Feature Representation for Sea Ice Mapping

by

Javier Noa Turnes

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2026

© Javier Noa Turnes 2026

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Juha Karvonen Senior Researcher at Finnish Meteorological Institute
Supervisor(s):	David A. Clausi Professor, Dept. System Design Engineering, University of Waterloo
	Linlin Xu Assistant Professor, Dept. Geomatics Engineering, University of Calgary
Internal Member:	Andrea Scott Associate Professor, Dept. Mechanical and Mechatronics Engineering, University of Waterloo
	Rodrigo Costa Assistant Professor, Dept. System Design Engineering, University of Waterloo
Internal-External Member:	Christine Dow Associate Professor, Dept. Geography and Environmental Management, University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

The following journal articles are included in this thesis. In each case, I made major contributions to the study design, methodology development, experimental evaluation, and preparation of the manuscripts:

Javier Noa Turnes, Mingzhe Jiang, Saeid Taleghanidoozdoozan, Linlin Xu, and David A. Clausi. ITT: Long-Range Spatial Dependencies for Sea Ice Semantic Segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18:13296-13309. (2025, Published).

Javier Noa Turnes, Muhammed Patel, Fernando Pena Cantu, Kyle Gao, Linlin Xu, and David A. Clausi. Is Pre-training Enough? SAR-IceFM: Towards Multi-Task Foundation Models for Sea Ice Classification. *Remote Sensing of Environment Journal*. (2026, Submitted).

Abstract

Sea ice monitoring is essential for climate research, Arctic navigation, and operational decision-making. Synthetic aperture radar (SAR) imagery is the primary sensing modality used by national ice services because of its independence of atmospheric and lighting conditions, and sensitivity to surface structure. However, SAR-based sea ice classification remains challenging due to spatially non-stationary statistics caused by incidence angle effects, seasonal transitions, and strong within-class variability. These factors complicate feature extraction and limit the robustness and transferability of conventional deep learning models.

This thesis investigates feature representation learning for sea ice classification in SAR imagery through both supervised and self-supervised paradigms. The first contribution introduces a supervised semantic segmentation framework that integrates convolutional neural networks (CNNs), transformers, and unsupervised region segmentation. The proposed Irregular Tokens on Transformers (ITT) architecture forms multi-scale, homogeneous tokens using Iterative Region Growing on Semantics (IRGS) and applies self-attention to capture long-range spatial dependencies. A multi-task training scheme combines pixel-level and region-level loss functions, encouraging region-consistent feature representations while preserving fine-grained boundaries. Experiments on multi-season RADARSAT-2 scenes demonstrate improved overall accuracy, sharper boundary delineation, and reduced predictive uncertainty compared to a CNN baseline. An expert audit conducted by the Canadian Ice Service further supports the operational relevance and stability of the approach across freeze-up and melt conditions.

While supervised learning delivers strong performance when annotations are available, SAR labeling remains costly and domain specific. The second contribution explores self-supervised pre-training toward a SAR foundation model for sea ice classification. By leveraging masked representation learning and multi-task objectives, the proposed framework learns transferable representations from unlabeled SAR imagery. The study evaluates whether large-scale pre-training alone is sufficient to address domain shifts across sensors and seasons, or whether task-specific adaptations remain necessary. Results show that self-supervised pre-training substantially improves downstream performance and generalization, but optimal accuracy is achieved when combined with structured fine-tuning aligned with sea ice semantics.

Overall, this thesis demonstrates that robust sea ice classification fundamentally depends on how feature representations are learned, and provides principled strategies for improving scalability, generalization, and operational viability in Arctic SAR applications.

Acknowledgements

I am deeply grateful to my advisor, Dr. David A. Clausi, for the invaluable opportunity to pursue my PhD at the University of Waterloo and for the mentorship that profoundly shaped both my academic trajectory and personal growth.

I am equally grateful to my co-advisor, Dr. Linlin Xu, for his invaluable support, insightful discussions, and continuous encouragement throughout these years.

It has been a true privilege to share this journey with my lab mates. I am especially thankful to those who welcomed me into the VIP Lab: Mingzhe Jiang, Neil Brubacher, Saeid Taleghanidoozdoozan, Yuan Fang, and Xinwei Chen, as well as those who joined afterward: Fernando Peña, Muhammed Patel, Lily de Loe, Jayden Hsiao, Niloofar Azad, Hudson Sun, Junfeng Lei, Kyle Gao, and Auguste Koh, among others. The professional collaboration and genuine friendships built during this time have been invaluable.

I also gratefully acknowledge the Systems Design Engineering Department for its well-structured administrative support and assistance throughout my studies.

I further wish to acknowledge the Canadian Ice Service for its valuable feedback and its willingness to engage with and support the long-term vision of this research. I also gratefully acknowledge the support of the Canadian Space Agency through the ROSS (Research Opportunities in Space Science) program.

I am deeply indebted to my parents and brothers. Any achievement of mine is a direct consequence of their guidance, sacrifice, and unwavering support.

Finally, my especial gratitude goes to the family I have built: my wife, Daliana, for her love, strength, and confidence in me, and my son, Oliver, who gives meaning to every accomplishment.

Dedication

To my son, Oliver,
born at the dawn of this chapter of my life.

Table of Contents

Examining Committee Membership ii

Author's Declaration iii

Statement of Contributions iv

Abstract v

Acknowledgements vi

Dedication vii

List of Figures xi

List of Tables xiii

1 Introduction 1

1.1 Motivation 3

1.2 Objectives 4

1.3 Thesis Structure 5

2 ITT: Long-range spatial dependencies for sea ice semantic segmentation 6

2.1 Introduction 7

2.2 Background . . . 9
2.3 Non-stationary statistics on SAR imagery . . . 11
2.4 Self-attention on irregular tokens for semantic segmentation . . . 12
2.4.1 Self-attention complementing convolutions . . . 12
2.4.2 Unsupervised segmentation: IRGS . . . 14
2.4.3 ITT: Irregular Tokens on Transformers . . . 14
2.5 Addressing non-stationary statistics . . . 17
2.6 Experiments . . . 18
2.6.1 Dataset . . . 18
2.6.2 Experimental setup . . . 18
2.7 Results and discussion . . . 21
2.7.1 Predictions on boundaries . . . 25
2.7.2 Uncertainty analysis . . . 27
2.7.3 Region-based semantic segmentation . . . 30
2.8 Audit conducted by the CIS . . . 34
2.9 Conclusion . . . 38

3 Is Pre-training Enough? SAR-IceFM: Towards Multi-Task Foundation Models for Sea Ice Classification 40

3.1 Introduction . . . 41
3.2 AI4Arctic Dataset . . . 44
3.3 MAE for sea ice . . . 46
3.4 Downstream training . . . 49
3.5 Experiments . . . 50
3.5.1 Architecture . . . 51
3.6 Results . . . 52
3.6.1 Overview of multi-task semantic segmentation of sea ice . . . 52
3.6.2 Foundation model transfer across data regimes . . . 54

3.6.3 Multi- vs. single-task training . . . 58
3.6.4 Multi- vs. single-scale training: resolution transferability . . . 59
3.6.5 Mask ratio ablation . . . 63
3.6.6 Comparison with other architectures . . . 64
3.7 Discussion . . . 66
3.8 Conclusion . . . 68

4 Conclusions and Future Directions **70**
4.1 Summary of Contributions and Findings . . . 70
4.2 Future work . . . 72

References **75**

APPENDICES **91**

A ArcticScope **92**

B Visualizations **95**

List of Figures

2.1 Non-stationary behavior on SAR scenes for sea ice mapping 11
2.2 ITT framework 15
2.3 Study area of the dataset 19
2.4 Self-attention block. 20
2.5 Semantic segmentation for scene 20101003 24
2.6 Semantic segmentation for scenes 20101017 (top) and 20100623 (bottom) . 25
2.7 Metrics around boundaries for different buffer thicknesses 26
2.8 Entropy calculated from the predicted probabilities 28
2.9 Predicted probabilities for class ice 29
2.10 Pixel- and region-based semantic segmentation comparison 32
2.11 Overall accuracy (OA) and boundary-buffer OA for hybrid segmentation . 33
2.12 CIS expert audit of operational capability 35
2.13 Seasonal error/frequency heatmap 36

3.1 Study area of the AI4Arctic dataset 45
3.2 MAE pre-training and multi-task fine-tuning 47
3.3 Semantic segmentation results using the ViT-h backbone 53
3.4 Combined score as a function of test-time resolution 60
3.5 Semantic segmentation results at different test-time resolutions 61
3.6 Difference metrics (%) of fine-tuned models 63
3.7 Semantic segmentation results using the ViT-h backbone 67

A.1 Human-in-the-loop active learning workflow 93
A.2 ArcticScope 94

B.1 Qualitative comparison of sea ice predictions across multiple architectures . 96

List of Tables

2.1 U-Net architecture 20
2.2 Overall accuracy (%) for each experimental protocol 22
2.3 Overall accuracy for pixel-based and region-based semantic segmentation . 31

3.1 Training settings of the proposed foundation model 51
3.2 Training times for different backbones 52
3.3 Performance gains from MAE pre-training compared to training from scratch 56
3.4 Performance of multi-task and single-task downstream models 58
3.5 Comparison of combined scores across architectures 65

Chapter 1

Introduction

The rapid expansion of satellite missions over the past two decades has transformed Arctic observation into a data-intensive discipline. SAR Constellations such as RADARSAT-2, Sentinel-1, TerraSAR-X, and the RADARSAT Constellation Mission (RCM) have increased spatial coverage, temporal revisit frequency, and radiometric stability for polar regions [1]. At the same time, climate variability, maritime route accessibility, offshore resource exploration, and ecosystem monitoring have intensified the demand for reliable sea ice information. Initiatives such as the AutoIce Challenge[1] further highlight the growing research interest in automated sea ice analysis, encouraging scalable approaches [2–4] for extracting concentration, stage of development, and floe size descriptors from satellite data.

The central limitation is no longer data availability, but the ability to transform large volumes of SAR observations into meaningful, spatially consistent representations that support operational and scientific decision-making. Shipping stakeholders are primarily concerned with navigable ice concentration, lead detection, and ridging patterns that affect hull stress [5, 6]. Environmental researchers rely on spatial variability in ice type, melt progression, and seasonal transitions to understand ocean-atmosphere coupling and habitat dynamics; meanwhile, biological studies depend on accurate mapping of ice extent and persistence to model ecosystem shifts [7]. Across these domains, the challenge is to extract stable, discriminative features from highly variable SAR backscatter patterns.

SAR is an active microwave imaging system that transmits pulses toward the Earth's surface and records the returned backscatter amplitude and phase. By coherently combining echoes acquired along the satellite trajectory, SAR synthesizes a large antenna

[1] https://ai4eo.eu/portfolio/autoice-challenge/

aperture, enabling high spatial resolution independent of weather conditions or solar illumination [8,9]. These properties have made SAR the primary modality for operational sea ice mapping in polar regions.

The suitability of SAR for sea ice analysis arises from the physical interaction between microwave radiation and the surface of the ocean. Backscatter intensity is influenced by surface roughness, dielectric properties, salinity, and incidence angle. Ice types, snow cover, melt ponds, and open water roughened by the wind produce distinct scattering behaviors [10]. However, this same sensitivity introduces complexity. The side-looking geometry induces spatially varying backscatter as a function of incidence angle. Speckle noise, inherent to coherent imaging, affects local statistics. Seasonal transitions alter dielectric properties and texture appearance [11]. Consequently, SAR sea ice classification is fundamentally a representation problem: the objective is not merely to label pixels, but to learn features that remain informative under physically driven variability.

Early SAR-based sea ice classification relied on handcrafted texture descriptors and statistical modeling. Gray-level co-occurrence matrix (GLCM) features [12, 13], Bayesian classifiers [14], random forests [15], and Markov random fields [16, 17] aimed to capture spatial consistency through engineered statistics. Systems such as MAGIC [18] combined unsupervised segmentation with pixel-based classification to reduce non-stationary effects. These approaches demonstrated robustness in structured scenarios but depended heavily on local texture blocks and assumptions about class homogeneity. The advent of deep learning shifted the paradigm toward data-driven feature learning. Convolutional neural networks (CNNs) such as LeNet [19] and U-Net [20] variants enabled hierarchical representation extraction. In SAR sea ice mapping, CNN-based methods [21, 22] improved dense prediction capabilities but remained constrained by receptive field size and supervision geometry. Transformers and self-attention mechanisms [23] introduced the capacity to model long-range dependencies directly. Nevertheless, adapting these architectures to SAR imagery with pronounced non-stationarity has remained non-trivial.

Representation learning seeks to learn feature transformations from raw data that capture meaningful structure while suppressing irrelevant variability [24]. In computer vision, deep neural networks learn hierarchical representations that progressively encode textures, shapes, and semantic patterns [25]. These representations can be learned through supervised or self-supervised paradigms. Supervised learning shapes features directly through labeled targets, often achieving strong performance under sufficient annotation. However, representation quality is intrinsically tied to label geometry and domain-specific supervision. Self-supervised learning (SSL), including methods such as SimCLR [26], MoCo [27], BYOL [28], and Masked Autoencoders [29], instead learns representations through reconstruction or contrastive objectives.

SSL has demonstrated that large-scale pre-training can produce transferable representations across tasks. Importantly, in sea ice monitoring, the ability to learn robust representations that can be effectively adapted with limited annotated data is critical for downstream applications. At the same time, sea ice annotation differs fundamentally from common computer vision benchmarks. Operational ice charts provide polygon-based summaries of dominant ice conditions rather than pixel-level ground truth. These characteristics are not shortcomings, but design choices optimized for operational efficiency. As SAR spatial resolution becomes finer, the discrepancy between annotation granularity and image detail becomes more pronounced. Therefore, recent developments [30, 31] in automated and semi-automated mapping indicate a gradual shift toward finer-scale products, motivating representation strategies that can exploit high-resolution SAR data while respecting operational labeling limitations.

Recent initiatives further illustrate the growing interest in operationalizing AI-based sea ice analysis. For example, the Automated Sea Ice Products (ASIP) project developed by the Danish Meteorological Institute aims to generate pan-Arctic sea ice concentration products using deep learning models [32, 33]. These products are intended for integration into the Copernicus Marine Service portfolio and represent one of the first large-scale efforts toward operational AI-based sea ice products derived from SAR imagery. However, current systems primarily reproduce ice concentration products in formats consistent with traditional ice charts, and their evaluation typically relies on quantitative comparison with chart-derived annotations. While this approach measures agreement with existing products, it provides limited insight into whether the learned feature representations capture the spatial structures and contextual relationships that analysts use during manual interpretation.

1.1 Motivation

The general motivation of this thesis is to investigate how feature representations learned from SAR imagery can be made robust, scalable, and transferable under spatial variability and supervision constraints. Rather than treating classification accuracy as the sole objective, this work examines how representation structure, whether shaped through supervised architectural design or large-scale self-supervised pre-training, affects robustness and generalization in Arctic sea ice mapping. The specific motivations of this thesis are driven by the following research gaps:

1. Existing SAR-based sea ice classifiers predominantly rely on local receptive fields or post-processing corrections to mitigate non-stationarity. Global contextual modeling

and region-consistent representation learning have remained underexplored, particularly in a unified framework that integrates pixel-level and region-level reasoning during training. Moreover, as automated sea ice products derived from machine learning gain momentum, the absence of direct evaluation by operational ice analysts limits understanding of how these representations translate into reliable predictions for real-world interpretation.

2. Foundation models in remote sensing are largely trained on multispectral optical imagery and land-use datasets. SAR-specific pre-training for Arctic sea ice remains limited in scale and diversity, and it is unclear whether large-scale self-supervised learning alone can overcome annotation sparsity and supervision geometry constraints.

1.2 Objectives

The general objective of this thesis is to establish principled and scalable representation learning approaches for SAR imagery that integrate domain knowledge, architectural innovation, and self-supervised learning to improve the reliability and adaptability of automated sea ice mapping. To address the identified research gaps, this thesis pursues the following specific objectives:

1. To design and evaluate a structured supervised representation learning framework that explicitly integrates global spatial context and region-level coherence into SAR-based ice-water classification. The method focuses on moving beyond purely local receptive fields by incorporating long-range spatial dependencies and region-aware learning mechanisms during training. It seeks to investigate how architectural integration of convolutional modeling, self-attention, and unsupervised segmentation improve robustness to spatial non-stationarity while preserving boundary precision. The goal is not only to improve segmentation performance, but to shape feature representations that are consistent across heterogeneous SAR scenes and operationally meaningful.

2. To develop and systematically assess a SAR-specific self-supervised foundation model capable of producing scalable and transferable feature representations for multi-task sea ice analysis. The approach examines whether large-scale pre-training on unlabeled Sentinel-1 data can yield representations that generalize across tasks, resolutions, and supervision regimes. It includes evaluating representation behavior under varying data scales, model capacities, fine-tuning strategies, and resolution transfer

settings. The goal is to determine the extent to which self-supervised learning reduces dependence on annotation geometry and sensor-specific distributions, and to clarify whether pre-training alone is sufficient to ensure robustness and adaptability.

1.3 Thesis Structure

The remainder of this thesis is organized as follows:

Chapter 2 introduces the Irregular Tokens on Transformers (ITT) framework, which integrates CNNs, transformers, and unsupervised segmentation to model long-range spatial dependencies under supervised learning. The chapter further evaluates region-consistent representation learning and examines the operational relevance of the approach through expert-driven validation aligned with sea ice charting practices.

Chapter 3 presents SAR-IceFM, a self-supervised SAR foundation model designed to learn transferable feature representations from large-scale unlabeled Sentinel-1 imagery. The chapter systematically investigates representation scaling, resolution transferability, and multi-task fine-tuning strategies, and analyzes the extent to which self-supervised pre-training enhances robustness under supervision and domain constraints.

Chapter 4 synthesizes the findings and discusses implications for future operational and large-scale sea ice monitoring systems.

Chapter 2

ITT: Long-range spatial dependencies for sea ice semantic segmentation

Deep learning models for SAR sea ice mapping often focus on local pixel neighborhoods, limiting their ability to capture large-scale patterns and account for within-class variability and spatial non-stationarity across full scenes. In this chapter, we propose strategies to leverage both global and local contextual information to improve automated ice-water classification. First, we introduce the Irregular Tokens on Transformers (ITT) framework, which uniquely combines convolutional neural networks (CNNs), transformers, and unsupervised segmentation to capture long-range spatial dependencies through self-attention applied to irregular, homogeneous regions. Second, we address non-stationary statistics through a multi-task training scheme that integrates a pixel-based loss for dense semantic segmentation with a region-based loss that promotes region-consistent feature representations. Evaluation on dual-polarized RADARSAT-2 scenes demonstrates reduced sensitivity to non-stationary patterns and lower uncertainty during melt and freeze-up conditions. We then extend the analysis to region-based inference and hybrid combinations of pixel- and region-level predictions, highlighting their complementary behavior. Finally, we assess the operational viability of the proposed approach through an expert-driven audit conducted by the Canadian Ice Service (CIS), which reveals both the strengths and season-dependent limitations of the ITT framework. The following sections contextualize the problem of sea ice mapping in SAR imagery and motivates the need for representation strategies that explicitly address spatial non-stationarity.

2.1 Introduction

Monitoring sea ice coverage is fundamental for understanding climate change and conducting human activities in northern territories. Dynamic sea ice patterns, with significant area during winter, influence Earth's albedo, atmospheric circulation, and environmental conditions globally [34]. Besides climate implications, sea ice acts as a key barrier to commercial shipping. Particularly in the Canadian Arctic, alternative shipping routes offer economic advantages during summer [35], but frozen seas disrupt sectors such as fishing, resource extraction, and tourism [36]. Given the environmental, economic, and logistical dependencies on sea transportation, monitoring and studying sea ice coverage are vital for mitigating risks to shipping operations and addressing hazards in the Arctic.

The CIS provides ice mapping information through various products, including ice concentration, types, age, and motion. Despite advancements in remote sensing technology, the current process still relies on the manual efforts of human experts. All national ice agencies, including CIS, use synthetic aperture radar (SAR) images to generate ice charts that offer a coarse description of ice concentration and predominant ice types within specific regions [37] (see Fig. 2.1). While such products aid analysts, they lack pixel-level information relevant to locate and describe the ice formation, for instance, when identifying leads and ice floes. Although some approaches focus on refining sea ice maps from ice charts [31], there remains the need for automated models that can generate pixel-level maps while handling vast amounts of data.

SAR imagery has been pivotal in sea ice mapping for decades. As active sensors, SAR systems provide the energy to illuminate target objects, and the operation frequencies are mostly unaffected by the atmosphere. The inherent penetration capability of the signal delivers essential information on material thickness and roughness [10], crucial for accurate ice classification. Nonetheless, the images show challenging visual artifacts such as speckle noise and the incidence angle effect given the side-looking geometry of SAR, which generates non-stationary patterns [18], mainly in surfaces with specular properties. Moreover, the signal ability to penetrate ice, variations in ice thickness, snow coverage, and substantial inter-class variance, particularly during melt or freeze-up periods, contribute to uncertainties when attempting to extract information from the scenes [11].

The historical approaches developed for sea ice classification using SAR images encompass a manifold of machine learning techniques. These include Bayesian estimation [14], support vector machines (SVMs) [18], and artificial neural networks (ANNs) [38]. Typically, these approaches are complemented by knowledge-driven strategies for feature extraction and unsupervised image segmentation [17,39]. More recently, other methods [21,40–46]

based on deep learning architectures have been applied to classify sea ice, showing significant improvements. Nevertheless, there are ongoing limitations in utilizing extensive contextual information to address non-stationarities. Capturing global information is constrained by the designed topology or by the small receptive field of the networks. Furthermore, despite the methodological progress reported in the literature, these approaches have not yet been adopted for operational use by national ice agencies, highlighting the gap between research developments and real-world sea ice analysis workflows.

In this chapter, we leverage both global and local contextual information for sea ice semantic segmentation, using two deep learning paradigms: CNN [19] and transformers [23, 47, 48]. CNNs excel at capturing local patterns by employing convolutions. Transformers, based on the self-attention mechanism, treat the image as a sequence of elements that represent different parts of the scene, allowing the use of long-range spatial dependencies. The literature contains comprehensive assessments [49–58] suggesting that a fusion of CNNs and transformers has the potential to improve classification. Furthermore, we provide explicit treatment to the boundary definition between target objects by including the unsupervised segmentation method Iterative Region Growing on Semantics (IRGS) [39], which has been thoroughly evaluated [15, 59] to identify homogeneous regions on SAR imagery. The specific contributions of this study are as follows:

1. We introduce the novel **I**rregular **T**okens on **T**ransformers (ITT) framework that combines CNNs, transformers, and unsupervised segmentation to successfully classify ice and open water in SAR images. The model effectively incorporates local and global context to address non-stationary statistics, ensuring confident predictions and more detailed segmentation around boundaries.

2. To the best of our knowledge, the method constitutes the first attempt to uniquely implement self-attention on tokens extracted from irregular, homogeneous, and multi-scale regions within the input image. Therefore, this suggests utility not only for sea ice classification, but also for other remote sensing problems. The code is publicly available[1].

3. To conduct a thorough evaluation, we present three experimental protocols employing dual-polarized images from the RADARSAT-2 sensor [60]. This design enables evaluation across a broad spectrum of seasonal ice-water signatures over two years.

4. Beyond quantitative metrics, we incorporate an expert-driven audit conducted by analysts from the CIS, providing an operational evaluation of classification quality.

[1] https://github.com/jnoat92/ITT_sea_ice.git

This assessment bridges methodological performance with practical deployment considerations.

The remaining sections of this chapter are structured as follows: Section 2.2 offers a concise literature review on recent methods for sea ice classification using SAR. Section 2.3 illustrates the manifestation of non-stationary behavior in the data. Section 2.4 derives the ITT method, which combines CNNs, transformers, and unsupervised segmentation. Addres-sing non-stationary statistics is discussed in Section 2.5. Section 2.6 outlines the experimental protocol employed to evaluate the proposed design. Results and discussions are presented in Section 2.7, while Section 4 provides a summary of the research outcomes and outlines potential future directions.

2.2 Background

The non-stationary nature of SAR signatures, influenced by noise and the sensor's incidence angle, presents a challenge for classification. Moreover, the descriptive capacity of a single pixel in SAR images is not as robust as that in multi-spectral or hyperspectral images. For SAR, analysts often rely on the relative intensity and texture appearances to discern ice types [12]. Many publications adopt a feature engineering pre-processing approach, specialy extracting texture features from gray-level co-occurrence matrices (GLCM) [61]. For example, Zakhvatkina et al. [14] selected texture features for ice type classification on HH polarized images. The authors used a Bayesian classifier, obtaining the conditional probabilities from the histogram of backscattering coefficients per class and assuming constant priors. Such considerations are strong constraints when the statistical distribution of the data is not stationary or with high class variance. In a different approach, Lu et al. (2023) proposed a random forest (RF) [62] classifier for ice-water classification, utilizing polarimetric and texture features from fully polarized SAR imagery. Despite an exhaustive feature search, the limitations arise as features are derived from small pixel blocks, thereby not leveraging global context.

