs [4].

In this work, we evaluate the effectiveness of the ViG architecture for patch-level classification of Antarctic iceberg SAR imagery. We explore how transfer learning, self-supervised domain adaptation, and supervised finetuning on the ViG architecture impacts its classification performance. Our goal is to evaluate if the graph-based visual representation learning of the ViG architecture can provide a more flexible and meaningful understanding of SAR imagery through the patch-based spatial relationships. We also aim to establish a foundation for future work in understanding iceberg dynamics, particularly in monitoring and investigating the role of small ice fragments that are often overlooked.

## 2. Related Works

Computer vision work exists in many areas related to climate phenomena, such as wildfires, cyclones, and, in the context of our project, glaciers. It is paramount for transforming how we monitor, predict, and project climate im-

pacts [2]. Existing endeavors related to icebergs demonstrate an established methodology. As mentioned before, the use of neural networks on SAR iceberg imagery predates this project, and these works typically consisted of Convolutional Neural Networks (CNNs). Moreover, these efforts typically carry some industry implications. For instance, Zhan et al.'s team presented the problem of drifting icebergs, specifically regarding ship navigation and oil rigs. The research boasted the CNN's ability to apply Transfer Learning to distinguish between ships and icebergs when only given limited training data and features [10]. Other research continues this trend of utilizing CNNs in the context of industry. In the case of Wells et al.'s team, they focused on the logistical effects of these geographical events: "Supply chain disruptors such as piracy and navigational obstacles like icebergs, pose a probable risk to economic development and national security." The team's model achieved a near 90% accuracy on their prescribed dataset [9].

Evidently, CNNs have ingrained themselves as a staple model for evaluating glacial activities. The repetition in practice bolsters this notion. However, despite its establishment in the field of glacial activity and computer vision, there may be room for improvement. CNNs may struggle to fully capture the irregular spatial structures found in SAR iceberg imagery due to the grid-based representation. Therefore, alternative approaches should be considered and evaluated. For our project, we investigate whether the Vision Graph Neural Network (ViG) can serve as an effective alternative to traditional CNN-based approaches. Our team's goal is to offer a neural network alternative while shifting the focus from the previously mentioned industry-related concerns to the icebergs themselves.

# 3. Methods

This section describes the dataset, preprocessing pipeline, ViG model architecture, training strategies, experimental setup, and evaluation metrics used to evaluate patch-level classification of Antarctic iceberg SAR imagery.

## 3.1. Dataset

We use SAR imagery of Antarctic icebergs collected from Sentinel-1 through the European Space Agency's Copernicus Browser [3]. The dataset consists of 90 SAR images capturing a variety of complex polar scenes containing icebergs, ocean regions, fragmentation patterns, and other iceberg structures.

Because the collected SAR dataset is unlabeled, we adopt a hybrid learning pipeline combining self-supervised learning and supervised finetuning. During the self-supervised stage, 40 unlabeled SAR images are used in a reconstruction-based representation learning task to adapt the model to the SAR domain. The remaining images are manually annotated for supervised patch-level classification. Of the labeled subset, 25 images are used for training, 7 for validation, and 18 for evaluation. Ground truth annotations are generated using Label Studio, as seen in Figure 1, where regions are manually segmented and exported as mask labels.

Each SAR image is divided into non-overlapping patches that serve as the fundamental units for classification. We evaluate both binary and multiclass classification settings. In the 2-class setting, patches are categorized as either ice or ocean. In the 4-class setting, patches are categorized as ocean, interior ice, boundary regions, or small ice fragments. Additionally, no-data regions are annotated to ensure they are excluded from training metrics and evaluation scores, rather than treated as a predicted class.

## 3.2. Model Architecture

In computer vision, CNNs have become the de facto standard network architecture for image analysis tasks, showcased by the previous works discussed in this paper. However, widespread use of an approach doesn't necessarily imply it is the best choice for all image domains. CNNs process images through grid-based convolution operations that are efficient at capturing local spatial features. Because CNNs operate on fixed spatial neighborhoods, they can struggle to capture the relationships of irregular shapes or fragmented structures, which are common in SAR imagery.

ViG provides a graph-based alternative for visual representation learning. Rather than modeling images through grid-based pixel relationships, ViG represents images as graphs made of interconnected image patches. As stated by Han et al [4], the popular CNN architecture treats the image as a grid sequence structure, which can limit their ability to flexibly model irregular and complex objects. ViG constructs a k-nearest neighbor (KNN) graph between image patches, which allows feature propagation between spatially related regions with graph convolutions operations. This graph-based alternative can provide a more versatile alternative than its rigid counterpart for capturing complex spatial patterns in SAR imagery. Previous work has shown that ViG can provide understandable representations and give meaning to images with its neighborhood structure through its k-nearest neighbor (KNN) graph [8].

ViG was originally designed for image-level visual recognition tasks using natural image datasets such as ImageNet [4], where the model predicts a single label for an entire image. In our work, we adapt the architecture for patch-level SAR classification by dividing SAR images into non-overlapping patches and assigning labels to each patch rather than the image as a whole. Each patch acts as a node within the graph structure, allowing the model to learn relationships between neighboring SAR regions while preserving local spatial context. We further adapt the ImageNet-

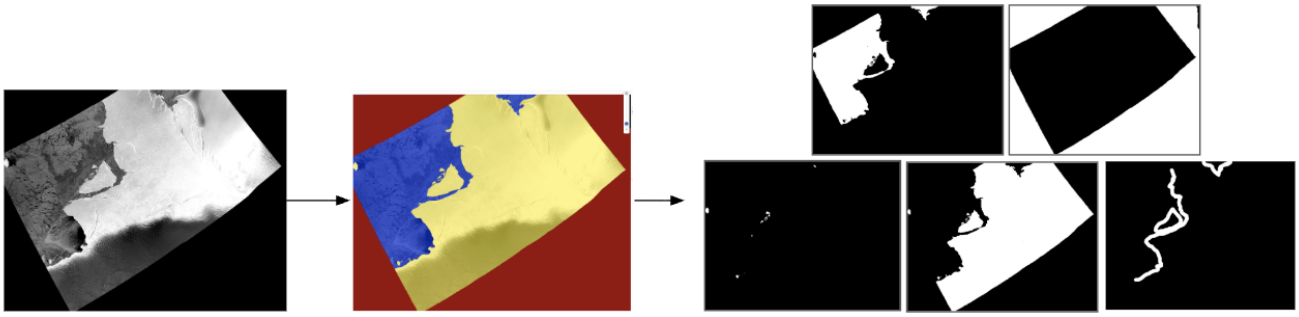

Figure 1. Annotation pipeline starting with the processed SAR imagery from the Copernicus browser, annotating in Label Studio, and the output masks.

pretrained ViG model to the SAR domain through a self-supervised reconstruction stage to help the model learn SAR specific representations before performing the downstream classification tasks.

The ability to capture complexities and bring meaning to iceberg imagery is integral for properly identifying components (ice, ocean, etc.) in SAR iceberg imagery. Without meaningful spatial representations, important fragmented structures within SAR imagery may become difficult to interpret and analyze.

### 3.3. Training Strategies

This work evaluates two training strategies for adapting the ViG architecture for patch-level classification: a direct supervised baseline and a self-supervised pretraining approach. Both strategies operate on the same underlying model architecture and patch-based formulation described in the previous sections, allowing controlled comparison of training methodology rather than structural changes.

The **baseline** strategy consists of directly fine-tuning an ImageNet-pretrained ViG model on labeled SAR patch data. The model is trained end-to-end using cross-entropy loss over patch-level predictions, with no freezing of layers. In this setting, the model is expected to learn SAR-specific representations solely through supervised signals. This approach serves as a direct transfer learning baseline from natural image pretraining to SAR the domain.

The self-supervised strategy introduces an intermediate pretraining stage using unlabeled SAR data. In this stage, the model is optimized using a reconstruction objective, where the network learns to reconstruct the input SAR image from its internal representations. The reconstruction loss is computed using pixel-wise mean squared error. This stage is intended to encourage the model to adapt its feature representations to SAR-specific texture and intensity distributions prior to supervised learning.

Following pretraining, the model is fine-tuned using the same supervised patch classification pipeline as the baseline. We call this strategy the **Deluxe**. No architectural changes are introduced between strategies during fine-tuning, ensuring that differences in performance arise from learned representations rather than model capacity.