Methods based on Markov random fields (MRFs) [16] have also contributed to sea ice mapping since these explicitly account for spatial interactions. Deng et al. [17] proposed an MRF for the unsupervised segmentation of SAR images leveraging spatial and feature information. The method shows robustness on non-stationary characteristics; however, since it is unsupervised, it is more valuable to guide analysts than to perform automatic inference. IRGS [39] (Section 2.4.2), also unsupervised, further improved the MRF using the image intensity values and introducing policies for edge penalties on the segmentation

and a region growing stage. As a result, the outcome follows natural boundaries on SAR images for sea ice.

Map-Guided Sea Ice Classification (MAGIC) [18] aids ice analysts through advanced class labeling algorithms implemented in a software package that incorporates texture feature extraction, unsupervised segmentation, and other experimental methods. It employs IRGS in two stages (local and global), reducing non-stationary effects for robust segmentation. MAGIC can enhance classification by incorporating predictions from a pre-trained classifier. Despite potential noisy outcomes, it refines segmentation through IRGS optimization, considering semantic classes (ice-water). This strategy combines region-based and pixel-based classification. Jiang et al. [15] extended the method to different ice types using an RF as the classifier. In practice, MAGIC serves as an effective post-processing refinement that corrects localized misclassifications using region-based optimization; however, global contextual dependencies are not explicitly incorporated during the learning stage, as the underlying classifiers rely primarily on local features.

Deep learning models enable the extraction of features simultaneously with training the classifier. Accordingly, Boulze et al. [21] propose a CNN for ice mapping on dual-polarized Sentinel-1 images. The approach modifies the popular LeNet [19] architecture designed to classify the center pixel of 50×50 subimages. Generally, the resulting predictions follow the coarse labels created from ice charts; however, the outcome is noisy and uncertain in areas with abundant texture. Chen et al. [63] treat data and model uncertainties with Bayesian CNNs, improving scene interpretation. Wang et al. [41] use the U-Net [20] architecture to produce dense predictions and diversify the receptive field of the model. Inspired by the combination of pixel and region-based classification proposed in the MAGIC system [18], Jiang et al. [40] suggests a scheme using the segmentation from IRGS and a CNN as a classifier. In this case, the same label (ice or water) is assigned to all the pixels within homogeneous regions, using an energy function and a decision rule that depends on the soft predictions from the CNN. No learning is involved in the approach, and, similar to Leigh et al. [18] and Jiang et al. [15], each region is expected to represent a unique class. We are changing that assumption in this chapter (Section 2.4.3).

Towards capturing more context, other approaches in the literature include attention mechanisms [45, 58], innovative training protocols [64], and graph neural networks [43]. Overall, methods for capturing global information are limited, especially for large SAR images. We address those concerns with a novel method by explicitly leveraging global and local context, integrating CNNs, transformers, and unsupervised segmentation.

2.3 Non-stationary statistics on SAR imagery

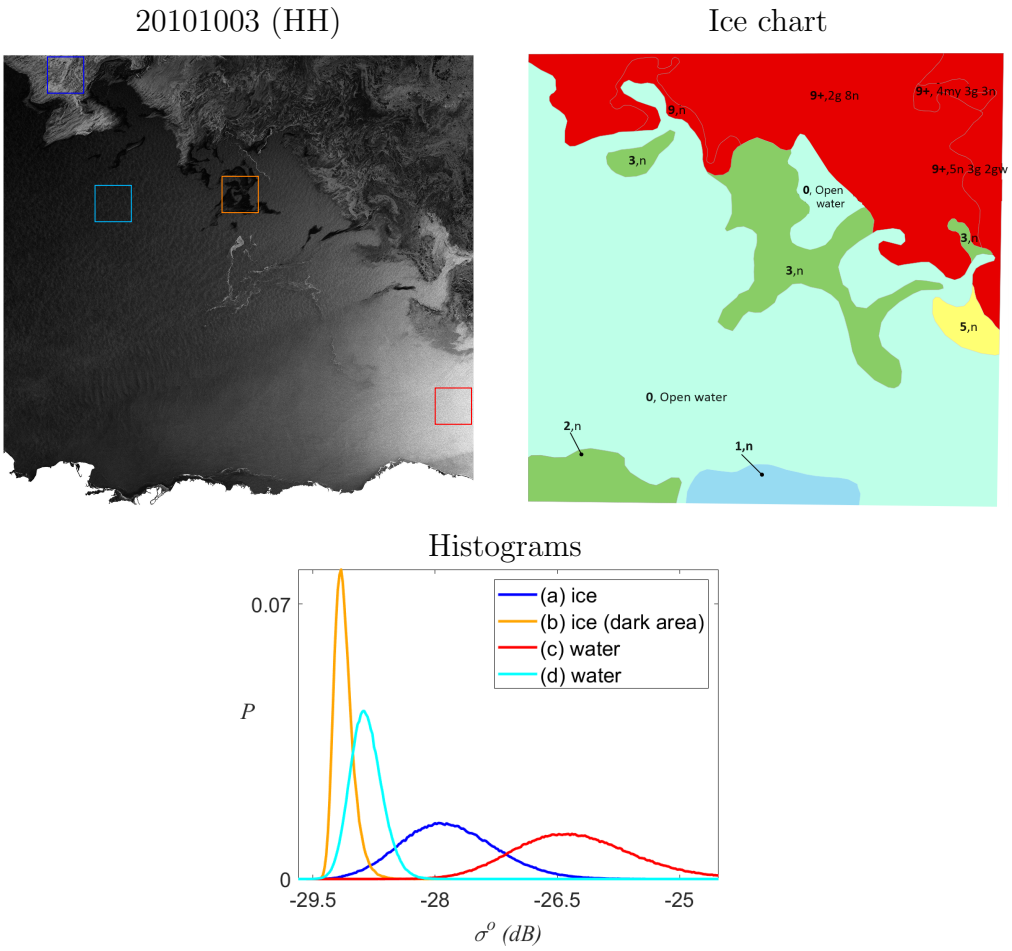

Figure 2.1: Non-stationary behavior on SAR scenes for sea ice mapping. The image on the left is a gray scale illustration of the HH polarization. Four scattered patches are indicated with their correspondent histogram. The ice chart shows a coarse segmentation of the scene indicating for each region the concentration of ice types. For example: the code "9+, 5n 3g 2gw" indicates high ice concentration (9+/10), with 5/10 new ice, 3/10 gray ice, and 2/10 gray-white ice.

Certain features in SAR images used for sea ice classification become evident only when analyzing a sufficiently large expanse of the landscape, as noted by Leigh et al. [18], particularly when distances exceed 50 km. Such extensive scenes manifest spatially

statistical variations, called non-stationary, attributed to factors such as the incidence angle effect, acquisition and synthesis imperfections such as antenna mechanism deviations and quantization errors [65], and the fuzzy nature of ice-water boundaries. Additionally, variations in weather conditions influence the characteristics of target objects, leading to changes in backscattering, mainly during the melt season.

To illustrate an example of non-stationary patterns, Fig. 2.1 shows, from a SAR scene, the radiometric calibrated HH polarization (σ^o). The image represents a region in the Beaufort Sea, recorded by the RADARSAT-2 sensor on October 3rd, 2010, using the ScanSAR Fine product with beam mode ScanSAR Wide. The figure also shows the normalized histograms from small patches highlighted in color, and the ice chart provided by the CIS. The image dimensions are around 10000×10000 pixels, and the patches are 800×800 pixels.

From the histograms, we can observe the high within-class variance on ice and water. The water inside the cyan patch looks darker and with more texture than water in the red patch. The brightness difference is an effect of the incidence angle, which increases as we move spatially from near range (right) to far range (left). The ice in the blue patch appears very heterogeneous, while the ice (new ice, dark area) in the orange patch is visibly dark and more homogeneous, although we often find texture attributes in regions covered by ice. Classifying these attributes results in predictions that are uncertain, ultimately leading to inaccurate outcomes and, separating the classes becomes difficult analyzing the patches independently. Conversely, observing the whole extension of the image provides a better recognition of ice and water in any region of interest. In that sense, the global information is substantially beneficial for classification models to handle non-stationarities.

2.4 Self-attention on irregular tokens for semantic segmentation

2.4.1 Self-attention complementing convolutions

CNNs generate local patterns by processing spatially adjacent input pixels from square windows, assuming pattern redundancy across the entire image. While this property enhances CNN efficiency through shared kernel weights, it introduces an inductive bias that makes the network scale variant and prone to failure when class statistics lack spatial stationarity. One way to address this limitation is by increasing the CNN receptive field [66], which defines the region of the input image that a single activation in the output layer

can capture [67]. This can be achieved through various techniques, such as using larger kernels, hierarchically incorporating additional layers, skipping some pixels (*stride*) during the sliding, or applying dilated convolutions [68]. However, these methods come with computational costs, which pose a significant challenge when processing large remote sensing images. Some approaches from the literature [20, 69–71] can diversify the receptive field; nevertheless, it still remains constrained by size and shape. For instance, if the ResNet-50 [69] architecture were used to classify the SAR image in Fig. 2.1, its receptive field, smaller than the drawn patches, would be limited to local coverage relative to the overall scene size.

The concept of self-attention was introduced in transformer networks for applications in natural language processing (NLP) [47]. It involves learning dependencies among feature vectors, called tokens, within an input sequence. The attention process, described in Equation 2.1, utilizes vectors for queries (Q), keys (K), and values (V) (analogous to retrieval systems), which are features derived from the tokens through multi-layer perceptron (MLP) layers. In essence, tokens exhibiting a strong relationship in the sequence deliver meaningful activations in the generated attention matrix QK^T, thereby aiding prediction tasks. The term d_k is the dimensionality of the key vectors and is used to avoid large gradients by scaling the values of QK^T.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.1}$$

Transformers applied to images [23, 48, 72, 73] treat the input as a sequence of square non-overlapping patches, each represented as tokens in an embedding space. The tokens form a sequence from which the model learns feature interactions, effectively capturing global or large-scale context from the input image without the inductive bias inherent in convolutions. There is, however, a downside to the tokenization process. Tokens may not correspond to a unique class when extracted from a regular grid. In addition, the area covered by each token defines how local the resulting features are. This aspect is critical, as reducing the area or increasing the input image size results in longer sequences of length L, which becomes difficult to scale given the attention complexity $O(L^2)$ in terms of both memory and computational requirements. For this chapter, we adopt from transformers only the self-attention mechanism and propose a novel approach to create tokens from irregular and homogeneous regions (Section 2.4.3).

2.4.2 Unsupervised segmentation: IRGS

On sea ice SAR images, the boundary definition separating two different classes shows a high uncertainty for semantic segmentation [63]. Therefore, maintaining a high accuracy in boundaries is challenging for classification methods. CNNs, for instance, could deliver high level features strongly correlated with the ground truth, not considering low level features obtained at early layers, where simple patterns such as edges are actually detected. Regarding transformers, the tokens represent square patches that contain more than one class, and the boundaries might not be recovered on the prediction.

A highly effective approach for achieving precise boundaries involves unsupervised segmentation based on pixel intensity values. The Iterative Region Growing using Semantics (IRGS) method [39], widely employed in full SAR scene segmentation [15, 18, 40, 43], exhibits a robust ability to delineate ice-water boundaries resembling visual interpretation. This method, grounded in MRF, utilizes edge strength between homogeneous regions and employs a region-growing technique [74] to yield stable and accurate segmentation. A key aspect of our methodology involves utilizing regions from IRGS to form tokens in transformers. With the appropriate over-segmentation, each region is expected to be homogeneous and represent a unique class, enabling a region-based classification that ensures well-defined boundaries. For this work, we used the implementation of IRGS developed for the operational tool MAGIC [18].

2.4.3 ITT: Irregular Tokens on Transformers

As to the first contribution, the approach introduced in this chapter involves the combination of three different components: CNNs, transformers, and unsupervised segmentation. The objective is to leverage each component's respective characteristics in a complementary manner, as outlined below:

- CNNs are appropriate for local features. They do not define boundaries in an explicit way, and the receptive field is limited.
- Transformers do not rely on 2D arranged samples; instead, they operate on 1D sequences derived from different parts of the scene, which allows for the extraction of global information, thereby expanding the receptive field.
- IRGS produces segments that follow natural boundaries, producing an over-segmented scene with closed irregularly shaped regions.

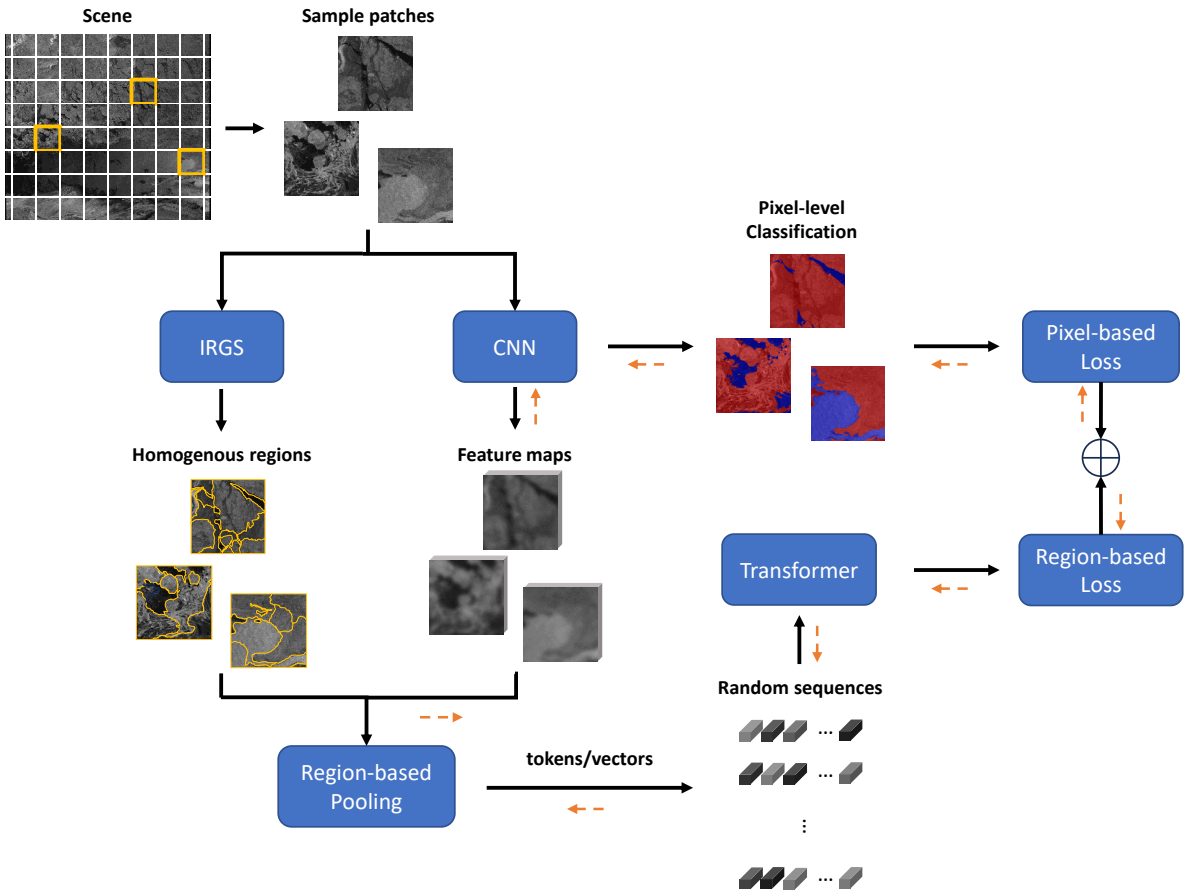

Figure 2.2: ITT framework: combination of CNNs, transformers, and unsupervised segmentation using IRGS to leverage both local and global information. The CNN branch produces a pixel-level classification, while the transformers learns region-level classes from embedded features. Both loss terms influence the CNN through back-propagation (dashed arrows).

Fig. 2.2 shows the ITT framework, which starts by extracting a batch of sample patches from arbitrary locations within the entire SAR scene to feed a CNN and produce feature maps. This procedure is a standard approach in processing remote sensing images for deep learning. Otherwise, resizing the image to lower dimensions would be necessary to meet memory requirements, and downsampling results in the loss of fine-grained information necessary for generating high-resolution prediction maps. Applying IRGS on the sample patches produces an over-segmentation with homogeneous regions that are likely to belong

to a unique class. These segments guide the creation of a sequence of region descriptors employed in the self-attention mechanism to gather information from various patches simultaneously. IRGS is applied only on the HV polarized channel since it is less sensitive to the incidence angle effect, thus offering, in general, more separability for ice and water [18]. Since the approach aims to over-segment the input images, artifacts such as low signal-to-noise ratio and sub-swath discontinuities in the HV channel do not compromise IRGS performance.

The region-based pooling generates vectors by averaging the CNN feature maps within each region. This step constitutes the tokenization for the transformer. Therefore, a token is created from each region and arranged to compose a sequence. Some factors, such as the input patch size, parameters utilized in unsupervised segmentation, noise, and the heterogeneity of objects in the image, contribute to a considerable variation in the resulting number of regions. For instance, with a single 128×128 pixel patch using 10 arbitrary classes on IRGS, the sequence length might extend to approximately a thousand tokens, posing computational challenges for transformers. To address this, we partition the sequence into several chunks of a fixed length to create a batch of shorter sequences that better suit the hardware constraints. If these new sequences are not the same length, the shorter ones are padded to ensure all sequences can be stacked in the same batch to feed the transformer. This requires to mask out the padded tokens during the attention.

A naive approach to refine the boundaries and reduce noisy outcomes for the CNN involves applying majority voting by choosing the most frequently predicted class within the regions provided by IRGS. Even assuming that the over-segmentation produces regions covering a single true class, from the CNN prediction, the most repeated class within a given region can be incorrect. Therefore, lacking a learning component, majority voting has limitations. The topology shown in Fig. 2.2 addresses this drawback by following two target tasks during training. Hence, we use the loss function $\mathcal{L} = \mathcal{L}_0 + \lambda\mathcal{L}_1$ consisting of a pixel-based loss $\mathcal{L}_0$ for the CNN (2.2) and a region-based loss $\mathcal{L}_1$ for the transformer (2.3). The hyperparameter λ defines the relative importance between the loss terms. Given an $N \times M$ input image and considering C classes, $\mathcal{L}_0$ calculates the cross-entropy loss between the predicted probabilities $\hat{y}_{nmc}$ and the ground truth y_{nmc} for each pixel nm. Similarly, $\mathcal{L}_1$ applies cross-entropy but considers predictions $\hat{y}_{tc}$ and the ground truth y_{tc} for the created tokens/regions (T) instead.

$$\mathcal{L}_0 = -\frac{1}{NM}\sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{c=1}^{C} y_{nmc}\log(\hat{y}_{nmc}) \tag{2.2}$$

$$\mathcal{L}_1 = -\frac{1}{T}\sum_{t=1}^{T}\sum_{c=1}^{C} y_{tc}\log(\hat{y}_{tc}) \tag{2.3}$$

The region-based loss $\mathcal{L}_1$ incorporates a learnable component intended to improve the region consistency of the CNN feature maps through backpropagation. Training in a multi-task scheme, both network branches benefit from each other. The transformer promotes more accurate CNN predictions on boundaries, and the CNN delivers deep features more correlated with the region-based target. In principle, any layer activation from the CNN, or even the input images, can be used to create the token sequences. At test time, we would use only the CNN for the semantic segmentation. Unlike previous approaches [15,18,40,63] using IRGS for sea ice classification, the semantic segmentation at inference time is not sensitive to the configuration of regions; therefore, fine-grained details are not excluded from the prediction as could happen by filling IRGS regions with a single class.

Regarding our second contribution, the tokenization from irregular regions is a relevant novelty for the ITT model. Current transformer-based methods for vision tasks use square patches to represent the tokens. With unsupervised segmentation, we define homogeneous regions at different scales and shapes to create the tokens, making the model more scale-invariant. Moreover, since the regions are likely part of a single class, the attention fosters learning class-specific features. This approach also opens the opportunity to leverage regional descriptors for other remote sensing tasks.

2.5 Addressing non-stationary statistics

From the proposed method, the CNN extracts local features while the transformer learns from scattered and distant regions on the whole scene. Nevertheless, Fig. 2.1 motivates that global information is needed to deal with non-stationary statistics. To achieve it, ITT must meet two conditions at training time.

1. The input patches must be scattered across the scene, and their quantity must exceed 1. Randomly selecting samples is sufficient to satisfy this criterion. The CNN processes a batch of patches in parallel, ensuring that pixels from one patch do not impact the processing of another patch. As a result, only context information within the receptive field is leveraged. In contrast, the transformer can process tokens/regions from different patches in the same sequence.

2. Any sequence that feeds the transformer must contain tokens from different input patches; otherwise, the field-of-view is reduced to the extent of the patch. To address this, sequences are generated randomly, allowing the consideration of dependencies between regions spatially distant from each other.

2.6 Experiments

2.6.1 Dataset

The dataset used in this work includes 33 dual-polarized SAR scenes captured between April 2010 and October 2011. As Fig. 2.3 shows, the scenes primarily cover regions of the Beaufort Sea, in the Arctic Ocean. Most dates during the winter are not included since the sea is fully covered by ice. The sensor used to capture the images is RADARSAT-2, and the product for each scene is ScanSAR Fine (SCF) with beam mode ScanSAR Wide A (SCWA). The image coverage is $500km^2$ ($500km$ of swath width at range and azimuth directions) with a nominal pixel spacing of $50m$ at ground range. The incidence angle for this product increases from 20 to 49 degrees from near range to far range, exhibiting the non-stationary effect in most scenes. Each scene provides data in two polarizations: HH and HV. Although the original image sizes are approximately 10000×10000 pixels, we employed downsampled versions obtained by averaging every 4×4 pixel consecutive windows. Working with the resulting resolution reduces computational demands and is adequate for producing detailed prediction maps. All images were calibrated to obtain the σ^o signatures.

We generated the ground truth maps, manually, based on the ice charts provided by the CIS and visual interpretation, only considering ice and water. The ground truth for each region/token was created implementing majority voting of the ground truth maps on IRGS regions. The occurrence of ice and water in the scenes is closely tied to the season, resulting in a dataset that exhibits considerable range of patterns throughout the year.

2.6.2 Experimental setup

To assess operational capabilities of ITT, we evaluated 33 models by taking on each case 32 scenes for training and the remaining scene for testing (leave-one-out policy). From each training scene we used 20% to validate the models. Specifically, we divided the training scenes on a grid of 5×3 pieces of the same size, from which we predefined 3 for validation

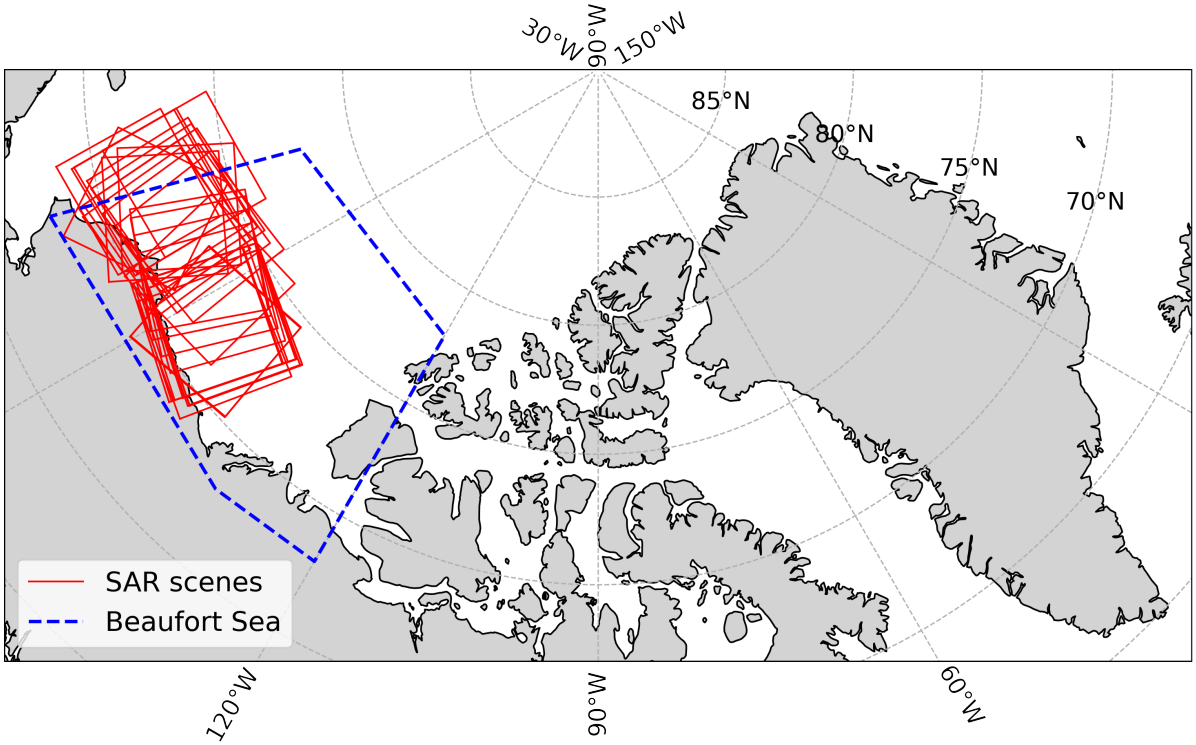

Figure 2.3: Study area of the dataset. The map shows the Western Hemisphere of the Arctic Ocean. The scenes are predominantly located on the Beaufort Sea.

and the remaining 12 for training. From these sets, patches of 224×224 pixels were extracted using a sliding window procedure with 5% overlap between consecutive patches. This results in around 5100 patches for training and 990 for validation per model.