This design separates representation learning from task-specific learning. The self-supervised stage biases the model toward capturing low-level structural and texture information in SAR imagery, while the supervised stage focuses on class discrimination. However, a potential limitation of reconstruction-based pretraining is that it may prioritize pixel-level fidelity over semantic separability, which can negatively impact discrimination between visually similar classes such as ice and ocean.

In short, we have 2 strategies:

- **Baseline**: Direct Finetuning
- **Deluxe**: Self-Supervised by Reconstruction and Direct Finetuning

Across both strategies, the classification task is formulated as dense patch-level prediction, where each image is divided into a fixed spatial grid and each grid cell is assigned a class label. The model outputs a grid of logits corresponding to predefined classes, while regions labeled as no data are excluded from optimization using an ignore index.

### 3.4. Experimental Setup and Configurations

The experimental pipeline is designed to evaluate training strategies under controlled and comparable conditions. All experiments use the same ViG backbone architecture and identical patch-based formulation, with variations restricted to training strategy, hyperparameters, and initialization method.

Input SAR images are resized to 224×224 resolution prior to being processed by the model. To maintain compatibility with ImageNet-pretrained weights, single-channel greyscale SAR imagery is replicated across three input channels.

Ground truth supervision is derived from manually annotated pixel-level masks. These annotations are converted

into patch-level labels through a structured pipeline that aggregates pixel-level class information within each grid cell. Patch labels are assigned using majority voting, with additional rule-based handling for ambiguous regions such as boundary and small ice classes. Regions labeled as no data are excluded from training and evaluation via an ignore index.

The ViG model is adapted to support patch-level prediction through a grid-based classification head that outputs grid-based predictions. For the self-supervised setting, an additional reconstruction head is introduced during pretraining, which maps latent representations back to image space. This component is removed during supervised fine-tuning, ensuring consistency across experimental conditions.

Experiments were conducted using multiple grid resolutions and class configurations to evaluate the effect of spatial granularity and class complexity on SAR patch classification performance. Grid sizes of 56×56 and 112×112 were explored, corresponding to different patch resolutions over the input image. Both binary class and multiclass settings were evaluated, including simpler ice/ocean configurations as well as finer-grained classifications incorporating boundary and small ice regions.

We also compared multiple pretrained initialization and transfer learning strategies. Experiments were conducted using both ImageNet-pretrained weights and SAR-adapted weights generated through the self-supervised reconstruction stage. Additional experiments examined different fine-tuning configurations, including full-network training and selective layer freezing during transfer from ImageNet-pretrained weights.

Training was performed using configuration-controlled hyperparameter settings defined through YAML files. Models were optimized using standard PyTorch optimization and learning-rate scheduling pipelines over multiple training epochs. Experiments were executed on GPU hardware to support both self-supervised pretraining and supervised fine-tuning stages.

No data augmentation is applied in the final experimental setup in order to isolate the effect of training strategy without introducing additional stochastic variation.

### 3.5. Training and Evaluation Metrics

During training, we monitored both quantitative and qualitative metrics to evaluate convergence behavior and model learning dynamics.

Training loss values were recorded across epochs to observe convergence stability and identify potential overfitting behavior. In general, the Baseline strategy demonstrated more stable convergence and stronger downstream classification performance compared to the Deluxe strategy. For the supervised finetuning stage, cross-entropy loss was monitored for patch-level classification, while the self-supervised reconstruction stage used a pixel-wise Mean Squared Error (MSE) loss to evaluate reconstruction quality.

In addition to numerical loss monitoring, qualitative visualizations of predicted outputs were inspected during training. These visualizations allowed us to observe how patch-level predictions evolved over time and whether the model learned meaningful spatial structures such as ocean regions, ice regions, and boundaries. As seen in Figure 2, the predicted patch-level classifications greatly improved over 10 epochs of supervised finetuning.