The proposed method is versatile and compatible with any CNN architecture; for simplicity, we opted for the widely recognized U-Net [20], given its recent success for sea ice classification on SAR imagery [3, 41]. Initially designed for semantic segmentation of biomedical images, U-Net features a fully convolutional neural network (FCN) that incorporates skip connections through concatenation between encoder and decoder layers. Details of the layers employed are outlined in Table 2.1, specifying the kernel size, number of filters (if applicable), and strides for each operation, including convolutions (C), max-pooling (P), and transposed convolution (T). Following each 3×3 convolutional layer, batch normalization and the ReLU activation function are applied. The 1×1 convolution stands for the pixel-based prediction.

Table 2.1: U-Net architecture: C, P, and T denote convolution, max-pooling, and transposed convolution, respectively.

Encoder	**Decoder**
C(3×3, 32, 1) $\times$ 2 P(2×2, 2)	T(3×3, 256, 2) C(3×3, 256, 1) $\times$ 2
C(3×3, 64, 1) $\times$ 2 P(2×2, 2)	T(3×3, 128, 2) C(3×3, 128, 1) $\times$ 2
C(3×3, 128, 1) $\times$ 2 P(2×2, 2)	T(3×3, 64, 2) C(3×3, 64, 1) $\times$ 2
C(3×3, 256, 1) $\times$ 2 P(2×2, 2)	T(3×3, 32, 2) C(3×3, 32, 1) $\times$ 2*
C(3×3, 512, 1) $\times$ 2	C(1×1, 2, 1)

* Layer providing features to create tokens.

For the transformer implementation, we adapted the Vision Transformer (ViT) [23] by eliminating the token embedding, as this step was unnecessary because of the presence of the region-based pooling. Tokens were generated from the second last layer of U-Net, ensuring that the feature maps had already attained the input resolution. The transformer structure consists of a sequence of 8 self-attention blocks. As Fig. 2.4 shows, each block contains two residual stages. The initial stage applies multi-head attention across 6 contiguous parts (heads) of the tokens, while the second stage encompasses two multi-layer perceptron (MLP) layers. Following the original transformer implementation, we apply layer normalization before each stage and the GELU activation function after each MLP layer. Finally, we add another MLP layer for the region-based loss.

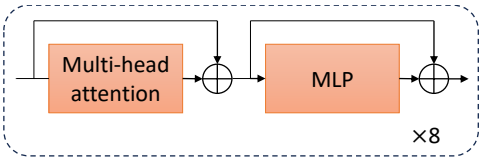

Figure 2.4: Self-attention block.

We have devised three protocols to evaluate the effectiveness of the ITT method, by gradually adding components during training:

- **U-Net**: Use only $\mathcal{L}_0$, which reduces the method to the CNN block. This protocol constitutes the baseline.

- **ITT**: Multi-task: Train from scratch using both losses ($\mathcal{L}_0 + \lambda\mathcal{L}_1$).
- **U-Net+ITT**: Multi-task: From pre-trained CNN (obtained from U-Net), train using both losses ($\mathcal{L}_0 + \lambda\mathcal{L}_1$).

The input batch size is 8, and the sequences length in the transformer is 400. IRGS generates regions for each input patch considering 50 arbitrary classes to ensure over-segmentation, resulting in hundreds of homogeneous regions for the chosen input size. The Adam optimizer [75] is employed with a learning rate $\alpha = 0.00013$. Following 10 or more consecutive training epochs without improvement on the validation set, the learning rate decreases by 10%. The training stops if there is no improvement after 15 epochs. The maximum number of training epochs is 100. These hyperparameters were tuned using a random search, training 40 times on 20 scenes and selecting the combination of values with the best performance in the validation set. Tuning was initially performed on U-Net and then on the multi-task protocol ITT to refine λ, which was finally 2.62. During inference, we used the whole scene, without slicing into patches, to feed the network.

The patch extraction and the unsupervised segmentation were done offline before training the models since IRGS only runs on CPUs. We used four Tesla V100-SXM2-16GB GPUs to train each model. Accordingly, on 100 epochs, it requires 3.5 hours and 40.3 hours to train the CNN and the whole model, respectively. The computation of tokens from irregular regions imposes a bottleneck that makes the multi-task protocols slower. The inference (U-Net inference) time for a scene using one GPU is 0.85 seconds.

2.7 Results and discussion

Based on the conducted experiments, the results are presented using overall accuracy (OA) at pixel level as evaluation metrics. The accuracy is sufficiently reliable as the dataset maintains mildly different proportions of ice and water, 60% and 40% respectively. Metrics such as precision and recall are in the same order and delivered the same conclusion as OA; without indication of biases related to class imbalance. Table 2.2 displays the OA for the established experimental protocols across each model, following the leave-one-out policy. The upper part of the table describes the ablation settings concerning the different components utilized in each protocol. The second column specifies the acquisition date for the test scene in the format *yyyymmdd*. The best score in each row appears in bold. For protocol U-Net, we applied majority voting for comparison purposes. Note that when working with large images of approximately 2500 × 2500 pixels, the relatively small

difference in accuracy among protocols represents a large number of pixels. For instance, a difference of 0.1% represents more than 6000 pixels.

Table 2.2: Overall accuracy (%) for each experimental protocol following the leave-one-out policy. Results applying majority voting (MV) are included. Rows are light-cyan for test scenes within the freeze-up season. Rows are light-red for test scenes within the melt season. The bottom line shows the average metric across all models, including the standard deviation.

Ablation settings	CNN	✓		✓	✓
	Transformer			✓	✓
	From pre-trained CNN				✓
	Multi-task			✓	✓
Model #	Test date	U-Net	U-Net+MV	ITT	U-Net+ITT
1	20100418	99.90	99.96	99.89	**99.98**
2	20100426	99.92	99.96	99.88	**99.98**
3	20100510	99.25	**99.31**	98.96	99.23
4	20100524	98.89	98.86	98.66	**99.11**
5	20100605	98.95	**99.20**	98.99	98.25
6	20100623	97.83	97.67	**99.01**	98.85
7	20100629	97.16	96.89	96.66	**97.75**
8	20100704	95.49	**98.57**	96.67	97.31
9	20100712	93.75	**94.53**	91.11	93.08
10	20100721	93.90	94.25	**94.58**	94.54
11	20100730	90.20	90.48	92.22	**92.33**
12	20100807	88.67	87.63	**92.12**	91.23
13	20100816	**91.40**	85.49	90.61	91.09
14	20100907	99.84	**99.89**	99.85	99.84
15	20100909	**95.09**	94.45	94.87	**95.09**
16	20101003	95.96	95.94	95.23	**96.01**
17	20101014	99.28	99.53	**99.78**	99.35
18	20101017	82.09	82.42	80.99	**93.35**
19	20101021	97.06	97.01	96.77	**97.46**
20	20101025	95.77	**96.20**	95.54	95.87
21	20101027	93.20	92.95	92.63	**93.45**
22	20101114	97.96	97.88	**97.98**	97.90
23	20101120	99.24	**99.30**	99.28	99.10

Continued on next page

Table 2.2 – continued from previous page

Model #	Test date	U-Net	U-Net+MV	ITT	U-Net+ITT
24	20101206	97.93	98.06	**98.14**	98.13
25	20101214	99.38	99.49	**99.63**	99.61
26	20110530	99.25	99.25	**99.34**	99.12
27	20110613	99.54	99.50	**99.55**	99.48
28	20110627	97.78	97.68	97.39	**98.10**
29	20110709	97.66	97.13	97.62	**97.82**
30	20110710	97.05	96.31	97.24	**97.29**
31	20110720	96.11	96.11	**96.47**	96.14
32	20110725	94.34	94.04	94.76	**94.89**
33	20111029	94.35	93.73	94.87	**95.24**
Average		96.19±3.85	96.05±4.25	96.28±3.87	**96.85±2.64**

Table 2.2 shows that U-Net had the best accuracy in two cases, outperforming the remaining protocols only on model 13. Applying majority voting as post-processing did not outperform U-Net on average, although it showed improvements in around half of the 33 models. In the remaining cases, the result worsens as some regions are completely assigned to the wrong class (see Fig. 2.5). The outcome for ITT is the result of training the whole model end-to-end from scratch. It produced a slight improvement on average, which suggests some positive influence from the transformer branch on the feature maps produced by the CNN. Finally, U-Net+ITT also performs the training in a multi-task scheme, although initializing the CNN parameters with those obtained using U-Net. U-Net+ITT delivered the most successful prediction, scoring 14 times the highest accuracy, being better than U-Net in 23 cases, and showing the highest average. U-Net+ITT was also the most stable approach since it shows the lowest variance. As expected, there are notable improvements with respect to the CNN baseline when all the ablation settings are combined. The most evident cases were model 12 scene 20100807 and model 18 scene 20101017, challenging scenes captured during the melt and freeze-up season, respectively.

Recalling the scene 20101003 analyzed in Section 2.3, Fig. 2.5 shows the dual-polarized SAR images, the segmentation from IRGS, the ground truth, and the prediction maps from each experimental protocol. In the Beaufort Sea, acquisition dates from October 2010 align with the freeze-up period [76]; therefore, the scene contains new ice (magnified area) as a challenging pattern to detect, mainly because ice signatures are not usually the darkest and smoother in the image.

U-Net predicted most of the new ice pixels as open water, and the majority voting could not improve such prediction. The outcome from U-Net within IRGS regions was

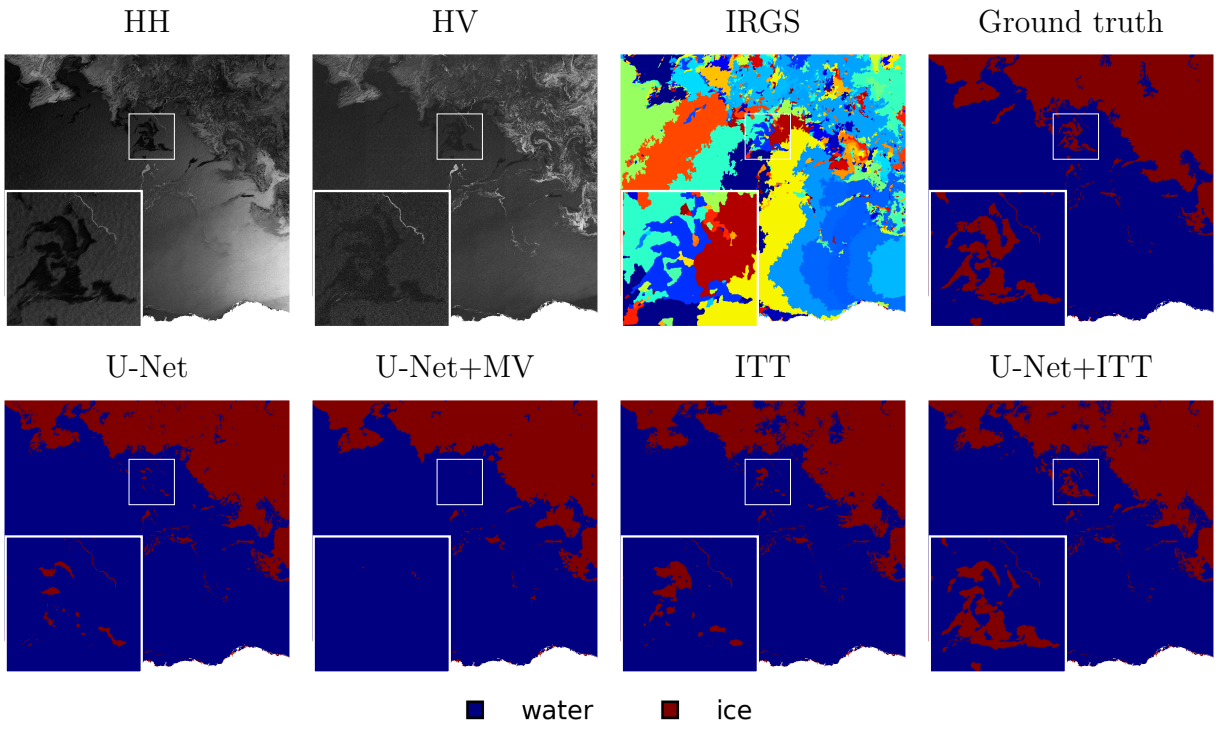

Figure 2.5: Semantic segmentation for scene 20101003. The first row shows the SAR images, the unsupervised over-segmentation from IRGS, and the ground truth. The second row shows the prediction maps.

incorrect since it does not leverage global information. In addition, the dependency from IRGS during inference relies on how precise the over-segmentation is. IRGS performance is affected by the spatial non-stationary statistics, which means the over-segmentation is better when applied at local level [18]. The predictions from ITT and U-Net+ITT are not sensitive to the unsupervised segmentation during the test. However, during training, they are influenced by the region-based loss, with a more reliable over-segmentation since IRGS operates on small patches. The effective capture of long-range contextual information and the pre-training of the CNN parameters lead to the best performance on protocol U-Net+ITT detecting the new ice. Although a segment with low local contrast, located at far range, remained challenging for any approach to recognize.

We can observe a similar result in Fig. 2.6. Scene 20101017, also from the freeze-up season, shows a sudden change of texture and illumination on open water that U-Net misclassified and only U-Net+ITT could rectify, more accurately matching the ground

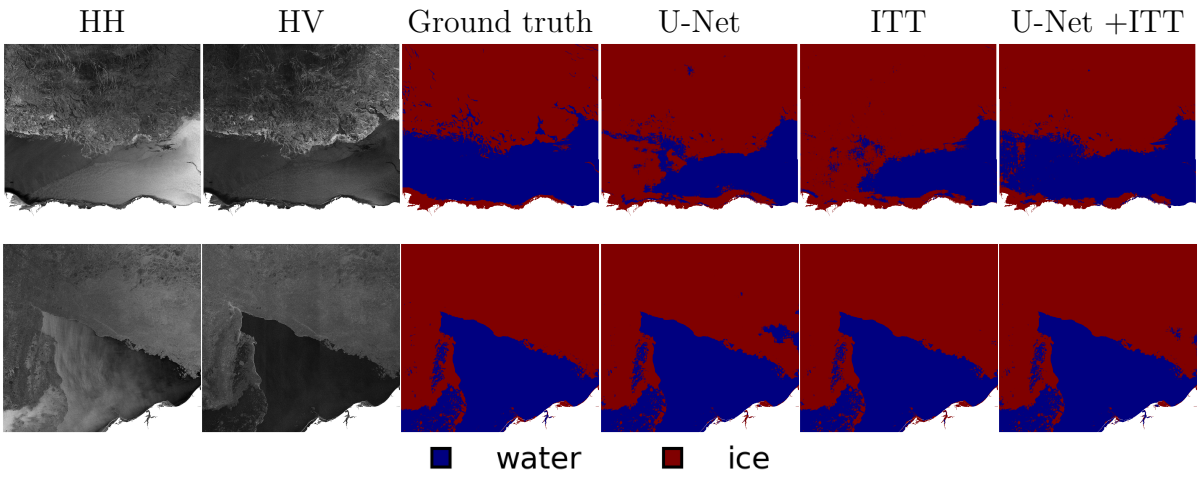

Figure 2.6: Semantic segmentation for scenes 20101017 (top) and 20100623 (bottom).

truth. With a limited receptive field, U-Net also failed to detect some small dark areas of ice at near range, even exhibiting contrasting intensity values compared to open water. In addition, a small area of ice with a dark appearance in the upper center of the scene was misclassified as water. For scene 20100623, U-Net mainly failed on a region of ice that appears homogeneous and without the textural features that are often present on ice. The large context utilized in ITT and U-Net+ITT enabled more accurate classification of the region; although, training from scratch was more effective in this case, which suggest that the initialization via pre-traning U-Net is still sub-optimal.

The most recent approach for ice-water classification using 21 of the scenes used in this chapter was developed by Jiang et al. [40], reporting evaluation metrics based on a sample of pixels per scene. For comparison purposes, we evaluated the correspondent prediction maps considering all pixels and our ground truth. On average, the OA was $94.02 \pm 6.67\%$. Compared with the U-Net+ITT protocol, which scored $96.76 \pm 3.02\%$ using the same scenes, our method is more accurate with a statistically significant difference.

2.7.1 Predictions on boundaries

The analyzed OA from Table 2.2 offers a general idea of the performance of each training protocol; nevertheless, it does not reflect the effect of the ablation components on the improvement of boundary consistency in semantic segmentation. Consequently, we created new metrics to evaluate the classification around boundaries. Starting with the ground

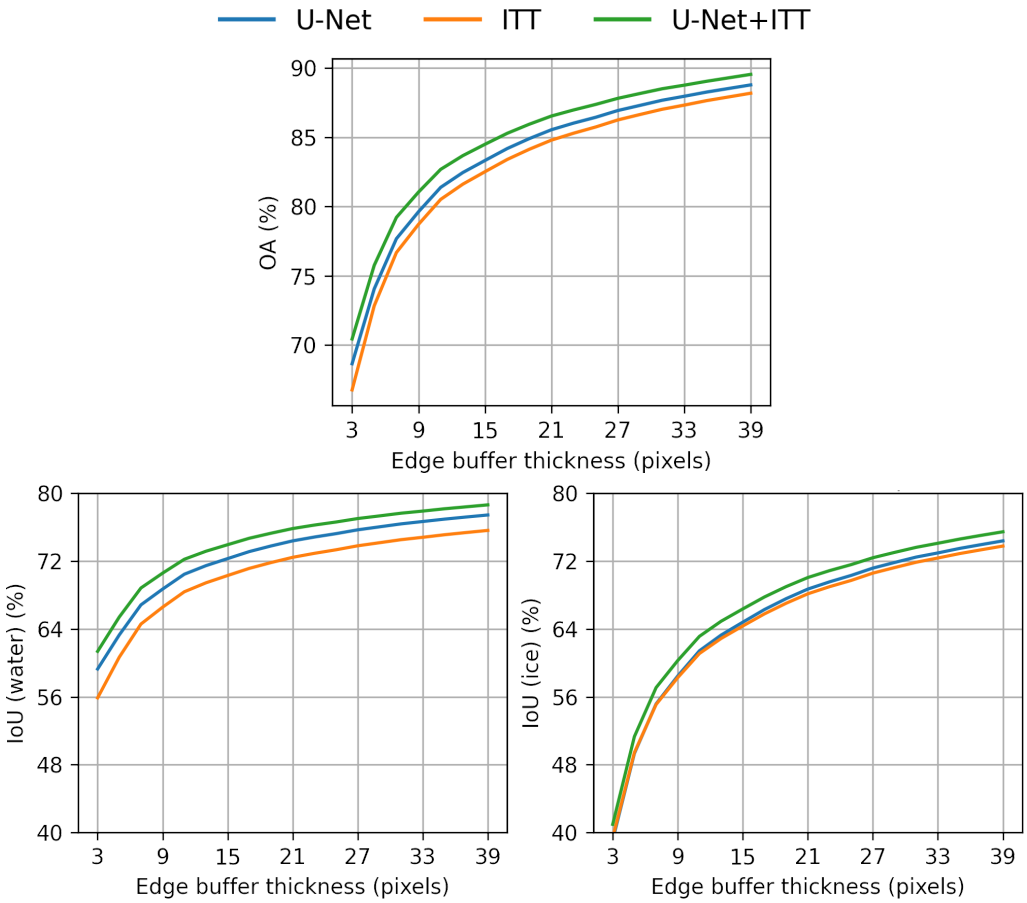

Figure 2.7: Metrics around boundaries for different buffer thicknesses.

truth maps, we extracted the contours to create a mask of one-pixel thick edges. Next, we progressively applied morphological dilation with a disk kernel to create different buffers that demarcate the area around boundaries separating classes. Finally, we computed OA and intersection over union (IoU) per class, only considering the pixels within the buffers.

Fig. 2.7 shows, for each experimental protocol, how the OA and IoU vary as the buffer thickness increases. Within the 1-pixel thick buffer, the intensity transitions in the images are ambiguous and can be either ice or water; therefore, this region has been excluded from the charts. Thickness values are odd numbers because the buffer expands by one pixel on either side of the boundaries. For narrower buffers, the OA is low, but it saturates towards the OA of all pixels as the buffer grows thicker. Consequently, the

relative difference among protocols becomes more noticeable when analyzing boundaries. In this scenario, the predictions from protocol U-Net+ITT were the most accurate, being the training configuration that outperformed U-Net.

The same patterns can be observed with the IoU per class; however, the superiority of U-Net+ITT is more consistent for water across the different buffers. In SAR images, statistics for the class water tend to be less stationary than ice. The incidence angle effect and textural artifacts caused by the wind make water regions look similar to ice, which hinders boundaries in between the classes. The IoU suggests that U-Net+ITT is less sensitive to feature variations for water. Overall, the incorporation of the region-based loss, operating on homogeneous regions, performs a refinement of the CNN feature maps so that the semantic segmentation is more detailed on object boundaries.

2.7.2 Uncertainty analysis

Improving the prediction around boundaries is a direct consequence of reducing uncertainty on semantic segmentation. Since we have improved predictions on boundaries through the success of IRGS unsupervised segmentation, the proposed method seems to provide greater confidence to the model. Fig. 2.8 illustrates the average entropy per scene calculated from the predicted probabilities $\hat{y}$ of the classes ice and water obtained on each model (2.4). The chart contains one group of bars per test scene, and each group indicates the entropy delivered on each experimental protocol. The hatched bars indicate the protocol that scored the highest OA.

$$H = -\sum_{c=1}^{C} \hat{y}_c \log(\hat{y}_c) \tag{2.4}$$

Considering that low entropy implies low uncertainty, we can observe a prominent increase in the metric during the melt season and a more modest increase during the freeze-up season. In the Beaufort Sea, the melt season began in June 2010 and May 2011, and the freeze-up season started in October 2010 and October 2011 [76]. During these periods, non-stationary effects become more prominent, and signatures of ice and water are less separable; therefore, the predictions tend to be more uncertain. Test scenes where the entropy is similarly low, and, accordingly, with higher accuracy for all protocols (see Table 2.2) are usually dated in winter when the landscape is often covered with ice or in summer when mostly water remains.

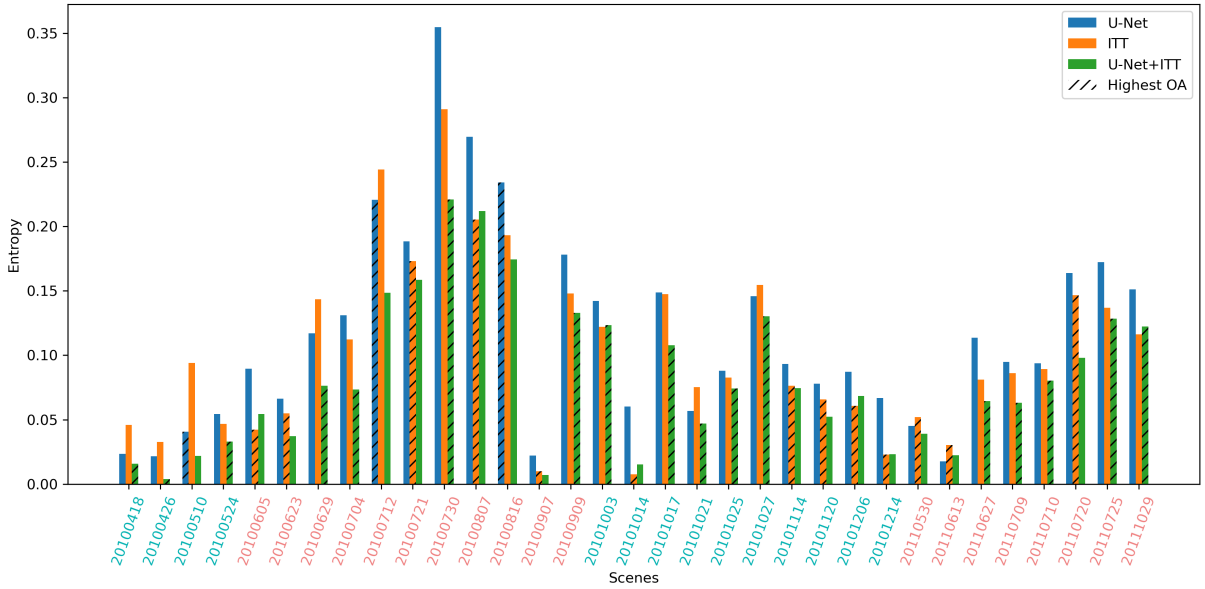

Figure 2.8: Entropy calculated from the predicted probabilities. Similar to Table 2.2, the color coding for the scene dates indicates the freeze-up (light-cyan) and melt (light-red) seasons. The hatched bar indicates the experimental protocol with the higher OA for a test scene. U-Net+ITT delivered the lowest entropy in all but one of the scenes.