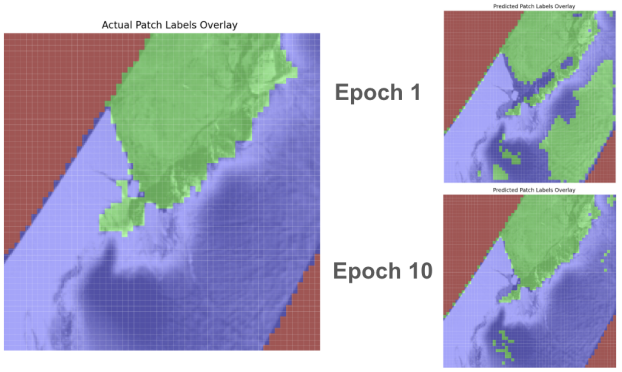


Figure 2. On the left is the ground truth labels. On the right are the predicted labels for epoch 1 and epoch 10 during Deluxe binary classification supervised finetuning.

## 4. Results and Analysis

After training, we evaluate both strategies on the patch-level SAR classification task on 2 different grid sizes. The evaluation process composed of analyzing quantitative data and visually inspecting the qualitative output images.

### 4.1. Evaluation Metrics

For final evaluation, we analyzed model performance using confusion matrices, accuracy, F1-score, and Intersection over Union (IoU). These metrics were selected to evaluate both overall classification performance and spatial prediction quality.

- **Confusion Matrices**: this metric provides class-level analysis by showing which classes are correctly classified and which classes are not. Correspondingly, this allows us to know exactly where the model makes correct predictions and where the model is getting confused.
- **Accuracy**: this metric measures the overall percentage of correctly classified patches
- **F1-score**: this metric calculated by precision and recall. This is particularly useful for imbalanced data distribution.

  - **Macro F1**: computes the F1-score independently for each class and then averages them equally. This metric will be particularly useful for 4-class analysis.
  - **Micro F1**: aggregates the total true positives, false positives, and false negatives across all classes.
- **Intersection over Union**: this metric measures the overlap between predicted class regions and the ground-truth regions making it useful for spatial boundary spots and segmentation task in general.

#### 4.1.1. Confusion Matrix

After recording our data, we create confusion matrices visualizations.

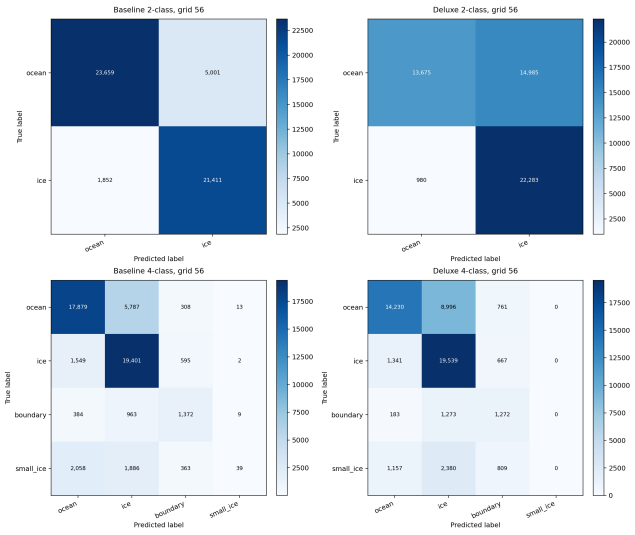


Figure 3. Confusion matrix for the 2-class

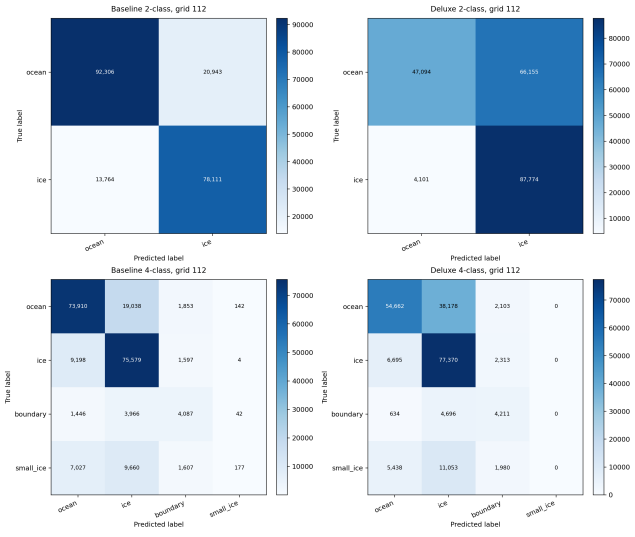


Figure 4. Confusion matrix for the 4-class

From the confusion matrix results, in both grid-size settings, the Baseline strategy is clearly stronger. Most predictions are concentrated along the diagonal, indicating correct classification. In contrast, the Deluxe strategy shows a significant number of errors in prediction.