As expected, U-Net delivered, in general, the highest entropy, which means the model is strongly affected by the uncertainty derived from the data. By leveraging global context, ITT lowered the entropy with regard to U-Net in the majority of the scenes. Nevertheless, U-Net+ITT produced the lowest entropy in all but one of the scenes when compared to U-Net, with a statistically significant difference. As U-Net scored the highest accuracy just in 3 scenes (see hatched bars), the confidence of the model obtained on U-Net+ITT increases towards more accurate predictions. Fig. 2.9 illustrates this statement by showing heat maps of the predicted probabilities for class ice on the SAR scenes 20100730, 20110627, and 20110720, captured during the melt season. Since the classification is binary, observing the heat maps from either ice or water soft labels provides information about the confidence of the predictions. The heat maps are expected to show, in blue, the probability 0 for water and, in red, the probability 1 for ice.

Scene 20100730 exhibits a wide variability of class statistics and the presence of ice floes at different scales. The classification for U-Net was detailed by detecting small floes; however, the presence of intermediate colors in the heat map, predominantly on ice, in-

dicates the prediction was considerably uncertain. Notice also how small floes located at near range (right) show weak boundaries, caused by the strong backscatter over water that the incidence angle effect creates. Conversely, U-Net+ITT could produce a more detailed and sharper prediction, that, even illustrated using soft-predicted labels, looks closer to the ground truth. In such a complex scene, creating tokens from irregular and multi-scale regions could also make the network more scale invariant.

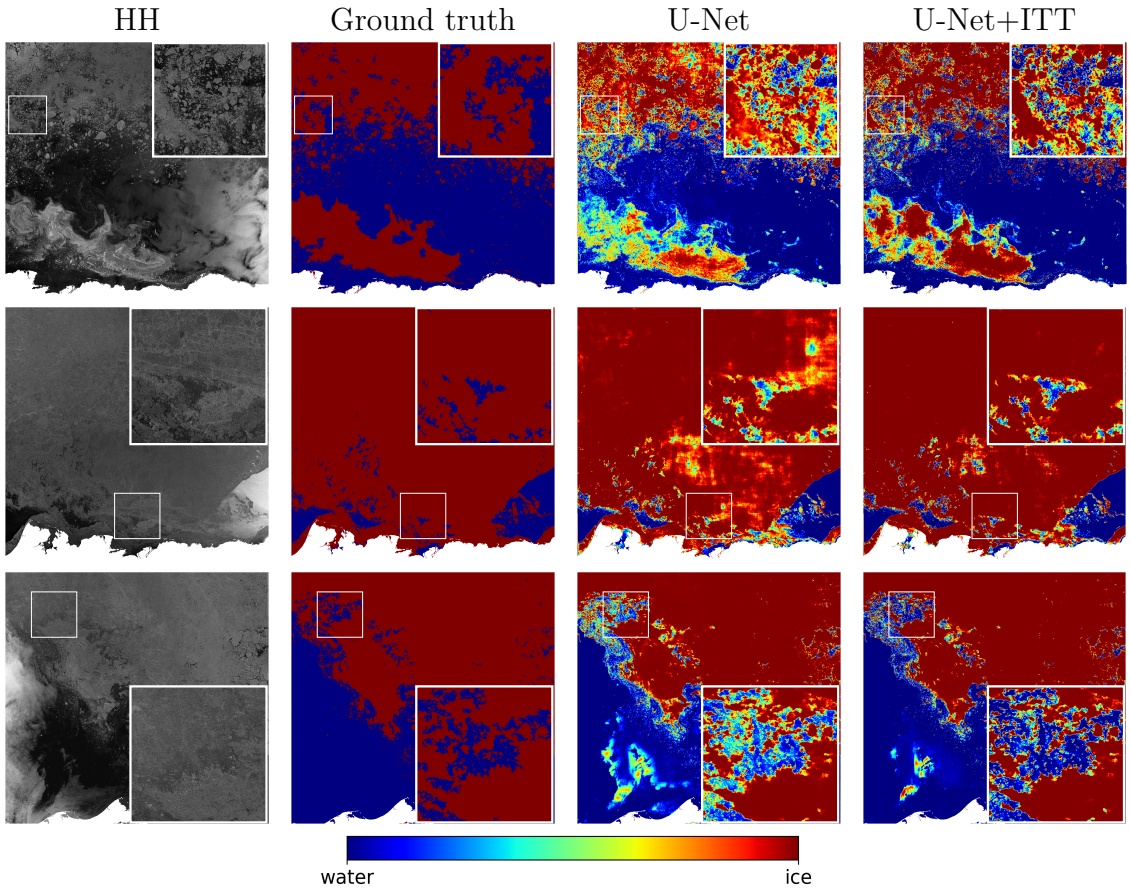

Figure 2.9: Predicted probabilities for class ice on scenes 20100730, 20110627, and 20110720. From left to right we observe: HH and HV images, ground truth, and ice soft-predicted labels for U-Net and U-Net+ITT. Probability 1 (pixels in dark red) are predicted as ice with minimum uncertainty. Probability 0 (pixels in dark blue) are predicted as water with minimum uncertainty. Intermediate colors indicate some level of uncertainty.

Both polarizations on scene 20110627 exhibit spatial non-stationarities and inter-class

mixing. Intermediate probabilities are more present on U-Net, while U-Net+ITT mainly displays extreme probability values that approach the ground truth better, with less uncertainty. The boundary strength in the predictions is consistent with the relative difference between the protocols shown in Fig. 2.7. We can observe on the magnified area of scene 20110720, mainly on the HH polarization, that boundaries are noisy and poorly visible, producing blurry class separation in the heat map of U-Net. In contrast, U-Net+ITT could delineate more accurately the ice-water boundaries. Notably, both protocols were uncertain on the dark area of open water on the bottom-left part of the image, probably confusing the pattern with new ice, which absorbs the SAR sensor's scattered energy even at near range. Since the acquisition date of the scene is in August, it is unlikely to be new ice, which suggests the time information can be relevant to improve the results. From the experiments we can confirm that the multi-task training leveraging the region-based loss $\mathcal{L}_1$ stimulates the pre-trained CNN to produce similar features within homogeneous regions, thus demarcating better between classes. Accordingly, reducing the uncertainty of the model improves the classification around boundaries, even in ambiguous class transition areas.

2.7.3 Region-based semantic segmentation

Although the ITT framework was introduced and evaluated primarily as a pixel-level semantic segmentation model, the proposed architecture also enables an alternative inference mode that produces region-based predictions. This capability arises naturally from the region-based loss employed during training and from the transformer branch operating on tokens derived from homogeneous regions. Region-based predictions are obtained by over-segmenting the full SAR scene using IRGS, producing a large set of irregular and approximately homogeneous regions. Feature maps extracted by the CNN are pooled within each region using average pooling, generating a sequence of region descriptors consistent with the tokenization strategy described in Section 2.4.3. Since the number of regions can reach several thousands for complex scenes, the resulting sequence is partitioned into fixed-length chunks to satisfy the transformer's computational constraints. Inference is then performed independently on these sub-sequences, and the predicted class labels are reassembled according to the regions from which the tokens were generated, yielding a region-based semantic segmentation map.

From this point onward, prediction heads are explicitly identified using suffixes to distinguish the inference modality associated with the ITT and U-Net+ITT training protocols. Pixel-level semantic segmentation results are reported as ITT_pixel and U-Net+ITT_pixel, while region-based semantic segmentation results are reported as ITT_region and U-Net+

Table 2.3: Overall accuracy (OA, %) for pixel-based and region-based semantic segmentation obtained using the ITT and U-Net+ITT training protocols.

Experimental Protocol	**Pixel-based**	**Region-based**
ITT	96.28±3.87	95.99±4.44
U-Net+ITT	**96.85±2.64**	96.52±3.49

ITT_region. Table 2.3 summarizes the overall accuracy (OA) obtained for each configuration. A comparison of the results indicates that pixel-based predictions outperform their region-based counterparts. This behavior is expected, as pixel-level inference is independent of the quality of the unsupervised over-segmentation and preserves fine-grained spatial variations that may be suppressed when enforcing region-wise label consistency. In contrast, region-based predictions depend on the assumption that each segmented region is dominated by a single semantic class, an assumption that may be violated on large scenes exhibiting strong spatial non-stationarity or complex textural transitions. At inference time, this makes region-based predictions explicitly sensitive to the quality of the unsupervised over-segmentation. Although IRGS is effective at delineating natural boundaries in SAR imagery, it does not guarantee purely homogeneous regions when applied to full scenes. As a result, mis-segmented or mixed regions can propagate errors to the final prediction by assigning an incorrect class uniformly to all pixels within a region.

For region-based predictions, pretraining the CNN remains beneficial. Initializing the CNN with weights learned from the pixel-only baseline leads to more discriminative region descriptors and improves the stability of transformer-based classification, as reflected by the higher OA achieved by U-Net+ITT_region compared to ITT_region. This behavior is consistent with the findings observed for pixel-based inference, reinforcing the role of CNN pretraining in shaping region-consistent feature representations.

An inherent characteristic of region-based semantic segmentation is that predicted class boundaries are strictly aligned with those produced by IRGS. This alignment implies a structural limitation: errors in region-based predictions appear as region-level misclassifications, rather than as spatially diffuse or noisy artifacts, revealing a trade-off between boundary sharpness and sensitivity to segmentation quality. Figure 2.10 illustrates this behavior on scene 20101017. In the pixel-based prediction (U-Net+ITT_pixel), certain open water areas are misclassified, likely due to the fine-grained nature of pixel-level inference, which makes it more sensitive to local noise and textural variability. In contrast, U-Net+ITT_region correctly labels these areas because they belong to a large open water region consistently delineated by IRGS. However, new ice areas exhibiting a dark backscat-

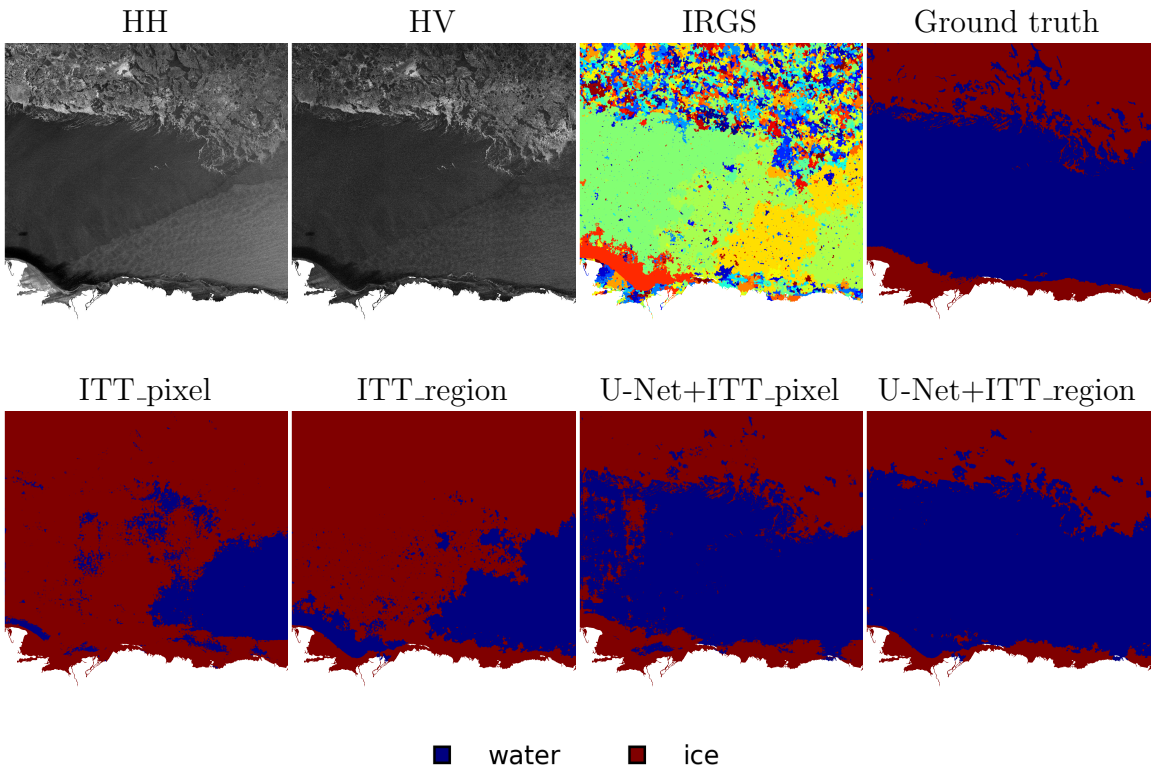

Figure 2.10: Pixel- and region-based semantic segmentation comparison for image patch on scene 20101017.

ter signature near the shoreline are entirely misclassified in the region-based prediction, as the corresponding regions are incorrectly segmented.

Hybrid segmentation

The complementary behavior observed between pixel-based and region-based inference motivates the evaluation of a hybrid strategy that combines the predicted class probabilities from both heads. In this experiment, we considerd only the U-Net+ITT protocol, which delivered the highest accuracy for both pixel-based and region-based semantic segmentation.

Let $\hat{\mathbf{y}}^{\text{pixel}}$ and $\hat{\mathbf{y}}^{\text{region}}$ denote the soft predictions (class probabilities) produced by the pixel and region heads, respectively. A convex combination of the two probability maps is

defined as

$$\hat{\mathbf{y}}^{\text{hybrid}} = \alpha\, \hat{\mathbf{y}}^{\text{pixel}} + (1-\alpha)\, \hat{\mathbf{y}}^{\text{region}}, \tag{2.5}$$

where $\alpha \in [0, 1]$ controls the relative contribution of each head. Since α and $(1-\alpha)$ are normalized weights, the resulting probabilities remain bounded between 0 and 1. The final semantic segmentation map is obtained by assigning to each pixel the class with maximum combined probability. We evaluate eleven configurations by varying α from 0 (pure region-based prediction) to 1 (pure pixel-based prediction) in increments of 0.1. For each configuration, we compute the overall accuracy (OA) over the entire scene and OA within boundary buffers of increasing thickness, following the procedure described in Section 2.7.1. Metrics are averaged across the 33 leave-one-out models, and the mean and standard deviation are reported. Figure 2.11 summarizes the results.

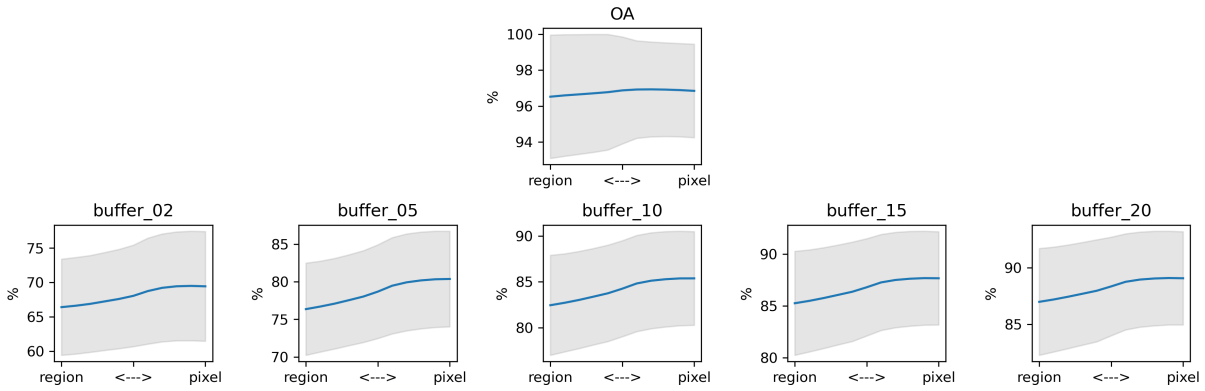

Figure 2.11: Mean and standard deviation of overall accuracy (OA) and boundary-buffer OA obtained by linearly combining pixel-based and region-based predictions from U-Net+ITT. Results are averaged over 33 leave-one-out models.

For both global OA and boundary evaluations, performance improves consistently as α shifts toward the pixel-based head. The mean accuracy increases monotonically with larger contributions of U-Net+ITT_pixel, while the corresponding standard deviation decreases, indicating greater stability across scenes. This trend is observed not only for full-scene segmentation but also within boundary buffers of increasing thickness, where the highest mean OA is achieved when the pixel-based prediction dominates. Collectively, these results demonstrate that the pixel-based head provides superior accuracy and robustness, confirming its dominant role in the overall semantic segmentation performance of U-Net+ITT.

2.8 Audit conducted by the CIS

The outcomes of the ITT framework motivated a subsequent technology transfer initiative, in collaboration with the CIS, aimed at integrating region-consistent ice-water classification into an operational human-in-the-loop workflow. A detailed description of this deployment and active-learning pipeline is provided in Appendix A.

In order to evaluate the operational capability of the proposed approach, CIS analysts audited the prediction maps produced by the U-Net+ITT_pixel and U-Net+ITT_region models through visual inspection. Each prediction map was assessed and ranked according to three qualitative categories: *high*, *sufficient*, and *not sufficient*, reflecting its suitability for operational use as detailed ice-water maps. Two quality criteria were considered: (i) the correctness of homogeneous regions, defined as contiguous ice-only or water-only areas in the SAR image that may be entirely misclassified, and (ii) the accuracy of boundary delineation for regions that follow natural ice-water interfaces. Figure 2.12 summarizes the audit results by counting the number of SAR scenes assigned to each operational category for the two prediction strategies and quality criteria. The grouped histograms report these counts separately for homogeneous-region correctness and boundary delineation. To account for seasonal variability, freeze-up and melt-season scenes are stacked within each bar, with dashed separators indicating the contribution of each season.

The histogram highlights systematic differences between prediction strategies across both quality criteria and seasons. For the evaluation of homogeneous regions, U-Net+ITT_pixel consistently attains a higher number of scenes ranked as *high* or *sufficient*, with this advantage being most pronounced during freeze-up conditions. This indicates more reliable large scale semantic consistency when predictions are derived directly at the pixel level, without enforcing region-level homogeneity. In contrast, U-Net+ITT_region exhibits a larger proportion of scenes ranked as *not sufficient*, particularly during the melt season, suggesting increased sensitivity to fragmented and heterogeneous ice conditions. Similar patterns are observed for boundary delineation, where pixel-based predictions receive more favorable rankings overall, while region-based predictions are more frequently downgraded during melt. Taken together, these results indicate that the operational performance of region-based aggregation is strongly influenced by seasonal ice morphology, whereas pixel-based predictions exhibit more stable behavior across seasonal transitions.

Qualitative Assessment

Ice analysts also performed a qualitative assessment of the predicted ice-water segmentation maps by visually inspecting each scene and recording text notes describing the main

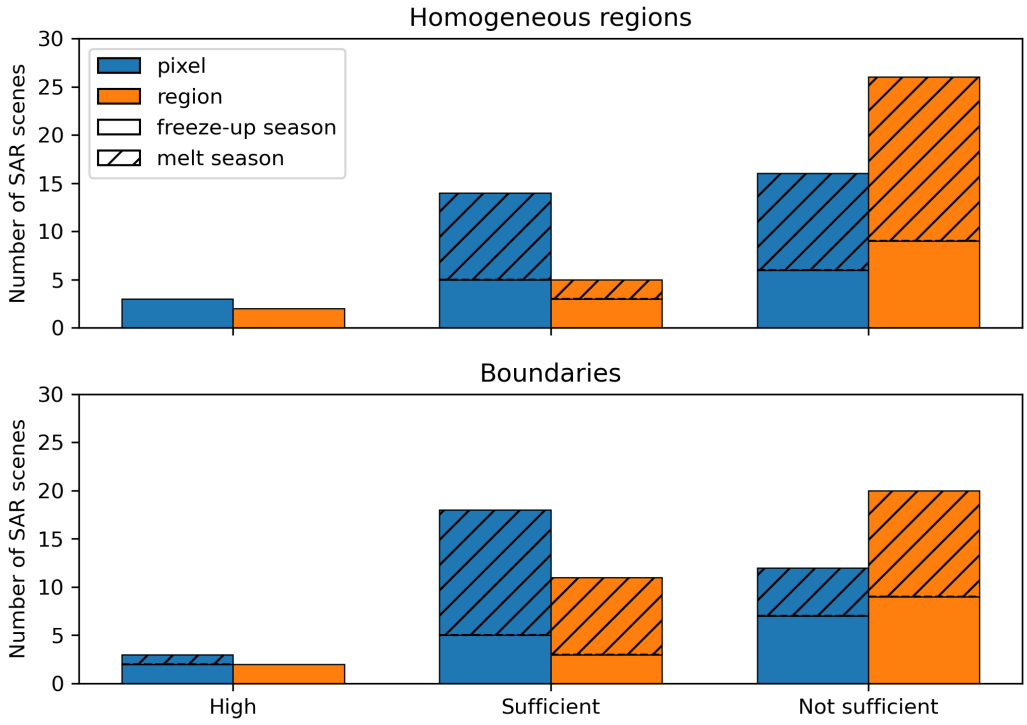

Figure 2.12: CIS expert audit of operational capability for U-Net+ITT pixel-based and region-based predictions. Bars show the number of SAR scenes ranked as high, sufficient, or not sufficient for homogeneous regions (top) and boundary delineation (bottom), showing the seasonal breakdown within each bar.

deficiencies observed in the predictions. These notes provide a high-level, expert-driven evaluation of failure mechanisms that are difficult to capture using pixel-wise metrics alone, such as missed small floes, incomplete representation of narrow ice features, or degraded boundary delineation. As such, the outcome of this evaluation is inherently descriptive and subjective, consisting of unstructured textual observations rather than categorical or numerical scores.

To enable systematic analysis of these qualitative assessments, we interpret the expert notes through a structured error taxonomy that groups recurring descriptions into a finite set of well-defined failure modes. The taxonomy is designed to reflect distinct error mechanisms relevant to binary ice water segmentation, rather than directional confusion (e.g., ice versus water), and allows multiple error types to be associated with a single prediction map. This step transforms the descriptive expert feedback into a consistent multi-label

representation that can be analyzed statistically while preserving the interpretability of the original observations. Only prediction maps for which analysts provided explicit notes are included in this analysis, ensuring that error frequencies are computed conditionally on active qualitative evaluation.