- **2-class:** The Baseline model performs well for both ocean and ice classification. Meanwhile, the Deluxe model struggles as it misclassified a lot of ocean patches as ice.
- **4-class:** We added boundary and small ice fragments, both strategies struggle with these smaller classes. This indicates that the boundary and small ice fragment classification is much harder compared to ice and ocean patches.
- **Grid 56 vs 112:** There is no significant changes when comparing smaller vs larger grid size.

#### 4.1.2. Accurary, F1, Intersection over Union

For Accurary, F1, Intersection over Union, we report these data in both table form and bar plot visualization.

The accuracy, F1-score, and IoU results show a similar pattern. The Baseline model performs significantly better than the Deluxe model across the reported metrics. When comparing grid sizes, Grid 56 produces better predictions than Grid 112. Smaller patch-size results in more patches leading to more incorrect predictions.

A couple of interesting insights can also be seen from the data:

- **IoU**: IoU drops compared to Accuracy and F1. This suggests that although the models can often classify the general region correctly, they still struggle with spatial overlap and boundary precision.
- **Macro F-1 in 4-class**: Macro F1 drops sharply compared to accuracy and micro F1. This gap indicates that the models perform well on dominant classes such as ocean and ice, but struggle with minority or harder classes such as boundary and small_ice.

### 4.2. Qualitative results

Next, we examined the qualitative results by visually comparing the annotated ground-truth overlay with the predicted output generated by each strategy.

As shown in Figure 6, the Deluxe model makes visible mistakes when predicting ocean regions. This observation is consistent with the confusion matrix results.

As shown in Figure 7, both models show mixed predictions around boundary areas between classes. This observation is supported the IoU results, which indicate that boundary-level performance remains challenging even when the overall classification metrics are reasonable. Additionally, the Baseline strategy demonstrated limited prediction of small ice fragment regions, while the Deluxe strategy failed to predict any small ice fragment regions. This suggests the fine-grained fragmentation patterns are

Table 1. Accurary, F1, IoU Data

| Strategy | Grid | Classes | Accuracy | Macro F1 | Micro F1 | Mean IoU |
|---|---|---|---|---|---|---|
| Baseline | 56 | 2 | **86.53%** | **86.56%** | **86.58%** | **76.31%** |
| Baseline | 56 | 4 | 73.27% | 52.27% | 73.48% | 40.81% |
| Baseline | 112 | 2 | 83.15% | 82.88% | 82.96% | 70.78% |
| Baseline | 112 | 4 | 73.37% | 50.51% | 73.36% | 39.35% |
| Deluxe | 56 | 2 | 69.05% | 68.42% | 69.34% | 52.25% |
| Deluxe | 56 | 4 | 66.34% | 45.73% | 66.58% | 33.97% |
| Deluxe | 112 | 2 | 65.87% | 64.61% | 66.06% | 48.14% |
| Deluxe | 112 | 4 | 64.82% | 45.04% | 65.06% | 33.05% |

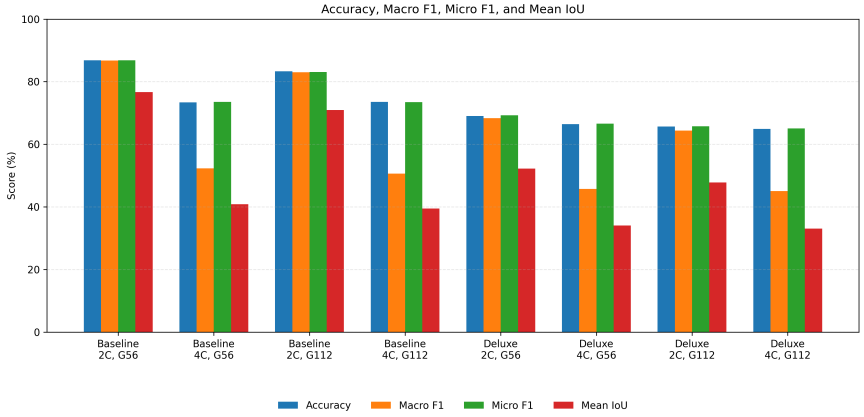


Figure 5. Accuracy, F1, and IoU visualization.