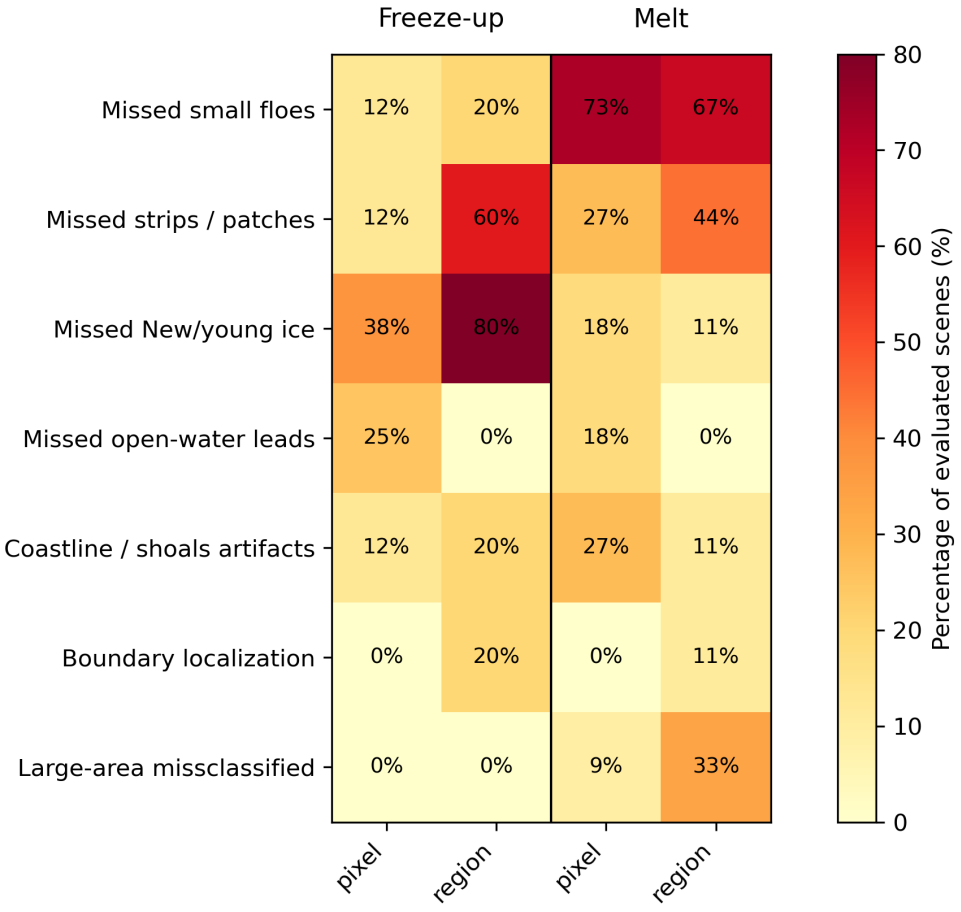

Figure 2.13: Seasonal error/frequency heatmap derived from qualitative expert evaluation of ice-water segmentation predictions. Values indicate the percentage of annotated prediction maps in which each error type is observed for a given season-model combination. Percentages are computed per condition, and multiple error types may co-occur within a single prediction.

Figure 2.13 visualizes the resulting seasonal error/frequency analysis. For each combination of season s (freeze-up and melt) and prediction strategy m (U-Net+ITT_pixel and

U-Net+ITT_region), we compute the frequency with which each error type e (each row in the heatmap) is observed across the set of annotated prediction maps. Specifically, the reported percentage is defined as:

$$f(e|s,m) = \frac{1}{N_{s,m}} \sum_{i=1}^{N_{s,m}} \mathbf{1}(e \in \mathcal{E}_i) \times 100 \tag{2.6}$$

where $N_{s,m}$ is the number of prediction maps for season s and prediction strategy m, $\mathcal{E}_i$ denotes the set of error types assigned to the i-th prediction, and $\mathbf{1}(\cdot)$ is an indicator function. Because multiple error types may co-occur within a single prediction, the percentages do not sum to 100% and are interpreted as frequencies of occurrence rather than mutually exclusive probabilities. We further note that zero-valued percentages indicate that the corresponding error type was not observed in the predictions for the given season-model combination, rather than implying its theoretical absence.

The analysis is explicitly separated into freeze-up and melt seasons, reflecting the fundamentally different ice morphologies and SAR backscatter regimes encountered during these periods. Freeze-up scenes are generally characterized by expanding, spatially coherent ice cover with comparatively stable boundaries, whereas melt-season scenes exhibit highly fragmented ice fields, narrow leads, strips[2], and patches[3] embedded within open water. Aggregating these regimes would mask season specific failure patterns and reduce the operational relevance of the analysis.

Overall, the seasonal error-frequency results provide insight into the interaction between model architecture and prediction strategy within the U-Net+ITT framework. The elevated error frequencies observed for region-based predictions during melt conditions are consistent with the sensitivity of this approach to the unsupervised IRGS segmentation. On large and heterogeneous SAR scenes, IRGS may produce spatially extensive regions that span multiple ice-water regimes, which are subsequently assigned a single semantic label during region-based inference. This effect is most pronounced during melt, when fragmented floes, narrow strips, and mixed backscatter patterns dominate, leading to both fine-scale omissions and occasional large area misclassification. In contrast, pixel-based predictions are not conditioned on a pre-segmentation of the scene and rely directly on image evidence, resulting in a more gradual degradation that remains largely confined to

[2]Narrow, elongated ice features embedded within open water or along the ice edge.

[3]In the sea ice jargon, patches are small, compact ice features completely surrounded by open water. They arise from fragmented or remnant ice and appear in SAR imagery as isolated regions of ice-like backscatter.

localized structures. These findings complement the quantitative results reported in Section 2.7.3, highlighting the dependent nature of region-based aggregation and the critical role of scene heterogeneity when coupling long-range representations with unsupervised region partitioning during inference.

Importantly, the expert audit further indicates that on the order of half of the pixel-based prediction maps meet CIS operational acceptability thresholds, being ranked as either *high* or *sufficient* across the evaluated criteria. This suggests that the U-Net+ITT pixel-level approach achieves a non trivial level of operational viability without additional post-processing. At the same time, the qualitative feedback highlights recurring, season-dependent deficiencies, particularly under melt conditions, that are not fully reflected by quantitative metrics alone. These observations emphasize the value of expert assessment for identifying targeted avenues for refinement as the ITT method progresses toward operational use.

2.9 Conclusion

This chapter has presented a novel method for sea ice classification that integrates CNNs, transformers, and unsupervised segmentation. The approach not only leverages the operational characteristics of each component, but also provides a topology to handle spatial non-stationary statistics in SAR images by capturing global information effectively. Furthermore, we are pioneering in the use of self-attention on tokens created from homogeneous regions of irregular shape and scale. From the experiments, two aspects were valuable to improve the chosen CNN baseline: pre-training the CNN module and training, afterwards, the whole method in a multi-task fashion combining pixel and region based loss terms. This configuration achieved the highest average overall accuracy and the most stable performance across scenes.

The resulting semantic segmentation is less sensitive to non-stationarities and exhibits sharper boundary delineation with reduced predictive uncertainty through region-consistent feature learning. Beyond pixel-level segmentation, we evaluated region-based inference and hybrid combinations of pixel- and region-level predictions, revealing a trade-off between boundary precision and sensitivity to over-segmentation quality, with pixel-based inference achieving the most stable and accurate performance across scenes. The operational viability of the framework was further assessed through an expert-driven audit conducted by analysts from the CIS, providing qualitative insight into segmentation quality and confirming the practical potential of the proposed approach.

This research represents an initiative to classify sea ice at a regional scale using deep learning, bringing the approach closer to the procedures currently employed by sea ice analysts for creating ice charts. Consequently, this work opens multiple avenues for future research. For instance, the robustness of irregular tokens to object scale variability will be further investigated. In addition, the usage of the positional encoding in transformers can be favorable. We hypothesize that the incorporation of region descriptors, such as geographical location, acquisition time or season, and geometric properties can further improve the classification. Finally, we will explore stochastic region-level inference strategies that enable regions to participate in multiple contextual sequences to produce complementary predictions without retraining the model.

Chapter 3

Is Pre-training Enough? SAR-IceFM: Towards Multi-Task Foundation Models for Sea Ice Classification

In the previous chapter, representation learning for sea ice classification was addressed through supervised modeling of spatial non-stationarity in synthetic aperture radar (SAR) imagery. Nevertheless, supervised strategies remain inherently limited by the availability and geometry of labeled data, particularly in Arctic environments where annotations are coarse, sparse, and operationally constrained. As satellite archives continue to expand while high-quality labels remain scarce, the central challenge shifts from designing better supervised architectures to learning more transferable and scalable feature representations. In the context of SAR, characterized by complex backscatter behavior, spatial non-stationarity, and polygon-based supervision, this motivates a fundamental question: can large-scale self-supervised pre-training yield representations that generalize across tasks, resolutions, and supervision regimes for Arctic sea ice analysis?

In this chapter, we introduce SAR-IceFM, an open-source[1] SAR-based foundation model trained on Sentinel-1 imagery for multi-task Arctic sea ice analysis. We systematically investigate representation learning across varying data and model scaling regimes, comparing pre-training against training from scratch and analyzing behavior under single-task and multi-task fine-tuning. In addition, we examine resolution transferability under single- and multi-scale training strategies on downstream generalization. Through this structured analysis, the chapter explores how representation quality interacts with super-

[1] https://github.com/jnoat92/SAR-IceFM

vision geometry, task heterogeneity, and resolution handling in SAR-based sea ice classification.

3.1 Introduction

Sea ice is a key component of the Earth's climate system. Its dynamics have far-reaching environmental, societal, and economic impacts. It regulates heat exchange between the ocean and atmosphere, drives ocean circulation, and sustains fragile Arctic ecosystems that support biodiversity and traditional ways of life [34]. For northern communities, safe and predictable sea ice conditions are essential for transportation, hunting, and food security. On a global scale, accurate sea ice information also benefits maritime industries, as diminishing ice cover is opening new shipping routes and expanding industrial activities in polar regions [35]. Meeting these diverse needs requires moving beyond manual charting, which cannot keep pace with the increasing volume of satellite observations. Automated sea ice mapping is therefore necessary to generate reliable descriptors through predictions that are both accurate and timely, ensuring that communities, policymakers, and industries have access to actionable information when it is most needed.

National ice monitoring agencies, such as the Canadian Ice Service (CIS), use Earth-orbiting satellites as the primary tool for monitoring sea ice. Through visual interpretation, a process requiring considerable training and experience, the agencies provide sea ice information to support marine transportation and safety. The different products and services include ice charts, forecasts, and onsite guidelines [37]. Ice charts are manually produced to depict complete or partial concentration of ice conditions at a particular time, summarizing key descriptors such as sea ice concentration (SIC), stage of development (SOD), and floe size (FLOE) using the standardized egg code [31]. As they are created through expert interpretation of satellite observations, ice charts represent the closest practical approximation to ground-truth annotations.

SAR is the primary data source for monitoring sea ice, as it operates independently of weather and illumination conditions. Its backscatter response is influenced not only by surface roughness but also by the dielectric properties of different ice types, and its penetration capability provides additional structural information. These characteristics make SAR imagery particularly well suited for sea ice monitoring. However, interpretation is not straightforward: ambiguities arise from seasonal and regional variability in sea ice properties, and further challenges stem from inherent SAR imaging artifacts, as shown in our previous work [30].

Recent studies [4, 77] highlight the growing shift toward deep learning models for automating sea ice classification from SAR imagery. In collaboration with the Canadian Ice Service, our team's work progressed from data curation and model design to the conclusion that self-supervised learning is necessary for robust SAR-based sea ice analysis [2, 30, 31, 78]. Nonetheless, developing models with true operational capability remains challenging, largely because supervised training or fine-tuning self-supervised methods for specific tasks typically requires substantial amounts of annotated data, which are scarce in the sea ice domain. While this limitation extends across remote sensing applications, some works have begun to incorporate self-supervised learning (SSL) to exploit the vast and continuously expanding collections of unlabeled SAR images. Models pre-trained in this way can be adapted to specific downstream tasks through fine-tuning, and when designed to enable broad transfer across applications, they are often referred to as foundation models [79, 80]. Such models unlock scalability by reducing reliance on large labeled datasets and reduce the need for domain-specific model designs.

Cloud-based data catalogs have facilitated access to open-source satellite datasets [81]. In computer vision, recent foundation models such as CLIP [82], DINOv2 [83], and Segment Anything (SAM) [84] have demonstrated strong transferability, enabling zero-shot or prompt-based adaptation across a wide range of tasks. Nonetheless, most existing foundation models in remote sensing are trained primarily with multispectral imagery and have been shown to significantly improve downstream task performances [85]. There is an extensive body of work leveraging existing foundational models, or creating and fine-tuning models for remote sensing [86–88]. Despite these advances, current remote sensing foundation models often do not incorporate SAR data or lack sufficient coverage of Arctic regions during training. Although Earth-scale multi-satellite models, such as Clay and AlphaEarth Foundations, demonstrate strong generalization across a wide range of tasks [89, 90], a significant domain gap remains between land-surface backscatter and sea ice backscatter. In particular, SAR imagery exhibits non-stationary statistics, speckle noise, and acquisition-dependent variability that are not well represented in existing training corpora. Consequently, the applicability of these models to SAR-based sea ice analysis remains limited, as they are predominantly trained on passive sensors and land-use data.

The literature contains only a limited number of works that explore self-supervised learning (SSL) for sea ice classification. Seston et al. [91] integrate Bootstrap Your Own Latent (BYOL) [92] with U-Net [20] and apply it to Sentinel-1 SAR imagery from the Canadian Arctic Archipelago for ice–water segmentation. Their approach reduces labeling requirements by approximately 50% compared to models trained without SSL pre-training. Similarly, Qin et al. [93] demonstrate significant improvements in segmenting sea ice SOD by pre-training a Vision Transformer (ViT) [23] backbone using a Masked Autoencoder

(MAE) [29]. Similarly, Alkaee et al. [94] establish benchmarks for SOD classification by exploring fine-tuning strategies based on existing remote sensing foundation models. Overall, current SSL approaches for sea ice remain limited to single-task settings.

Beyond sea ice, several SSL approaches in the SAR domain have emerged as promising candidates for developing foundation models. Wang et al. [95] introduced the Feature-Guided MAE-SAR (FGMAE-SAR), a feature-aware SSL framework that reconstructs hand-crafted SAR features rather than raw pixels. Pre-training FGMAE-SAR on medium-resolution imagery has been shown to improve downstream task performance and is particularly effective in noisy SAR contexts. In another large-scale effort, Allen et al. [96] applied MAE to SAR amplitude data covering 8.7% of the Earth's land surface, and fine-tuned the pre-trained weights on land-use tasks, substantially reducing labeling needs while generalizing to unseen geographical regions. Weijie et al. [97] introduced the SAR-JEPA framework using SSL to learn transferable representations, targeting land-use downstream tasks.

Multi-modal SSL has also been investigated for SAR by pairing it with optical imagery. Wang et al. [98] adapted the self-distillation with no labels (DINO) method [99] for joint SAR–optical pre-training, enabling representations that transfer across modalities. In parallel, modular architectural designs aimed at improving multi-scale robustness such as Patch Incomplete Mask (PIMask) [100] and Rotated Varied-Size Attention (RVSA) [101], originally proposed for other remote sensing tasks, highlight further opportunities for extracting meaningful representations of sea ice.

In this chapter, we explore the development of foundation models using MAE trained on SAR imagery from the Arctic. Specifically, we introduce SAR-IceFM, a SAR-based foundation model for sea ice classification, developed in parallel with a sea ice analysis tool and intended to operate under practical data availability, supervision, and resolution constraints. To the best of our knowledge, this is the first work to evaluate the effectiveness of self-supervised pre-training on multiple downstream sea ice tasks simultaneously. Through ablation studies, we further demonstrate that fine-tuned models substantially outperform training from scratch. Our contributions can be outlined as follows:

1. We introduce SAR-IceFM, a foundation model for SAR sea ice analysis, and investigate how pre-training effectiveness varies with data regime, model capacity, and the pre-training/fine-tuning balance, thereby clarifying the conditions under which pre-training leads to measurable improvements.

2. We conduct a empirical assessment of single-task and multi-task fine-tuning for SIC, SOD, and FLOE, benchmarking task-specific performance and revealing task-

dependent trade-offs and limitations of shared encoders under coarse, polygon-based supervision.

3. We evaluate resolution transferability under single- and multi-scale training, empirically demonstrating how supervision geometry, training consistency, and inference resolution influence generalization across spatial scales, and clarifying when multi-scale training improves robustness and when it introduces instability.

The remainder of this chapter is structured as follows. Section 3.2 describes the AI4Arctic dataset and how it is used to train the foundation model. Section 3.3 presents the proposed MAE training protocol for sea ice classification using SAR imagery. Section 3.4 details the fine-tuning process, while Section 3.5 provides information about the architecture, training settings, and evaluation of the foundation model. Section 3.6 presents the results along with ablation studies, and Section 3.7 discusses the challenges and opportunities for building a foundation model for sea ice. Finally, Section 3.8 concludes the chapter with a summary of key findings.

3.2 AI4Arctic Dataset

The AI4Arctic dataset [102] combines high-resolution σ^o calibrated SAR imagery with lower-resolution passive microwave measurements and numerical weather prediction data. The SAR component is derived from the Sentinel-1 mission, using Extra Wide Swath (EW) Level-1 Ground Range Detected (GRDM) products at medium resolution and dual polarization (HH and HV). Complementary data sources include 14 brightness temperature observations from the Advanced Microwave Scanning Radiometer (AMSR2) and 6 reanalysis-based weather parameters such as winds, temperature, water vapor, and cloud liquid water. In addition, a distance-to-land mask is provided to mitigate coastal effects that can influence coarse-resolution sensors.

As illustrated in Figure 3.1, the dataset covers waters around Greenland and the Canadian Arctic. The temporal span extends from 2018 to 2021, with seasonal distribution reflecting the availability of operational ice charts, resulting in denser coverage during periods critical for navigation. Each Sentinel-1 scene covers approximately 400×400 km^2, corresponding to roughly $10{,}000 \times 10{,}000$ pixels at 40 m spacing. The dataset includes 513 training scenes and 20 test scenes, a configuration originally adopted for the AutoIce challenge [4].

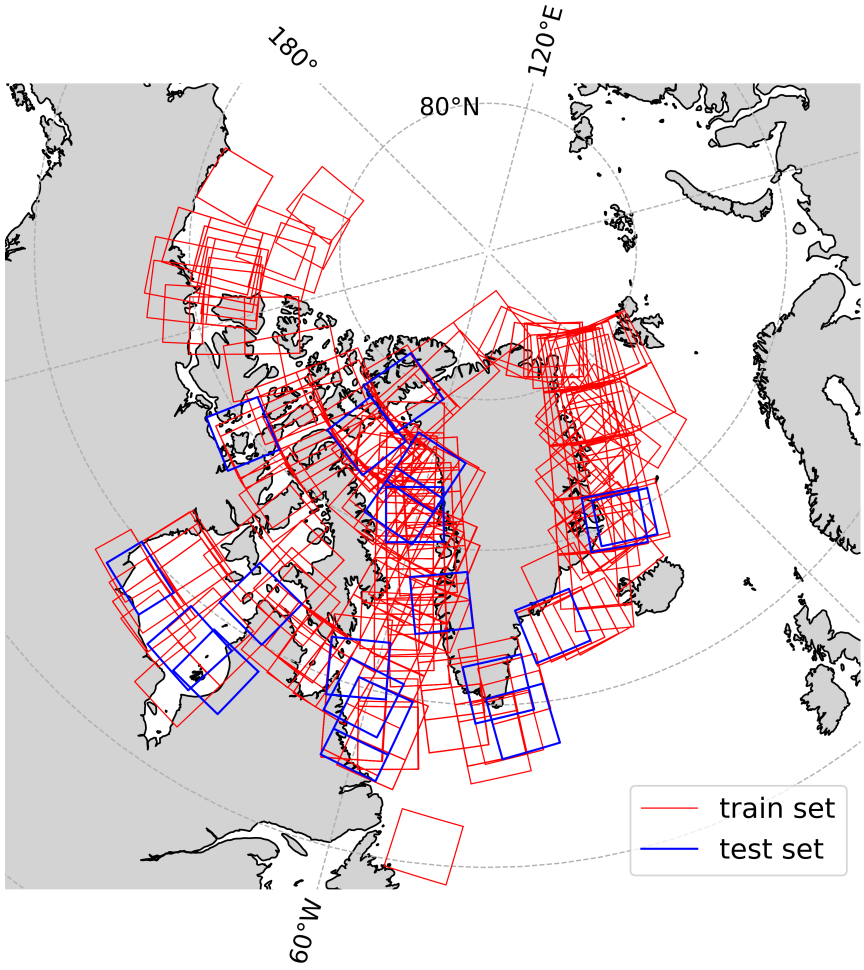

Figure 3.1: Study area of the AI4Arctic dataset. Scenes designated for training are outlined in red, while test scenes are outlined in blue. Those scenes that share the same satellite footprint and cannot be distinguished from each other correspond to different acquisition times.

Annotation labels are derived from ice charts produced by trained analysts at the CIS and the Danish Meteorological Institute (DMI) Greenland Ice Service. Each ice chart consists of manually drawn polygons of relatively homogeneous conditions, with descriptors encoded in the World Meteorological Organization (WMO) egg code. The ice charts are aligned to the Sentinel-1 grid and converted into discrete classes for each descriptor: SIC (11 classes), SOD (5 classes), and FLOE (6 classes).

While the dataset is valuable, it is important to highlight that ice charts are not strict ground truth, as they rely on subjective interpretation and are constrained by time and operational demands. Furthermore, the polygon-based approach assumes homogeneity within polygons, but local variability on SAR backscatter patterns can be substantial (see labels on Figure 3.2). The coarseness of these approximations causes accuracy to vary spatially, with potential mislabeling and discrepancies across national services [103]. Although label noise can be a limitation, it is also a realistic challenge as ice-chart based labels are products with operational relevance. AI4Arctic remains the most extensive open benchmark dataset with a rich semantic structure for Arctic sea ice. With large scale, diversity, and public availability, the dataset can drive advances in representation learning.

3.3 MAE for sea ice

Growing empirical evidence [95, 100] indicates that reconstructive self-supervised methods, such as MAE, are effective for building foundation models in remote sensing. Unlike contrastive SSL, which relies on carefully defined positive and negative samples, reconstruction directly exploits the inherent spatial redundancy of remote sensing data. This distinction is particularly relevant for SAR-based sea ice analysis, where visually similar regions may correspond to different physical ice conditions, making the definition of negative samples ambiguous. In contrast, reconstruction encourages the model to learn representations that capture the underlying spatial structure of the data without relying on semantic assumptions.

MAE training is illustrated in Figure 3.2 (pre-training stage). Formally, let an input image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ be partitioned into a sequence of N non-overlapping patches $\{\mathbf{x}_i \in \mathbb{R}^{P \times P \times C}\}_{i=1}^{N}$. These patches are linearly projected into a sequence of patch embeddings $\{\mathbf{z}_i \in \mathbb{R}^{D}\}_{i=1}^{N}$, forming the input tokens for the transformer. A random subset of patches is then masked, such that only a fraction $\rho \in (0, 1)$ of patches is retained as visible tokens, while the remaining $(1 - \rho)$ patches are replaced by mask tokens. Let $\mathcal{V}$ and $\mathcal{M}$ denote the sets of visible and masked indices, respectively.

The MAE encoder $f_\theta(\cdot)$ operates only on the visible tokens:

$$\mathbf{h}_\mathcal{V} = f_\theta\left(\{\mathbf{z}_i \mid i \in \mathcal{V}\}\right) \tag{3.1}$$

where $\mathbf{h}_\mathcal{V}$ denotes the latent representations of the visible patches. By restricting the encoder to a subset of tokens, MAE significantly reduces computational complexity while forcing the model to learn global context from partial observations.

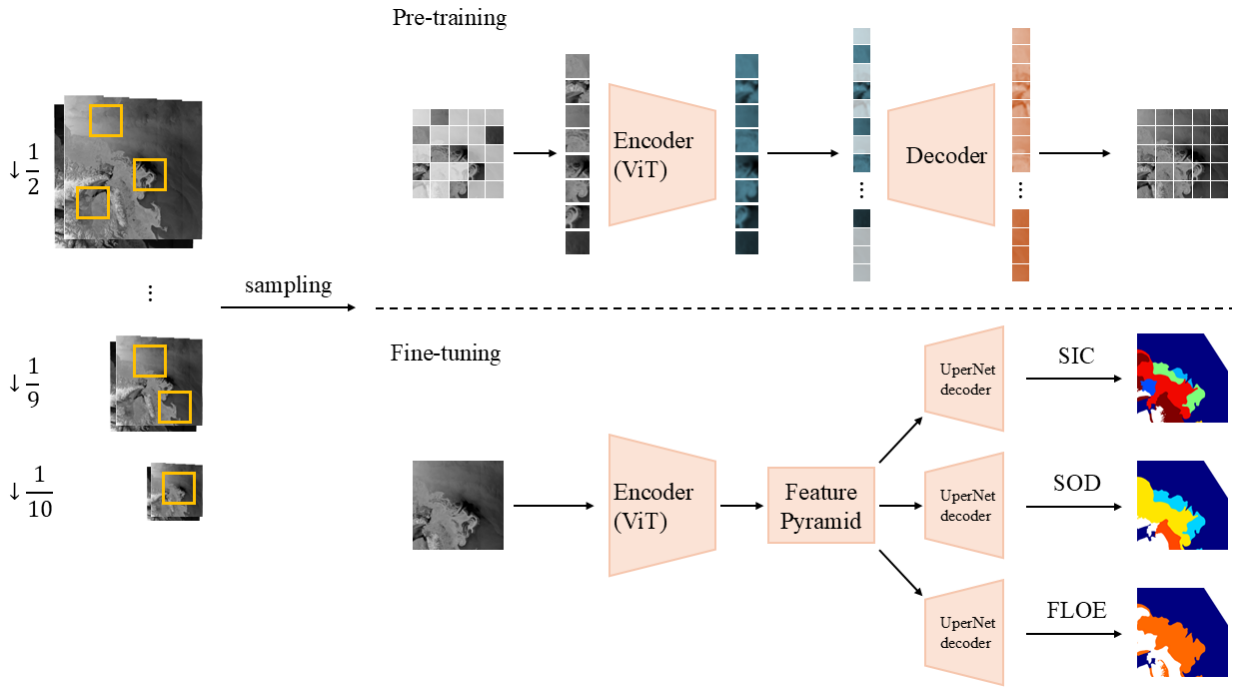

Figure 3.2: MAE pre-training and multi-task fine-tuning for the sea ice foundation model. Multiple scales for SAR inputs are used to capture global context and local detail.