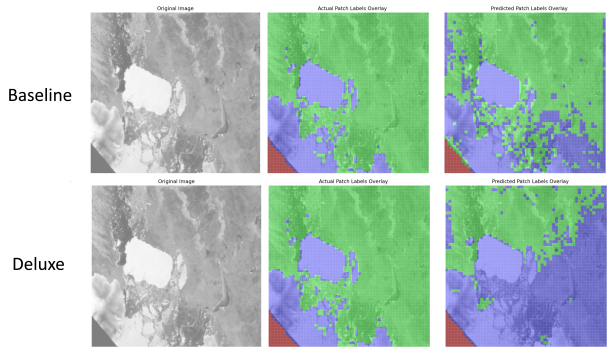


Figure 6. 2-Class Baseline Grid 56

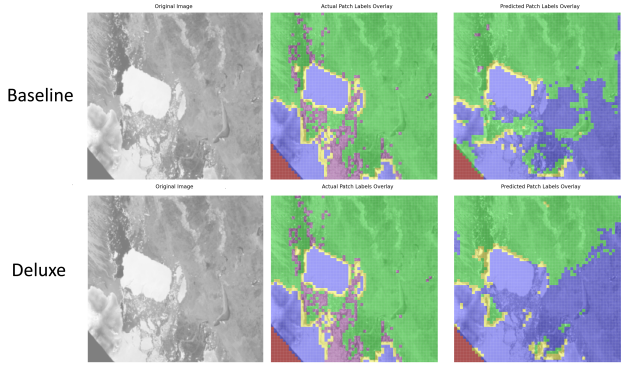


Figure 7. 4-Class Baseline Grid 56

difficult for to recognize with both training approaches, but particularly with the self-supervised reconstruction strategy added to the training.

### 4.3. Analysis

Overall, the quantitative and qualitative results show that the Baseline strategy outperforms the Deluxe strategy. Although the Deluxe pipeline was designed to improve node representations through a self-supervised reconstruction pretext task, it performed worse.

There are two hypothesis for this result:

- **Pretext task and downstream task misalignment:** In the deluxe pipeline, the downstream task is too different from the pretext task. Although reconstruction can help preserve SAR texture information, it may not teach the model the most useful features for separating ice, ocean,

boundary, and small ice fragment regions. As a result, the Deluxe model may learn representations that are useful for reconstruction but less useful for classification.

- **Limited dataset quality and diversity:** Another possible explanation is the limited size and diversity of the dataset. With more labeled and unlabeled SAR scenes, the Deluxe pipeline may have had a better chance to learn useful domain-specific representations.

The grid-size comparison can also be explained through the patch-level classification setup. Grid 56 performs better than Grid 112 because each patch contains more spatial context. In contrast, Grid 112 divides the image into smaller patches which amplifies classification mistakes.

## 5. Limitations and Future Work

Several limitations affected the performance of this work.

First, the diversity of the dataset was limited for both supervised and self-supervised learning. Although the supervised patch-level approach generated a large number of samples from 56x56 and 112x112 grid configurations, many patches originated from the same underlying SAR scenes. As a result, the model was exposed to limited environmental and structural diversity during training. This issue was particularity significant in the self-supervised reconstruction stage, where the unlabeled SAR dataset likely did not contain enough varied iceberg scenes, fragmentation patterns, and ocean conditions to learn generalized SAR representations.

Another limitation was the dataset distribution bias between training and evaluation splits. The evaluation dataset contained a larger concentration of small ice fragment regions compared to the training dataset, which likely contributed to weaker generalization on minority classes. Both Baseline and Deluxe strategies struggled to consistently identify small ice fragments and boundary regions, with the Deluxe strategy showing no small ice fragment predictions during evaluation. Additional examples of fragmented iceberg regions would likely improve downstream classification performance.

The presence of label noise introduced during manual annotation also presents a challenge in this work. SAR imagery often contains ambiguous transitions between ocean, ice, and fragmented regions, making precise segmentation difficult even for human annotators. As a result, some ground truth labels may contain inconsistencies or inaccuracies that negatively affect both optimization and evaluation. Furthermore, manual annotation of SAR imagery is highly time-consuming, limiting the speed at which larger datasets can be generated.