The decoder $g_\phi(\cdot)$ receives both the encoded visible tokens and a set of learned mask tokens $\{\mathbf{m}_i\}_{i\in\mathcal{M}}$, reconstructing the full set of patches:

$$\{\hat{\mathbf{x}}_i\}_{i=1}^{N} = g_\phi\left(\mathbf{h}_\mathcal{V}, \{\mathbf{m}_i\}_{i\in\mathcal{M}}\right) \tag{3.2}$$

where $\hat{\mathbf{x}}$ represents the reconstructed image. The model is trained by minimizing a reconstruction loss computed only over the masked patches:

$$\mathcal{L}_{\text{MAE}} = \frac{1}{|\mathcal{M}|} \sum_{i\in\mathcal{M}} \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2 \tag{3.3}$$

This objective encourages the model to infer missing content from surrounding context, thereby learning representations that capture both local structure and long-range spatial dependencies. Importantly, the masking ratio is typically high (e.g., 75% or more), which increases the difficulty of the reconstruction task and promotes the learning of more informative latent features.

From a representation learning perspective, MAE can be interpreted as learning a latent function f_θ that encodes the conditional distribution of missing patches given visible context. This interpretation emphasizes that the encoder does not explicitly reconstruct the missing content, but instead learns a representation sufficient to characterize the conditional distribution of masked regions given the observed context. Such a formulation is particularly well suited to SAR imagery, where spatial dependencies arise from both physical scattering mechanisms and acquisition geometry and must be inferred from partial observations. In the context of sea ice analysis, this reconstruction-based learning is advantageous as it does not rely on explicit semantic labels, but instead leverages the structural regularities of SAR data. By reconstructing masked regions, the model captures textural patterns, spatial continuity, and contextual relationships that are critical for distinguishing between different ice types, enabling the learning of transferable representations from large collections of unlabeled data.

The MAE applies spatial masking, treating all input channels with equal importance during pre-training. In multimodal remote-sensing data, this design leads the model to emphasize the modalities that are most abundant in the input. Since the majority of channels in our dataset correspond to passive microwave and meteorological variables, the model may learn representations that are disproportionately influenced by those sources. This effect is critical because the HH and HV SAR backscatter channels, although fewer in number, carry the most relevant information for sea ice characterization [2]. Without adjustments, the MAE may underutilize these channels and produce suboptimal representations for downstream tasks; therefore, as shown in Figure 3.2, we restrict the input to the SAR channels only.

Working with the AI4Arctic dataset, our previous research [78] has shown the benefits of capturing large spatial context by downsampling SAR scenes. Providing the model with broader context not only improves consistency with the coarse resolution of ice charts but also mitigates non-stationary effects such as incidence-angle artifacts [30]. At the same time, focusing exclusively on coarse scales risks overlooking the fine-grained structures that are essential for identifying small-scale ice patterns. Since our goal is to build a foundation model, we seek to integrate both perspectives: leveraging global context while preserving local detail. With this goal, we evaluate whether multi- or single-resolution training is more appropriate for learning robust feature representations for sea ice mapping (see Section 3.6.4).

For multi-scale training, the design relies on multiple versions of the dataset that differ only in a scale factor (1/2, 1/3, ..., 1/10). The 1/2 version is generated using a 2×2 averaging kernel, which reduces noise while maintaining the visibility of small floes. The remaining scales are derived from this base resolution by bilinear interpolation, while the ice

charts are resized using nearest-neighbor interpolation to preserve categorical boundaries. Larger scale factors yield a greater number of patches. Therefore, to avoid bias toward any particular scale during training, a weighted sampler is used to select patches, by assigning sampling probabilities inversely proportional to the number of patches in each dataset version.

3.4 Downstream training

After pre-training, the ViT encoder from the MAE is transferred to the downstream model and fine-tuned. The encoder is followed by a feature pyramid bottleneck, which aggregates hierarchical representations, and three independent UperNet-based decoders [104], each specialized for one task: SIC, SOD, and FLOE. This multi-task learning design enables the model to capture common features shared across tasks through the encoder and bottleneck, while the decoders produce exclusive features required for task-specific predictions. The loss functions are MSE for SIC and cross-entropy (CE) for SOD and FLOE. For an input image patch of N pixels, the MSE loss for SIC is:

$$\mathcal{L}^{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2 \tag{3.4}$$

while the CE loss for SOD and FLOE is:

$$\mathcal{L}^{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log\hat{y}_{i,c} \tag{3.5}$$

where y and $\hat{y}$ denote one-hot encoded labels and prediction probabilities, respectively, and C denotes the number of classes. Previous studies [78, 105] have shown that sea ice concentration is more effectively estimated as a regression task, motivating the use of the MSE loss. Multi-task training optimizes a joint loss comprised of the weighted sum of the three components:

$$\mathcal{L}_{total} = \lambda_{SIC}\,\mathcal{L}^{MSE}_{SIC} + \lambda_{SOD}\,\mathcal{L}^{CE}_{SOD} + \lambda_{FLOE}\,\mathcal{L}^{CE}_{FLOE} \tag{3.6}$$

The λ terms are weight hyperparameters that balance the contribution of each individual task specific loss term. Since the magnitude of MSE is significantly higher than that

of CE, a larger value is assigned to the CE losses. Following Chen et al. [78], the weights were set at: $\lambda_{SIC} = 1$, $\lambda_{SOD} = 3$, and $\lambda_{FLOE} = 3$. For single-task training, each loss term is optimized one at a time by separate training runs using the same model architecture.

Figure 3.2 illustrates the downstream model. Multi-resolution inputs are used during pre-training to help the model to learn features that appear across different spatial scales. Since most of the representation learning already occurs during pre-training, including multi-resolution during fine-tuning is not strictly necessary. Exposure to multiple scales during MAE pre-training can equip the encoder with a degree of scale robustness [101], while fine-tuning primarily adapts these features to task-specific outputs. However, this assumption does not necessarily hold under coarse polygon-based supervision and patch-based inference, and we discuss these limitations in Section 3.6.3. To ensure a fair comparison, multi-resolution inputs are retained during fine-tuning so that from-scratch and pre-trained models are trained under identical conditions.

3.5 Experiments

For the experiments, we divided the available training scenes into pre-training, fine-tuning, and validation sets. The selection was pseudo-random by ensuring that all classes appeared in each set. Twenty scenes were reserved for validation, and the remaining ones were partitioned into different ratios for pre-training and fine-tuning (e.g., 60% / 40%). For each split, we evaluated the benefit of pre-training with MAE. Given the large number of experiments designed to evaluate representation transferability across tasks, resolutions, and training regimes, each configuration was executed once. This design balances experiment breadth and computational cost while enabling systematic comparison across configurations, focusing the analysis on comparative trends rather than run-to-run variance. Our implementation builds on the public repositories mmselfsup[2] and mmsegmentation[3], with adaptations for large remote sensing images and variable input channels.

For full-scene evaluation at validation and test, each SAR scene was first scaled by a factor of 1/5 and then processed in 384×384 pixel patch, with a 20% overlap between consecutive patches. Overlapping predictions were merged by averaging logits to reconstruct the scene-level output, which was subsequently upsampled to restore the original resolution. During training, validation was performed every 1000 iterations.

[2] https://github.com/open-mmlab/mmselfsup
[3] https://github.com/open-mmlab/mmsegmentation

Table 3.1: Training settings of the proposed foundation model that obtained the highest combined score on the validation set.

Settings	**Pre-training**	**Fine-tuning**
Number of iterations	30000	
Input patch size	384×384	
Batch size	64	16
Data augmentation	Random crop, random flip	
Optimizer	AdamW	
Learning rate	0.0002	0.0001
Scheduler	Cosine annealing with warmup	
Weight decay	0.01	
Attention dropout rate	0.3	-
Mask ratio	0.75	-

Table 3.1 summarizes the hyperparameters and training settings on our experiments. Configurations related to fine-tuning were obtained by maximizing the performance on the validation set training the model from scratch using ViT-b as encoder (see Section 3.5.1) and 20% of the data. MAE processes only the visible tokens, which reduces memory usage and allows for larger batch sizes compared to fine-tuning, where all tokens are used. Additionally, the MAE loss is based on MSE, which generally has a larger magnitude than the weighted loss employed in downstream tasks. This combination of larger batch size and higher loss magnitude makes it possible to use a higher learning rate during pre-training, leading to faster and more consistent convergence.

3.5.1 Architecture

Different sizes of the encoder are utilized: ViT-b (base), ViT-l (large), and ViT-h (huge), with 86M, 307M, and 632M parameters, respectively. The encoder consists of a patch embedding layer that produces tokens from 16×16 patches followed by a sequence of self-attention layers, with the same architectural settings as proposed by Dosovitskiy et al. [23]. A Vision Transformer (ViT) encoder is adopted due to its scalability with model size and its modular, patch-based representation, which facilitates integration with large self-

supervised frameworks that rely on interchangeable transformer backbones. The MAE decoder is composed of eight self-attention layers, with embedding size of 512, and 16 attention heads. For the decoder, we found beneficial to use attention dropout (see Table 3.1), as it discouraged overfitting to visible patches and prevented trivial reconstructions.

The feature pyramid bottleneck, specifically designed to connect ViT backbones to prediction heads, resizes backbone feature maps from 4 different stages using transpose convolutions for upsampling and max-pooling for downsampling. Once the dimensions match, the features are stacked and passed to independent prediction heads based on the UPerHead convolutional architecture. Each head enriches the top backbone feature by merging a Pyramid Pooling (PSP) and a Feature Pyramid Network (FPN). A final 1×1 convolution produces per-pixel predictions.

All experiments were conducted on NVIDIA A100-40GB GPUs. ViT-B was trained on a single GPU, whereas ViT-L and ViT-H required two GPUs. Training and inference times are reported in Table 3.2. Pre-training was performed on 80% of the data, followed by fine-tuning on the remaining 20%. Training time varies with the dataset partition because of the weighted sampler, which balances across resolutions and becomes more costly as dataset size increases.

Table 3.2: Training times for different backbones using an 80/20 pre-training/fine-tuning split with a mask ratio of 75% for MAE.

Backbone	**Pre-training (hrs)**	**Fine-tuning (hrs)**
ViT-b	12.9	3.8
ViT-l	35.9	6.7
ViT-h	36.9	6.9

3.6 Results

3.6.1 Overview of multi-task semantic segmentation of sea ice

The multi-task segmentation results obtained with the two training strategies are shown in Figure 3.3: a model (ViT-h backbone) trained from scratch on 60% of the training set, and the same model fine-tuned on the same portion of data after pre-training with MAE on the remaining 40%.

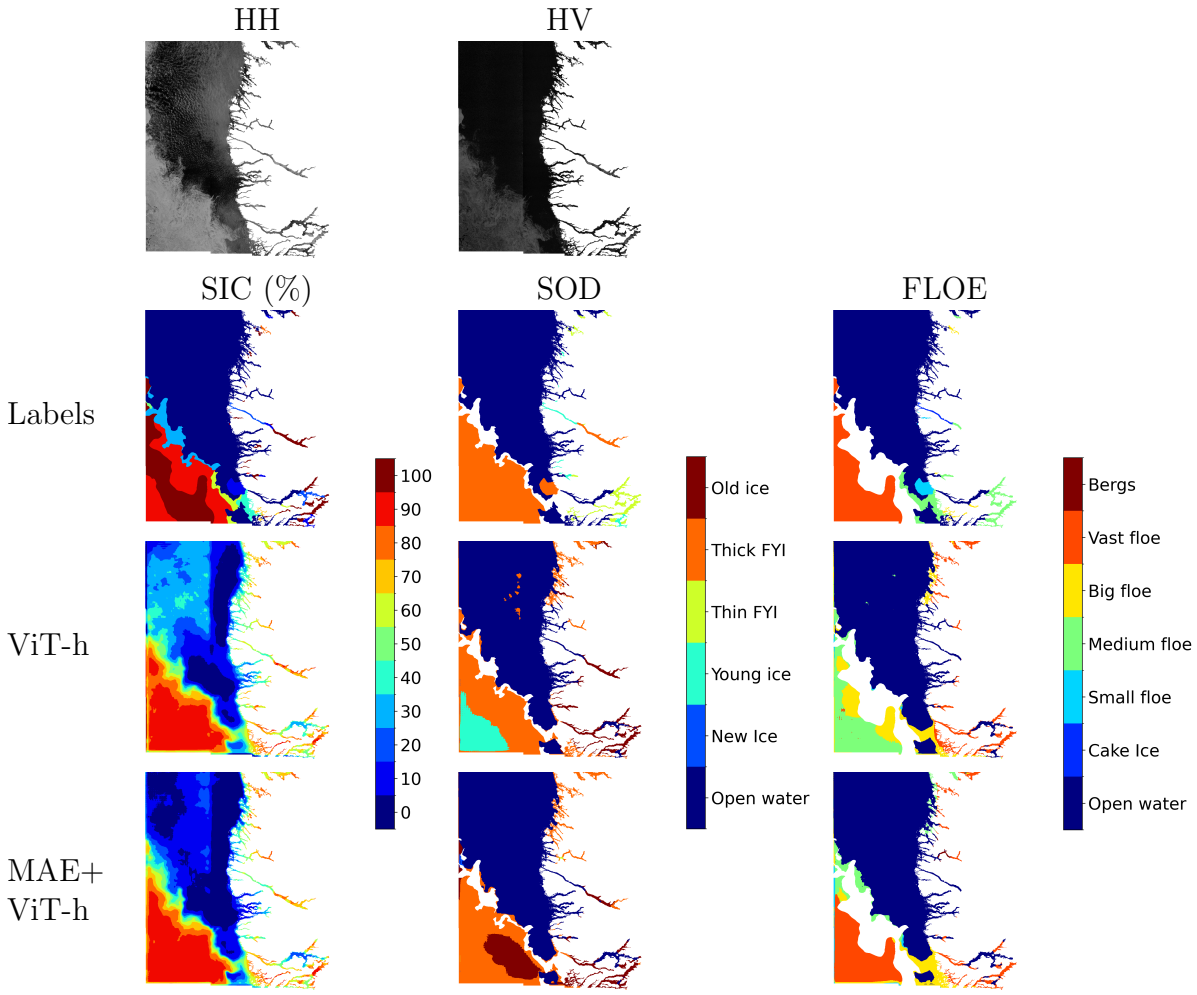

Figure 3.3: Semantic segmentation results using the ViT-h backbone: trained from-scratch versus fine-tuned after MAE pre-training. From top to bottom: SAR images (April 30, 2021, southwest Greenland), reference ice chart labels, predictions from the model trained from scratch, and predictions from the fine-tuned model. Masked areas correspond to land or unannotated polygons.

The sea ice descriptors under study (SIC, SOD, FLOE) are not arbitrary labels but ordered categories: for example, SOD progresses from open water to old ice as a continuum of ice age and thickness, and FLOE increases gradually with floe size. This ordering implies that neighboring classes are more similar and often spatially adjacent, thin first year ice

(FYI), for instance, is more likely to resemble or be located next to young ice or thick FYI than to old ice. Consequently, even when the prediction is not exact, a model that outputs a near class still captures meaningful information about the ice state. This makes it particularly relevant to assess how pre-training improves the capability to exploit such correlations.

For all three sea ice descriptors, pre-training with MAE produced predictions that more closely resembled the reference charts. Even when misclassifying, the fine-tuned model typically assigned labels that were nearer to the true class compared with the from-scratch model. For example, in the SIC case, the reference shows a large polygon of open water (0% concentration). The from-scratch model predicts values ranging from 0 to 50%, whereas the fine-tuned model provides a more homogeneous estimate of 0–20%. The 100% concentration polygon was not perfectly captured, but the fine-tuned model consistently predicted 90%. The open water region also highlights sensitivity to sub-swath discontinuities, particularly visible in the cross-polarized (HV) channel.

For SOD, the largest polygon is annotated as thick FYI and appears homogeneous in the SAR image. Both models predict two stages within this polygon: one common stage (thick FYI) and one differing stage. The from-scratch model misclassifies part of the region as young ice, while the fine-tuned model misclassifies it as old ice. Considering that young ice is smoother, more saline, and thus produces lower backscatter than thick FYI, the prediction toward old ice is a more plausible error. Old ice shares a higher backscatter level with thick FYI, making the fine-tuned model's output more consistent with the SAR signature of the area. In the illustrated image, the benefits of self-supervised pre-training are most evident for FLOE, where the major polygons are classified correctly.

3.6.2 Foundation model transfer across data regimes

To evaluate the performance of our method, we use three primary metrics corresponding to the different sea ice descriptors. For SIC, we report the R^2 score, a standard metric for regression tasks that quantifies how well the predicted values match the ground truth. For SOD and FLOE, we use the F1 score, which balances precision and recall to assess the accuracy of predicted classes. Following the AutoIce evaluation conventions [4], these three metrics are combined into a single overall metric called the combined score, computed as a weighted average with weights of 2/5 for SIC, 2/5 for SOD, and 1/5 for FLOE, reflecting the relative importance of each sea ice parameter in the evaluation. For each experimental protocol, the model delivering the higher combined score on the validation set is chosen during training.

Table 3.3 reports the performance gains obtained by fine-tuning multi-task models that were pre-trained with different proportions of data, compared to training the same models from scratch. For each row, the percentage of data used during pre-training and fine-tuning is specified. Columns show the differences in the evaluation metrics for the three sea ice parameters SIC-R^2, SOD-$F1$, FLOE-$F1$, as well as the combined score. Positive values indicate that pre-training improved performance relative to training from scratch, while negative values indicate a decrease. The table also compares the three backbone sizes (ViT-b, ViT-l, ViT-h). Overall, the table demonstrates how pre-training with different data proportions affects downstream performance across metrics and backbones. For comparison purposes, we found it clearer to illustrate relative differences in the metrics, while absolute scores are discussed in Section 3.6.6 and Table 3.5.

Experiments on pre-training/fine-tuning split.

At the 99/1 split, where almost all data is used for pre-training and only 1% is left for supervised fine-tuning, most differences are negative. Small positive gains appear only for ViT-h on SOD (+1.9%) and FLOE (+0.6%). These results confirm that extremely limited fine-tuning data is insufficient to specialize the pretrained representations, even when the backbone is large. However, the fact that ViT-h still recovers slight improvements suggests that larger models benefit more from MAE's representational prior under extreme data scarcity.

As the fine-tuning ratio increases to more balanced regimes (95/5, 90/10, 80/20, and 60/40), positive gains become more frequent and more consistent. Examples include +8.6% SIC improvement for ViT-l at a 90/10 split, +4.0% FLOE improvement for ViT-l at the same split, and +3.4% combined-score improvement for ViT-h at 60/40. These results indicate that moderate amounts of fine-tuning data are required to adapt the pretrained features to the multi-task setting, while pre-training still provides an advantageous initialization.

The largest gains occur when fine-tuning dominates (40/60 and 20/80). For ViT-h, improvements reach +10.3% in SIC, +4.0% in SOD, +4.3% in FLOE, and +6.6% in combined score at the 40/60 split. At 20/80, gains remain substantial across all metrics (SIC +5.9%, SOD +5.1%, FLOE +4.9%, combined score +5.4%). Overall, MAE pre-training is most beneficial when combined with abundant supervised training, enabling the pretrained encoder to be effectively specialized to the downstream tasks.

Table 3.3: Performance gains from MAE pre-training compared to training from scratch, reported as percentage differences across sea ice concentration SIC (R^2), SOD (F1), FLOE (F1), and the combined score. Results are shown for three ViT backbones (b, l, h) under different pre-training/fine-tuning data splits. Positive values (gray-shaded cells) indicate improvements from pre-training.

Pre-training (%) / Fine-tuning (%)	**SIC-R^2 (%)**			**SOD-$F1$ (%)**			**FLOE-$F1$ (%)**			**Combined Score (%)**		
	ViT-b	ViT-l	ViT-h	ViT-b	ViT-l	ViT-h	ViT-b	ViT-l	ViT-h	ViT-b	ViT-l	ViT-h
99 / 1	-17.2	-14.0	-5.0	-0.1	-1.1	**+1.9**	-0.5	-2.0	**+0.6**	-7.0	-6.4	-1.1
95 / 5	**+1.2**	-15.8	-0.3	-0.3	-2.4	**+3.0**	**+1.2**	-0.9	-0.7	**+0.6**	-7.5	**+0.9**
90 / 10	-6.4	**+8.6**	-0.1	-0.4	**+0.1**	-1.6	-2.3	**+4.0**	**+1.6**	-3.2	**+4.3**	-0.3
80 / 20	**+5.6**	-2.7	-1.0	-1.4	**+0.5**	**+1.2**	**+3.0**	**+0.3**	**+1.8**	**+2.3**	-0.8	**+0.4**
60 / 40	-1.0	-0.4	**+5.2**	**+0.3**	-2.5	**+2.2**	**+2.3**	-2.1	**+2.1**	**+0.2**	-1.6	**+3.4**
40 / 60	-2.9	**+1.0**	**+10.3**	-0.1	-0.4	**+4.0**	-0.6	-2.4	**+4.3**	-1.3	-0.2	**+6.6**
20 / 80	**+2.0**	**+1.9**	**+5.9**	**+1.9**	**+0.9**	**+5.1**	**+1.0**	**+0.9**	**+4.9**	**+1.7**	**+1.3**	**+5.4**

Experiments on backbone scaling.

Across backbone sizes, the effect of MAE pre-training varies substantially. ViT-b exhibits the weakest gains, with improvements appearing only when sufficient fine-tuning data is available (e.g., positive combined-score gains at 80/20 and 20/80). At high pre-training ratios such as 99/1 or 95/5, ViT-b often shows negative transfer, suggesting that its limited capacity restricts its ability to leverage pretrained features when downstream supervision is scarce. ViT-l displays more mixed behavior: it benefits strongly from pre-training at certain intermediate splits, such as the +8.6% gain in SIC at 90/10 or the +4.3% combined score improvement at the same split, yet remains sensitive to data imbalance and occasionally under-performs when fine-tuning data is insufficient. In contrast, ViT-h consistently yields the largest and most stable improvements across nearly all pre-training and fine-tuning configurations. Its substantial capacity allows it to retain transferable representations from MAE while still adapting effectively during supervised training. The magnitude and consistency of gains for ViT-h indicate that larger backbones are better suited to benefit from self-supervised pre-training in the context of multi-task sea ice mapping.

Observations by task.

The impact of MAE pre-training varies across downstream tasks. SIC, the only regression task, shows strong dependence on backbone size: ViT-h achieves some of the strongest improvements across many of the moderate and fine-tuning–dominant splits (e.g., +10.3% at 40/60), whereas ViT-b and ViT-l exhibit gains only in specific settings. This suggests that modeling large-scale concentration patterns requires representational capacity that smaller backbones may not adequately provide. SOD exhibits generally smaller but often positive gains, particularly for ViT-h, even when fine-tuning data is limited. Because SOD relies more on local texture and less on global spatial information, the ability of MAE to capture generic SAR structures appears sufficient to benefit this task across a broader range of splits. FLOE displays the most consistent improvements after SIC, especially for ViT-l and ViT-h, with notable gains at moderate splits (e.g., +4.0% for ViT-l at 90/10 and +4.3% for ViT-h at 40/60). FLOE prediction depends on identifying both fine-grained floe textures and contextual shape patterns, for which pretrained representations are particularly valuable. Overall, MAE pre-training contributes positively across all tasks, with the largest improvements occurring in tasks requiring richer spatial or contextual understanding.

3.6.3 Multi- vs. single-task training

To assess whether downstream sea ice tasks benefit more from multi-task or from single-task learning, we compare both strategies using the ViT-h backbone on the 40/60 pre-training/fine-tuning data split, a configuration previously shown to yield the strongest transfer performance (see combined score in Table 3.3). Unlike earlier multi-task experiments, where model selection was based on the combined score, we select here, for every task, the checkpoint that performs best on its own validation metric. This ensures a strictly fair comparison with single-task baselines, which are by definition optimized for one task only. Table 3.4 summarizes the results for models trained from scratch (ViT-h) and with MAE pre-training (MAE+ViT-h). Orange shading highlights the strongest scores when comparing multi- vs. single-task training, while blue shading marks the performance gain obtained through fine-tuning relative to training from scratch.

Table 3.4: Performance of multi-task and single-task downstream models using a shared ViT-h backbone under a 40/60 pre-training/fine-tuning split. Orange cells mark stronger performance between training regimes, and blue cells represent MAE pre-training gains.

Method		**SIC-R^2**		**SOD-$F1$**		**FLOE -$F1$**		**Combined score**	
		value	dif	value	dif	value	dif	value	dif
Multi-task	MAE+ViT-h	74.8	**+10.7**	72.5	-0.2	63.4	**+1.1**	71.6	**+4.4**
	ViT-h	64.1		72.8		62.4		67.2	
Single-task	MAE+ViT-h	73.5	**+4.5**	77.7	**+4.2**	65.7	**+2.7**	73.6	**+4.0**
	ViT-h	69.0		73.6		63.1		69.7	

Across nearly all metrics, single-task training outperforms multi-task training, with a single exception: SIC-R^2 for the multi-task fine-tuned model. Despite this overall trend favoring task-specific optimization, the table also shows that MAE pre-training consistently strengthens both training regimes. Fine-tuned models exhibit positive gains in nearly all metrics when compared to their from-scratch counterparts, indicating that the pretrained encoder provides a robust initialization that supports more stable and effective adaptation to downstream objectives. This effect is particularly evident for SIC-R^2, where pre-training yields the largest improvements, but noticeable gains are also observed for SOD and FLOE under single-task learning. Even in the few cases where differences are marginal, as in SOD for the multi-task model, the pretrained backbone prevents the notable degradations typically found in multi-task settings initialized from scratch, underscoring the value of encoding generalizable SAR-specific priors.

Importantly, the fact that single-task models continue to outperform multi-task models even after MAE pre-training suggests that pre-training and training regime influence the encoder in fundamentally different ways. Pre-training primarily enhances the quality and richness of the learned representations, supplying a strong foundation of structural and textural features that are broadly useful across tasks. In contrast, the choice between multi-task and single-task learning governs how this shared representational space is allocated and refined during fine-tuning. Single-task models can dedicate the full encoder capacity to the requirements of a single objective, whereas multi-task models must distribute that same capacity across all downstream tasks, which can limit their ability to specialize despite starting from a strong pretrained initialization.

Finally, the combined score reflects these dynamics: although pre-training narrows the gap between the two regimes, the single-task models still provide the strongest overall downstream performance. The improvement brought by MAE pre-training does not diminish when each task is trained independently; rather, it compounds with the benefits of specialization. These results indicate that, for the present set of heterogeneous sea ice tasks, encoder-level sharing is not sufficient to outweigh the advantages of task-specific adaptation. They further suggest that achieving the full potential of multi-task learning in this context may require architectures or optimization strategies capable of more explicitly managing task interactions, such as adaptive loss balancing or the incorporation of additional modalities, including language or ancillary geophysical variables from other sensors, to provide complementary structure and reduce reliance on a single shared encoder.

3.6.4 Multi- vs. single-scale training: resolution transferability

To study how changes in image resolution affect downstream performance, we evaluate several model configurations that differ in whether pre-training and fine-tuning are performed using multiple resolutions (scaling factors between 1/2 and 1/10) or a single resolution (1/5). At test time, each model is evaluated at a single resolution, ranging from full resolution (1) to 1/9, and all predictions are resampled to 1/5 before metric computation. This normalization step ensures that differences in evaluation reflect model behavior rather than resolution-induced artifacts, such as sharper boundaries at higher resolutions or excessive smoothing at coarser scales. The experiments therefore assess resolution transferability, defined as the ability of a model trained under a given scale regime to generalize when inference is performed at a different spatial resolution.

Figure 3.4 shows the combined score as a function of test time resolution for all evaluated methods. Across most configurations, performance decreases as the test resolution

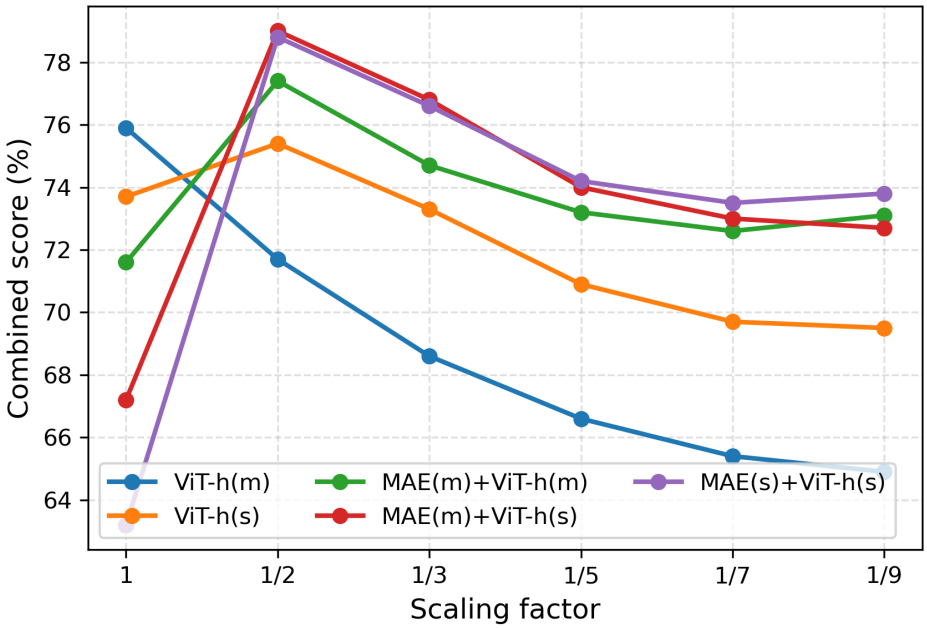

Figure 3.4: Combined score as a function of test-time resolution for different training strategies. Suffix (m) denotes multi-scale training and (s) single-scale training at the MAE pre-training or downstream fine-tuning stage. Trends discussed in the text primarily refer to intermediate and coarse resolutions (1/2–1/9), while the full-resolution case (1) exhibits distinct behavior due to distribution mismatch with the pre-training scale range.

becomes coarser, with higher scores generally obtained at finer scales. This trend persists despite the polygon-based and spatially coarse nature of ice-chart annotations, indicating that higher-resolution SAR inputs still provide informative local structure, such as textural variation and scattering patterns, that can be exploited even under coarse supervision. Under patch-based ViT inference, resolution directly affects the effective receptive field, making these fine-scale features particularly relevant. In this context, although multi-scale training is commonly associated with improved scale robustness, our results show that under coarse polygonal supervision and patch-based inference, single-scale training yields more stable generalization to unseen resolutions, as additional scale variability is not consistently supported by the label geometry. Multi-scale training is therefore likely to become beneficial only when supervision more closely follows natural image boundaries or when scale-aware architectural mechanisms are introduced. This behavior is consistent with prior remote sensing studies showing that coarse supervision does not prevent gains from higher-resolution imagery [106]. A notable exception occurs at full resolution, where

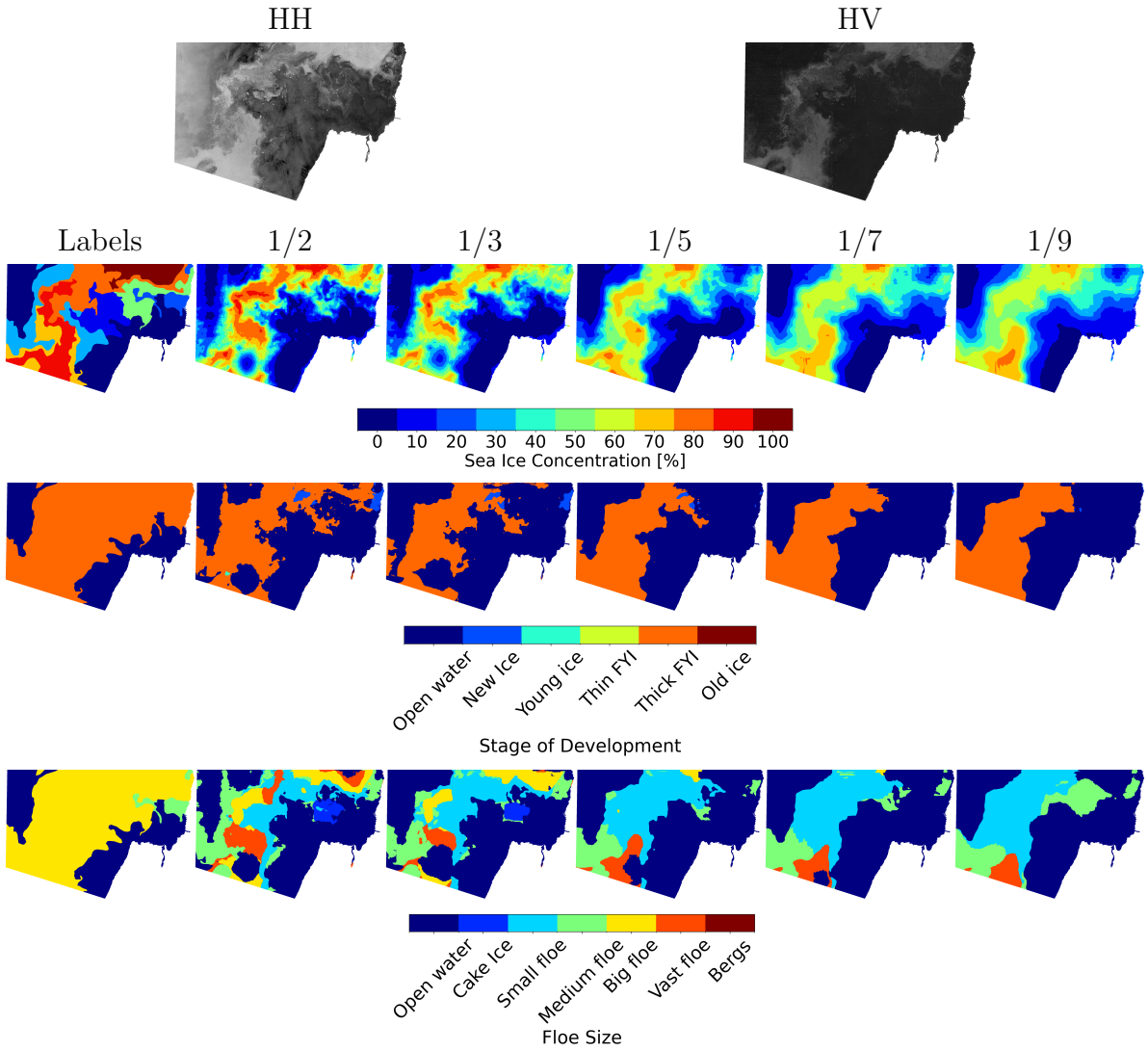

Figure 3.5: Semantic segmentation results at different test-time resolutions using the MAE(m)+ViT-h(s) model. Suffixes (m) and (s) denote multi-scale and single-scale training, respectively, at the MAE pre-training or downstream fine-tuning stage. From top to bottom, the rows show the SAR input image (July 19, 2020, Foxe Basin), SIC, SOD, and FLOE. From left to right, columns correspond to the ground-truth ice-chart labels and predictions obtained from inputs at different spatial resolutions, as indicated by the scaling ratios above each column.

from-scratch models outperform MAE-pretrained variants, suggesting a distribution mismatch between MAE pretraining, conducted on down-scaled inputs, and native-resolution SAR statistics.

The qualitative results in Figure 3.5 provide additional insight into the effect of test-time resolution on model predictions. At higher resolutions, predictions tend to capture finer spatial detail than the polygon-based ice charts, often producing boundaries that are visually sharper and more aligned with underlying SAR texture. While increased spatial detail can be beneficial, it also makes predictions more sensitive to image artifacts such as sub-swath discontinuities, particularly visible in cross-polarized channels. As the test resolution decreases, predicted regions become progressively coarser and more geometrically aligned with the ice-chart polygons. However, apparent geometric consistency is frequently accompanied by incorrect class assignments within large regions, indicating a loss of discriminative information rather than improved semantic agreement. This pattern aligns with the quantitative trends reported in Fig. 3.4, where coarser resolutions lead to systematic performance degradation despite increased visual similarity to the ground-truth geometry.

When excluding the full-resolution case, clear differences emerge when comparing training strategies. From-scratch models exhibit strong sensitivity to the scale regime used during training. In particular, the single-scale model ViT-h(s) consistently outperforms its multi-scale counterpart ViT-h(m) across most test resolutions, with ViT-h(m) showing a sharp performance drop as resolution decreases. While ViT-h(m) performs competitively at full resolution, its lack of robustness at coarser scales indicates that multi-scale exposure during supervised training introduces variability that the encoder cannot reliably reconcile under coarse supervision. By contrast, training at a fixed resolution provides a more stable inductive bias, enabling more gradual performance degradation under resolution shifts.

MAE pre-training substantially reduces the resolution sensitivity exhibited by from-scratch models. Except at full resolution, MAE-pretrained models consistently outperform their from-scratch counterparts and exhibit flatter performance curves across scales. The strongest and most stable results are obtained by MAE(s)+ViT-h(s), with MAE(m)+ViT-h(s) closely tracking its performance, indicating that MAE induces scale-robust representations that are largely preserved during single-scale fine-tuning. The reduced performance of MAE-based models at full resolution likely reflects the absence of native-resolution samples during pre-training, rather than a limitation of the approach itself. Overall, the results indicate that while inference resolution systematically affects performance, multi-scale training is not inherently advantageous under coarse, polygon-based supervision. Instead, MAE pre-training combined with consistent downstream resolution provides the most reliable resolution transferability, motivating future work on scale-aware pre-training and

resolution-conditioned architectures to better handle native-resolution SAR imagery.

3.6.5 Mask ratio ablation

We conducted an ablation study on the MAE mask ratio, defined as the fraction of input tokens that are hidden during MAE pre-training for the model to reconstruct. Exploring different mask ratios reveals how pre-training best supports each downstream task and how the optimal setting interacts with backbone size. Among the different pre-training/fine-tuning splits, we chose 80/20 for the ablation experiments. Similar to Table 3.3, Figure 3.6 illustrates the gains obtained after fine-tuning the different pre-trained backbones compared with training from scratch.

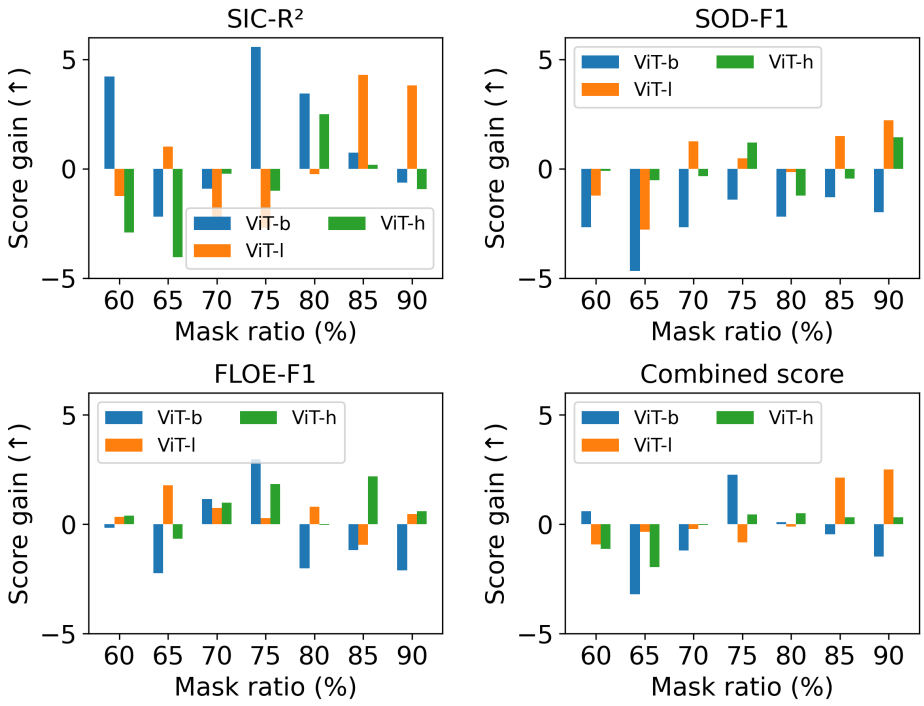

Figure 3.6: The charts report the difference metrics (%) of fine-tuned models, pre-trained with different mask ratios, with respect to the model trained from scratch. MAE and downstream stages are trained with 80% and 20% of the training set, respectively.

The ablation results show that the impact of the MAE mask ratio varies across backbones and downstream tasks. For SIC-R^2, moderate masking continues to benefit smaller backbones, with ViT-b achieving its strongest gains at intermediate ratios, while ViT-l

shows improvements only at select high ratios. In contrast, ViT-h exhibits mostly modest or negative gains across mask ratios, suggesting that under the 80/20 data split, larger backbones do not consistently benefit from aggressive masking. These trends indicate that the optimal mask ratio for SIC depends on both model capacity and data regime.

For SOD-$F1$ and FLOE-$F1$, gains are generally weaker and more scattered. SOD shows predominantly negative or marginal improvements for ViT-b and only isolated positive gains for ViT-l and ViT-h, indicating increased sensitivity to masking compared with earlier experiments. FLOE displays a similarly mixed pattern: ViT-b benefits only sporadically, ViT-l remains relatively stable with small gains, and ViT-h achieves moderate improvements at a few mask ratios but without a consistent trend. Overall, these results suggest that moderate masking is more favorable than very high masking for these classification tasks.

The combined score reflects these task-level behaviors. ViT-b achieves its best improvement at a moderate mask ratio, while ViT-l shows stronger gains at higher ratios, but not inconsistently. ViT-h exhibits only small improvements across a limited range of ratios, confirming that aggressive masking no longer yields clear advantages for the largest backbone. Collectively, the results reveal a weaker and less predictable relationship between mask ratio and downstream performance, with optimal settings varying by task, backbone, and data split. Based on the ablation results, we selected a mask ratio of 75% as a balanced configuration, as it achieves the highest average combined score gain across backbones while avoiding large performance drops across tasks.

3.6.6 Comparison with other architectures

Table 3.5 compares the combined score achieved by a range of segmentation architectures across different pre-training and fine-tuning data regimes. The evaluated baselines include convolutional models (U-Net, ResNet-50, DeepLabv3+, ConvNeXt), and transformer-based architectures trained from scratch (SegFormer and ViT variants). The proposed MAE+ViT models incorporate self-supervised masked autoencoder (MAE) pre-training followed by supervised fine-tuning, enabling a direct comparison between foundation-style initialization and purely supervised training. Visualizations about this experiments can be found in Appendix B

Rather than focusing solely on absolute maxima, dominance in this comparison is assessed through consistency across data regimes, as indicated by the top-three rankings highlighted in Table 3.5. Under this criterion, MAE-pretrained Vision Transformers, particularly MAE+ViT-h, emerge as the most consistently strong approach. MAE+ViT-h

Table 3.5: Comparison of combined scores across architectures. Baseline models (U-Net, ResNet-50, DeepLabv3+, ConvNext, SegFormer) are trained from scratch without pre-training, while the proposed MAE+ViT models incorporate self-supervised pre-training. Blue shading highlights the top three scores within each data regime, emphasizing cross-regime dominance rather than isolated peak performance.

Method	**Combined score (%)**							
	From-scratch training (% data)							
	1	5	10	20	40	60	80	100
U-Net	54.2	66.3	65.6	68.2	70.8	70.1	72.4	68.8
ResNet-50	53.3	65.5	67.4	62.6	65.0	58.9	60.4	62.3
DeepLabv3+	44.2	56.7	57.2	55.3	61.9	59.5	57.9	65.6
ConvNext	63.8	67.1	72.3	71.1	70.6	72.2	72.0	72.9
SegFormer	57.4	70.9	72.1	68.8	73.0	72.7	74.6	75.6
ViT-b	65.6	69.8	72.0	70.1	71.1	71.8	74.2	72.5
ViT-l	64.0	71.8	70.1	69.6	72.3	71.2	72.2	74.1
ViT-h	60.1	71.2	73.7	70.6	68.0	66.6	69.5	68.2
	Pre-training (%) / Fine-tuning (%)							
	99/1	95/5	90/10	80/20	60/40	40/60	20/80	0/100
MAE+ViT-b	58.6	70.4	68.9	72.3	71.3	70.5	76.0	-
MAE+ViT-l	57.6	64.4	74.4	68.8	70.7	71.0	73.5	-
MAE+ViT-h	58.9	72.1	73.3	71.0	71.4	73.2	74.9	-

ranks within the top three combined scores in the majority of data splits and attains the best overall score in two regimes (95/5 and 40/60). When considering all splits where MAE-based models are applicable, MAE+ViT variants achieve top-three performance in five out of seven regimes, demonstrating robust competitiveness across a wide range of supervision levels. Importantly, this dominance is not limited to fine-tuning–heavy settings. MAE-pretrained models perform competitively already in low-to-moderate supervision regimes (95/5, 90/10, and 80/20). These results indicate that MAE pre-training provides transferable representations that can be effectively specialized even when supervised data is limited, when fine-tuning data is not vanishingly small.

Among the from-scratch baselines, SegFormer achieves strong performance and attains the highest score in the 60/40 regime, while also dominating the fully supervised (0/100) setting. However, the fully supervised case is not directly comparable to MAE-based models, which by design rely on a pre-training stage and are therefore absent from the 0/100 regime. The comparison between supervised and foundation-style approaches is therefore more informative in the partially supervised regimes.

Overall, the comparison highlights a clear distinction between architectures relying exclusively on supervised training and foundation model behavior. While certain supervised models can achieve higher peak performance in specific regimes, MAE-pretrained Vision Transformers, especially MAE+ViT-h, offer greater stability and consistency across data regimes, aligning more closely with the goals of foundation models: scalable pre-training, adaptability to multiple downstream tasks, and robustness under varying levels of supervision.

3.7 Discussion

Our experiments provide key insights into the development of a foundation model for Arctic sea ice segmentation, highlighting several factors central to its effectiveness.

Across all experiments, MAE pre-training improves performance relative to training from scratch when sufficient supervised data is available for fine-tuning. Gains are limited under extreme data scarcity, indicating that self-supervised pre-training does not replace the need for downstream adaptation. Model capacity plays a critical role in this process: larger backbones consistently benefit more from MAE pre-training, while smaller models are more sensitive to negative transfer under limited supervision. These trends align with broader observations in foundation models, where large encoders are better able to retain general representations while reallocating capacity during task-specific fine-tuning.

The comparison between multi-task and single-task learning reveals clear task-dependent trade-offs. Despite sharing a common encoder, single-task models consistently outperform multi-task counterparts for SOD and FLOE, while SIC is the only descriptor that occasionally benefits from joint training. This behavior reflects differences in spatial scale, semantic granularity, and label noise across tasks. Shared encoders without task-aware mechanisms are insufficient to reconcile these competing objectives, even when initialized with MAE pre-training. While pre-training strengthens both regimes, it does not eliminate the need for task-specific adaptation, suggesting that effective multi-task learning for sea ice analysis will require more explicit mechanisms for balancing task interference.