The complexity of SAR imagery itself also presents a challenge. Unlike natural images, SAR images contain noise, irregular spatial patterns, and fragmented structures that can vary significantly between scenes. In many cases, visually similar regions such as ocean surfaces and fragmented ice can overlap in texture and intensity, increasing classification difficulty.

Additionally, the reconstruction-based self-supervised objective used in the Deluxe strategy may not align well with the downstream classification task. While reconstruction encourages preservation of low-level SAR texture information, it may not learn features that are semantically useful for distinguishing between classes such as ice and ocean. This mismatch between the pretext task and downstream classification task likely contributed to the weaker performance of the Deluxe strategy. A larger and more diverse unlabeled SAR dataset, or alternative self-supervised objectives more closely aligned with patch-level classification could improve future performance.

Finally, computational and time constraints limited the number of experiments that could be conducted. Future work could explore larger datasets, alternative self-supervised objectives, additional augmentation techniques, overlapping patch strategies, stronger annotation guidelines, semi-automated labeling pipelines, and broader comparisons against CNN and transformer-based baselines.

## 6. Conclusion

Although the experimental results did not fully support our original hypothesis, they still provide useful direction for future work.

### 6.1. Architecture

Regarding the model architecture, we conclude that adapting the Vision Graph Neural Network (ViG) backbone for patch-level SAR iceberg image classification is viable. Although many challenges remain, our limited experiment achieved a strong baseline performance. Specifically, the Baseline model archived 86.5% accuracy in SAR iceberg imagery object classification task. This experiment provides a good foundation for future work to further adapt graph-based vision models in this domain.

### 6.2. Strategies

About the strategies, at this stage, we demonstrate that directly fine-tuning the ImageNet-pretrained ViG model on labeled SAR iceberg data results in better performance than the Deluxe strategy. Nevertheless, future work should explore new training pipelines or improved variants of the Deluxe strategy.

Overall, this work demonstrates that graph-based visual representation learning remains a promising direction for Antarctic SAR iceberg analysis. While many challenges remain in adapting self-supervised learning and improving boundary and small ice fragment recognition, the results establish a foundation for future research in graph-based

SAR understanding and iceberg structure analysis and dynamics.

## References


[1] John C. Curlander and Robert N. McDonough. *Synthetic Aperture Radar: Systems and Signal Processing*. Wiley, 1991. 1

[2] Daniel Pazmino Vernaza. Introduction to computer vision for climate change, 2025. 2

[3] Jedrzej Bojanowski Dennis Clarijs Jurry de la Mar Dávid D. Kovács, Jan Musial and András Zlinszky. Copernicus data space ecosystem establishes public cloud processing for earth observation data., 2026. accessed January 27, 2026. 2

[4] Kai Han, Yunhe Wang, Jianyuan Guo, Yehui Tang, and Enhua Wu. Vision gnn: An image is worth graph of nodes. *arXiv preprint arXiv:2206.00272*, 2022. 1, 2

[5] IPCC. Climate change 2021: The physical science basis, 2021. 1

[6] Kaggle. Statoil/c-core iceberg classifier challenge, 2018. 1

[7] Salman Khaleghian et al. Sea ice classification of sar imagery based on convolutional neural networks. *Remote Sensing*, 13(9), 2021. 1

[8] Olivia Patterson and Rebecca Williams. Graph-based modeling of iceberg dynamics from synthetic aperture radar imagery. *10.32473/flairs.39.1.141697*, 2026. 2

[9] Kevin Wells, Vasit Sagan, and Yusupujiang Aimaiti. Differentiating vessel and iceberg with cnn using sar imagery for arctic navigatability. *https://ieeexplore.ieee.org/document/10282581*, 2023. 2

[10] Cheng Zhan, Licheng Zhang, Zhenzhen Zhong, Sher Didi-Ooi, Youzuo Lin, Yunxi Zhang, Shujiao Huang, and Changchun Wang. Deep learning approach in automatic iceberg - ship detection with sar remote sensing data. *https://arxiv.org/abs/1812.07367*, 2018. 2