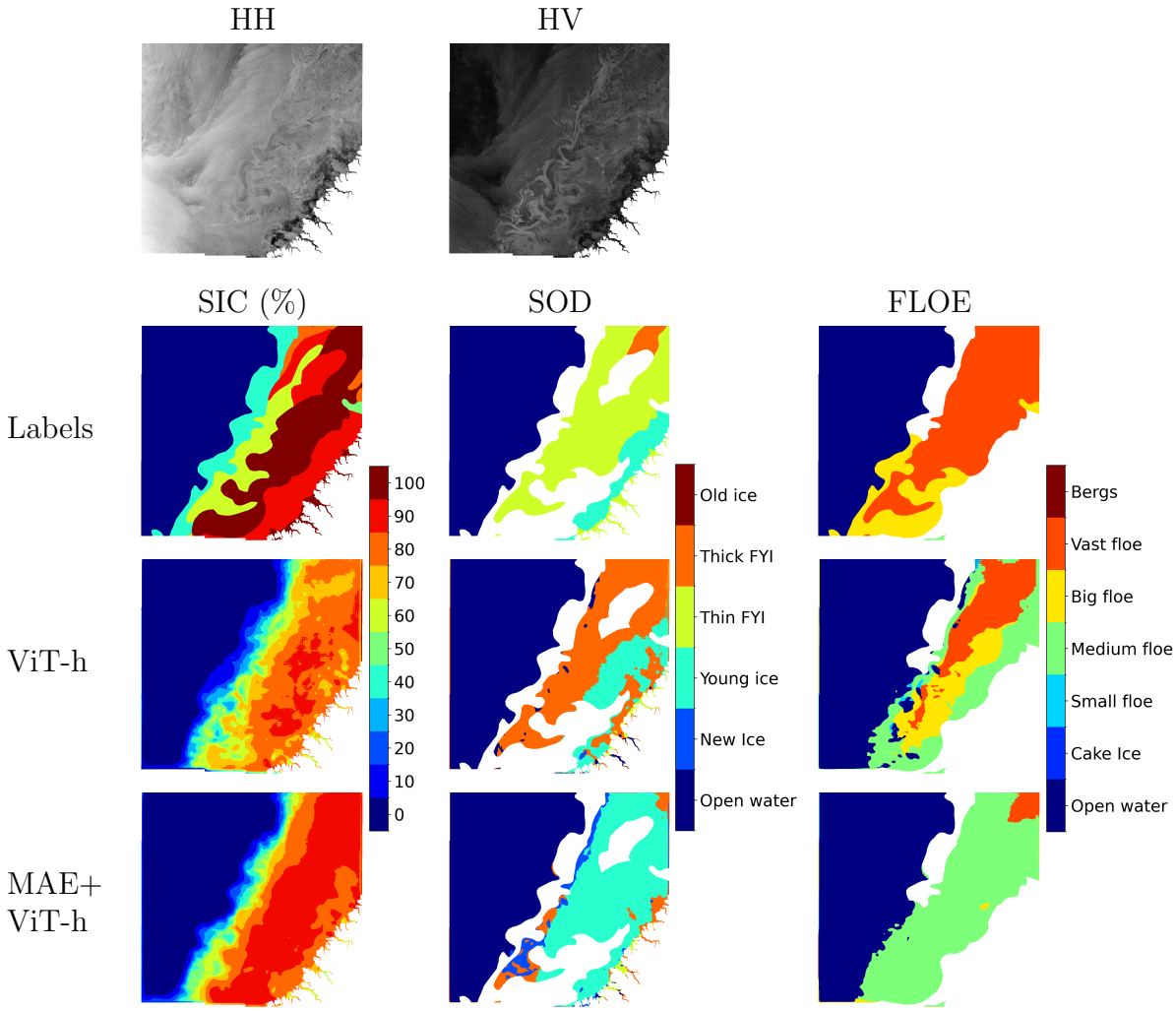

Figure 3.7: Semantic segmentation results using the ViT-h backbone: trained from scratch versus fine-tuned after MAE pre-training. **From top to bottom**: SAR images (February 12, 2020, Canadian Eastern Arctic), reference ice chart labels, predictions from the model trained from scratch, and predictions from the fine-tuned model. Masked (white) areas correspond to land or unannotated polygons.

The interpretation of performance metrics must be situated within the limitations of the annotation strategy. Ice charts provide polygon-based approximations of sea ice conditions, manually drawn under operational constraints. These annotations assume homogeneity

within polygons, whereas SAR imagery often reveals fine-scale heterogeneity. As illustrated in Figure 3.7, this mismatch becomes evident when comparing predicted segmentations to reference polygons. The predictions tend to follow natural textural and structural boundaries in the SAR images rather than polygon edges, leading to discrepancies that do not necessarily reflect errors in a geophysical sense. Interestingly, after fine-tuning, the model predictions align more closely with polygon homogeneity, but not always with the proper class (SOD and FLOE), resulting in higher scores for SIC only. This illustrates how metric improvements can partly reflect adaptation to annotation conventions rather than genuine advances in capturing ice dynamics.

Resolution transfer experiments further highlight the interaction between supervision geometry and representation learning. While higher test-time resolutions generally improve performance, multi-scale training does not consistently enhance generalization under polygon-based supervision. In many cases, single-scale training provides more stable resolution transfer, particularly for from-scratch models. MAE pre-training reduces sensitivity to resolution changes and leads to flatter performance trends across scales, indicating that generative pre-training captures structural properties of SAR imagery that support scale robustness. However, the benefits of explicit multi-scale training remain limited in the absence of scale-aware architectures or higher-fidelity annotations that better encode geometric detail.

3.8 Conclusion

This chapter presented a systematic investigation of self-supervised pre-training for developing foundation models tailored to Arctic sea ice classification from SAR imagery, as well as the creation of SAR-IceFM, an open-source SAR-based foundational model for sea ice analysis. By leveraging masked MAE pre-training on multi-resolution Sentinel-1 data, we addressed a key limitation of existing remote sensing foundation models, which remain largely biased toward multispectral imagery and land-based applications. The proposed training and evaluation framework explores a unified treatment of sea ice concentration, stage of development, and floe size within a shared backbone, and extensive experiments on the AI4Arctic dataset demonstrate that MAE pre-training consistently improves performance relative to training from scratch, with gains amplified by increased model capacity.

The results also reveal important constraints on the effectiveness of pre-training. Larger architectures are better able to extract and reuse transferable representations, reinforcing the central role of scaling in foundation models. However, downstream performance is strongly shaped by the structure of the available annotation. Polygon-based ice charts

impose assumptions of homogeneity that do not fully reflect the fine-scale heterogeneity observed in SAR imagery, complicating the interpretation of quantitative metrics. As a consequence, improvements in evaluation scores may partly reflect adaptation to annotation conventions rather than unambiguous advances in capturing physical ice processes.

Taken together, the findings provide a clear answer to the question motivating this work: pre-training is necessary but not sufficient. While MAE-based pre-training strengthens representation learning and improves robustness across data regimes, it does not, by itself, resolve challenges arising from heterogeneous task objectives, resolution variability, and coarse supervision. Effective multi-task foundation models for sea ice analysis therefore require careful alignment between pre-training strategy, task formulation, resolution handling, and annotation quality, rather than reliance on pre-training alone.

Looking forward, several directions emerge to advance foundation models for sea ice classification. Improved annotation strategies that better capture sub-polygon variability would reduce label noise and enable more faithful evaluation of model predictions. Alternative learning formulations that explicitly encode the ordered or continuous nature of sea ice descriptors may further enhance physical consistency. Finally, extending pre-training to larger and more diverse data, including multiple SAR missions and complementary modalities such as optical and passive microwave data, offers a promising path toward more generalizable and operationally robust foundation models for Arctic sea ice monitoring.

Chapter 4

Conclusions and Future Directions

This thesis investigated how feature representations learned from SAR imagery can be made robust, scalable, and operationally meaningful for Arctic sea ice mapping. The central premise was that sea ice classification is not merely a prediction problem, but fundamentally a representation learning problem under spatial non-stationarity and supervision constraints. To address this challenge, two complementary strategies were developed and evaluated: a structured supervised framework and a self-supervised SAR foundation model. Together, they examined how representation structure, shaped either through architectural design or large-scale pre-training, affects robustness, transferability, and operational reliability.

4.1 Summary of Contributions and Findings

- Design of the Irregular Tokens on Transformers (ITT) framework, combining CNN feature extraction, IRGS-based region tokenization, and transformer self-attention applied to irregular homogeneous regions. Self-attention over spatially sparse tokens captured long-range dependencies more effectively than purely convolutional baselines, improving robustness to spatial non-stationarity.

- Introduction of a multi-task training scheme integrating pixel-level and region-level supervision to promote region-consistent feature representations. The region-based loss regularized CNN feature maps, leading to improved boundary consistency, reduced predictive entropy, and greater stability across melt and freeze-up seasons.

- Systematic evaluation under leave-one-out protocols and operational expert audit by the CIS. The pixel-level U-Net+ITT configuration achieved the highest average accuracy and reduced variance across scenes, with statistically significant improvements over prior methods using the same data. Operational validation confirmed sharper boundary delineation and reduced uncertainty in complex seasonal conditions.

- Key supervised representation insight: Explicitly structuring supervision to incorporate regional coherence and global context reshapes feature spaces toward spatial consistency and operational reliability. Robustness emerges not merely from deeper architectures, but from embedding spatial relationships directly into the learning process.

- Development of SAR-IceFM, a self-supervised foundation model trained on large-scale unlabeled Sentinel-1 SAR imagery using masked representation learning and multi-task objectives. Self-supervised pre-training substantially improved downstream segmentation performance, increasing robustness under limited annotated data.

- Systematic investigation of representation scaling with respect to data size, model capacity, and fine-tuning strategies. Representation quality improved consistently with increasing scale; however, gains depended on task-aligned fine-tuning. Pre-training alone did not eliminate supervision geometry constraints.

- Evaluation of resolution transferability and cross-condition robustness. Even under coarse, polygon-based supervision, finer SAR resolutions consistently improved performance, indicating that high-resolution inputs provide discriminative features that remain informative despite region-level labeling. In contrast, multi-scale training did not consistently enhance robustness, as scale variability cannot be fully exploited when annotation geometry does not follow fine spatial details.

- Key self-supervised representation insight: Large-scale self-supervised learning produces transferable SAR representations, however, pre-trained representations must be adapted in a way that respects the geometry and meaning of sea ice annotations to achieve optimal performance. Foundation models reduce annotation dependence; however, domain-aware fine-tuning remains essential.

Taken together, these findings demonstrate that robust sea ice mapping is fundamentally a representation learning problem, where performance depends on how representations are structured and aligned with the data. While structured supervision improves spatial

consistency and interpretability, and self-supervised pre-training enhances scalability and transferability, neither is sufficient in isolation. Effective models must integrate both principles.

4.2 Future work

The findings presented in this thesis directly fulfill the stated objective of advancing principled and scalable representation learning strategies for SAR-based sea ice mapping. However, several research venues remain open. Future work may extend this thesis along the following directions:

Model

- **Scale Robustness of Irregular Tokens.** Although ITT demonstrated improved robustness to spatial non-stationarity, the behavior of irregular region tokens under strong object scale variability warrants further investigation. Future work should analyze how token granularity interacts with varying spatial resolutions and heterogeneous object sizes, particularly in melt-season conditions where floe fragmentation increases. Developing scale-aware tokenization strategies or adaptive region merging mechanisms may further stabilize representations across scales.

- **Embedding Region Descriptors into Positional Encoding.** The current ITT formulation relies primarily on spatial structure derived from unsupervised segmentation. A promising extension is to incorporate region-level descriptors, such as geographic coordinates, incidence angle, acquisition time, or seasonal metadata, into transformer positional encoding. By embedding physically meaningful contextual features into the attention mechanism, representations may become more domain-aware and less sensitive to spatial non-stationarity driven by acquisition geometry.

- **Contextual Monte Carlo Ensembling (CMCE).** Transformer inference over large-scale imagery is constrained by sequence length limitations, requiring partitioning into subsets of regions. Conventional deterministic partitioning restricts each region to a single contextual neighborhood, limiting the expressive power of self-attention. CMCE proposes a stochastic inference strategy for ITT in which regions are repeatedly sampled into different contextual groupings during inference. Because transformer representations depend on token co-occurrence, each grouping produces context-dependent predictions. Aggregating these predictions through confidence-weighted averaging enables more robust region-level inference without retraining the

model. Future research should evaluate CMCE under large-scale Arctic scenes, analyze its uncertainty behavior, and explore adaptive sampling strategies that prioritize ambiguous regions. This approach may provide a principled way to extend contextual modeling beyond fixed sequence constraints.

- **Temporal Representation Learning.** This thesis primarily focused on single-date SAR imagery; however, sea ice evolution is inherently dynamic and governed by temporal processes. Future research should therefore incorporate multi-temporal SAR sequences to explicitly model seasonal transitions and short-term ice evolution, potentially through spatio-temporal transformers. Temporal representation learning could reduce uncertainty during melt and freeze-up periods, improve robustness under rapidly changing ice conditions, and enable forecasting-oriented applications. Extending long-range spatial modeling toward spatio-temporal representation learning represents a natural and necessary progression for advancing robust Arctic sea ice monitoring systems.

Data

- **Operational Integration and Human-in-the-Loop Systems.** The use of high-resolution SAR imagery motivates the generation of finer-grained annotations beyond coarse polygonal charts. A practical extension of this work is the development of semi-automated annotation pipelines in which model predictions serve as initial estimates that ice analysts can refine. See progress on this direction in Appendix A.

Training:

- **Integration of SAR-IceFM and ITT.** A key future direction is the integration of large-scale self-supervised pre-training (SAR-IceFM) with region-aware supervised structuring (ITT). In such coherent representation learning pipeline:

  1. Self-supervised pre-training would establish transferable feature structures.
  2. Large-scale contextual modeling could be extended beyond binary ice-water segmentation to include concentration, stage of development, and floe size descriptors.

Applications:

- **Cross-Sensor Representation Alignment.** Current SAR-IceFM pre-training is predominantly based on Sentinel-1 data. A natural extension is the development of cross-sensor foundation models capable of aligning representations across RADARSAT-2, RCM, and Sentinel-1. This would involve investigating feature-level alignment strategies, calibration during pre-training, and domain adaptation during fine-tuning to improve cross-sensor generalization. Such work would directly address sensor-dependent data distributions and move toward a unified SAR representation space for Arctic monitoring.

- **Domain Generalization Across Arctic Regions.** Sea ice characteristics vary substantially across Arctic basins due to differences in ocean circulation and seasonal dynamics. Regions such as the Beaufort Sea, Hudson Bay, and Baffin Bay exhibit distinct ice morphology, floe fragmentation patterns, and melt–freeze cycles, which can lead to distribution shifts in SAR backscatter and spatial texture. As a result, models trained on data from one region may not generalize reliably to others. Future work should investigate domain generalization strategies that enable feature representations to remain robust across geographically distinct Arctic environments.

References

[1] Tyler G. R. Reid, Todd Walter, Robert Guinness, Sarang Thombre, Heidi Kuusniemi, and Norvald Kjerstad. *Navigation in the Arctic*, pages 1947–1970. Wiley-IEEE Press, 2021.

[2] Xinwei Chen, Muhammed Patel, Fernando J. Pena Cantu, Jinman Park, Javier Noa Turnes, Mingzhe Jiang, Linlin Xu, K. Andrea Scott, David A. Clausi, and Weimin Huang. The influence of input variable selection on deep learning-based sea ice parameter inversion from multi-sensor satellite data. In *OCEANS 2023 - MTS/IEEE U.S. Gulf Coast*, pages 1–4, 2023.

[3] X. Chen, M. Patel, F. J. Pena Cantu, J. Park, J. Noa Turnes, L. Xu, K. A. Scott, and D. A. Clausi. Mmseaice: a collection of techniques for improving sea ice mapping with a multi-task model. *The Cryosphere*, 18(4):1621–1632, 2024.

[4] A. Stokholm, J. Buus-Hinkler, T. Wulf, A. Korosov, R. Saldo, L. T. Pedersen, D. Arthurs, I. Dragan, I. Modica, J. Pedro, A. Debien, X. Chen, M. Patel, F. J. P. Cantu, J. N. Turnes, J. Park, L. Xu, K. A. Scott, D. A. Clausi, Y. Fang, M. Jiang, S. Taleghanidoozdoozan, N. C. Brubacher, A. Soleymani, Z. Gousseau, M. Smaczny, P. Kowalski, J. Komorowski, D. Rijlaarsdam, J. N. van Rijn, J. Jakobsen, M. S. J. Rogers, N. Hughes, T. Zagon, R. Solberg, N. Longépé, and M. B. Kreiner. The autoice challenge. *The Cryosphere*, 18(8):3471–3494, 2024.

[5] Christine König, Thomas König, Suman Singha, Anja Frost, and Sven Jacobsen. Combined use of space borne optical and sar data to improve knowledge about sea ice for shipping. *Remote Sensing*, 13(23), 2021.

[6] Hegang Ji, Jian Zhang, Junjie Liu, Shi Hua, Renwei Liu, and A.T. Bekker. Combined effects of thermal stress and ice impacts on polar ship hulls: A numerical study. *Applied Ocean Research*, 160:104654, 2025.

[7] Delphine Lannuzel, Letizia Tedesco, Maria Van Leeuwe, Karley Campbell, Hauke Flores, Bruno Delille, Lisa Miller, Jacqueline Stefels, Philipp Assmy, Jeff Bowman, et al. The future of arctic sea-ice biogeochemistry and ice-associated ecosystems. *Nature Climate Change*, 10(11):983–992, 2020.

[8] John A Richards and JA Richards. *Remote sensing digital image analysis*, volume 3. Springer, 1999.

[9] Swapan Kumar Saha. *Aperture synthesis: methods and applications to optical astronomy*. Springer Science & Business Media, 2010.

[10] Dipankar Mandal, Avik Bhattacharya, and Yalamanchili Subrahmanyeswara Rao. *Radar Remote Sensing for Crop Biophysical Parameter Estimation*. Springer Remote Sensing/Photogrammetry. Springer Singapore, Singapore, 2021.

[11] Sebastian Gerland, David Barber, Walt Meier, Christopher J. Mundy, Marika Holland, Stefan Kern, Zhijun Li, Christine Michel, Donald K. Perovich, and Takeshi Tamura. Essential gaps and uncertainties in the understanding of the roles and functions of Arctic sea ice. *Environmental Research Letters*, 14(4):043002, apr 2019.

[12] David Anthony Clausi. Texture Segmentation of SAR Sea Ice Imagery. *Methods*, 37(2):176, 1996.

[13] D.A. Clausi and M.E. Jernigan. A fast method to determine co-occurrence texture features. *IEEE Transactions on Geoscience and Remote Sensing*, 36(1):298–300, 1998.

[14] Natalia Yu. Zakhvatkina, Vitaly Yu. Alexandrov, Ola M. Johannessen, Stein Sandven, and Ivan Ye. Frolov. Classification of sea ice types in envisat synthetic aperture radar images. *IEEE Transactions on Geoscience and Remote Sensing*, 51(5):2587–2600, 2013.

[15] Mingzhe Jiang, David A. Clausi, and Linlin Xu. Sea-Ice Mapping of RADARSAT-2 Imagery by Integrating Spatial Contexture With Textural Features. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:7964–7977, 2022.

[16] Stuart Geman and Christine Graffigne. Markov random field image models and their applications to computer vision. In *Proceedings of the international congress of mathematicians*, volume 1, pages 1496–1517. Berkeley, CA, 1986.

[17] Huawu Deng and D.A. Clausi. Unsupervised segmentation of synthetic aperture Radar sea ice imagery using a novel Markov random field model. *IEEE Transactions on Geoscience and Remote Sensing*, 43(3):528–538, mar 2005.

[18] Steven Leigh, Zhijie Wang, and David A. Clausi. Automated ice-water classification using dual polarization SAR satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 52(9):5529–5539, 2014.

[19] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2323, 1998.

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[21] Hugo Boulze, Anton Korosov, and Julien Brajard. Classification of Sea Ice Types in Sentinel-1 SAR Data Using Convolutional Neural Networks. *Remote Sensing*, 12(13):2165, jul 2020.

[22] Huan Wang, Qian Hu, Jianning Chi, Chengdong Wu, and Xiaosheng Yu. Multi-receptive-fields convolutional network for remote sensing images super-resolution. In *2021 33rd Chinese Control and Decision Conference (CCDC)*, pages 1525–1530, 2021.

[23] 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 *ICLR 2021 - 9th International Conference on Learning Representations*, oct 2021.

[24] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[25] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[26] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine*

Learning, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020.

[27] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[28] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc., 2020.

[29] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, June 2022.

[30] Javier Noa Turnes, Mingzhe Jiang, Saeid Taleghanidoozdoozan, Linlin Xu, and David A. Clausi. Itt: Long-range spatial dependencies for sea ice semantic segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18:13296–13309, 2025.

[31] Xinwei Chen, Muhammed Patel, Linlin Xu, Yuhao Chen, K. Andrea Scott, and David A. Clausi. A weakly supervised learning approach for sea ice stage of development classification from ai4arctic sea ice challenge dataset. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–1, 2025.

[32] Tore Wulf, Jørgen Buus-Hinkler, Suman Singha, Nishka Dasgupta, Athanasios Athanasiadis, and Matilde Brandt Kreiner. A decade of sea ice concentration retrieved from sentinel-1. *Remote Sensing of Environment*, 337:115252, 2026.

[33] Tore Wulf, Jørgen Buus-Hinkler, Suman Singha, and Matilde Brandt Kreiner. Operational sar-based sea ice concentration products for copernicus marine service. In *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 68–71, 2024.

[34] Yanling Han, Yekun Liu, Zhonghua Hong, Yun Zhang, Shuhu Yang, and Jing Wang. Sea ice image classification based on heterogeneous data fusion and deep learning. *Remote Sensing*, 13(4):1–20, feb 2021.

[35] Larissa Pizzolato, Stephen E.L. Howell, Chris Derksen, Jackie Dawson, and Luke Copland. Changing sea ice conditions and marine transportation activity in Canadian Arctic waters between 1990 and 2012. *Climatic Change*, 123(2):161–173, 2014.

[36] Dimitrios Dalaklis, Megan L. Drewniak, and Jens Uwe Schröder-Hinrichs. Shipping operations support in the "High North": examining availability of icebreakers along the Northern Sea Route. *WMU Journal of Maritime Affairs*, 17(2):129–147, 2018.

[37] Junqian Wang, Claude R. Duguay, David A. Clausi, Véronique Pinard, and Stephen E.L. Howell. Semi-automated classification of Lake Ice Cover using dual polarization RADARSAT-2 imagery. *Remote Sensing*, 10(11), 2018.

[38] Rudolf Ressel, Suman Singha, Susanne Lehner, Anja Rösel, and Gunnar Spreen. Investigation into different polarimetric features for sea ice classification using x-band synthetic aperture radar. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(7):3131–3143, 2016.

[39] Qiyao Yu and David A. Clausi. IRGS: Image segmentation using edge penalties and region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(12):2126–2139, 2008.

[40] Mingzhe Jiang, Linlin Xu, and David A. Clausi. Sea Ice–Water Classification of RADARSAT-2 Imagery Based on Residual Neural Networks (ResNet) with Regional Pooling. *Remote Sensing*, 14(13):3025, 2022.

[41] Yi-Ran Wang and Xiao-Ming Li. Arctic sea ice cover data from spaceborne synthetic aperture radar by deep learning. *Earth System Science Data*, 13(6):2723–2742, jun 2021.

[42] Yibin Ren, Xiaofeng Li, Xiaofeng Yang, and Huan Xu. Development of a Dual-Attention U-Net Model for Sea Ice and Open Water Classification on SAR Images. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.

[43] Mingzhe Jiang, Xinwei Chen, Linlin Xu, and David A. Clausi. Semi-Supervised Sea Ice Classification of SAR Imagery Based on Graph Convolutional Network. *International Geoscience and Remote Sensing Symposium (IGARSS)*, 2022-July:1031–1034, 2022.

[44] Xinwei Chen, K Andrea Scott, Mingzhe Jiang, Yuan Fang, Linlin Xu, and David A Clausi. Sea ice classification with dual-polarized SAR imagery: a hierarchical pipeline. In *Proceedings - 2023 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACVW 2023*, pages 224–232, 2023.

[45] Rui Huang, Changying Wang, Jinhua Li, and Yi Sui. DF-UHRNet: A Modified CNN-Based Deep Learning Method for Automatic Sea Ice Classification from Sentinel-1A/B SAR Images. *Remote Sensing*, 15(9), 2023.

[46] Ali Radman, Masoud Mahdianpari, Brian Brisco, Bahram Salehi, and Fariba Mohammadimanesh. Dual-Branch Fusion of Convolutional Neural Network and Graph Convolutional Network for PolSAR Image Classification. *Remote Sensing*, 15(1):75, 2023.

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017-December(Nips):5999–6009, 2017.

[48] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, mar 2021.

[49] Yundong Zhang, Huiye Liu, and Qiang Hu. Transfuse: Fusing transformers and cnns for medical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2021*, pages 14–24, Cham, 2021. Springer.

[50] Liang Gao, Hui Liu, Minhang Yang, Long Chen, Yaling Wan, Zhengqing Xiao, and Yurong Qian. STransFuse: Fusing Swin Transformer and Convolutional Neural Network for Remote Sensing Image Semantic Segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:10990–11003, 2021.

[51] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H.S. Torr, and Li Zhang. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6877–6886. IEEE, jun 2021.

[52] Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, Matthew Lungren, Lei Xing, Le Lu,

Alan L. Yuille, and Yuyin Zhou. Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, 91:103280, 2024.

[53] Kun Yuan, Shaopeng Guo, Ziwei Liu, Aojun Zhou, Fengwei Yu, and Wei Wu. Incorporating convolution designs into visual transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 579–588, 2021.

[54] Rui Li, Shunyi Zheng, Ce Zhang, Chenxi Duan, Jianlin Su, Libo Wang, and Peter M. Atkinson. Multiattention Network for Semantic Segmentation of Fine-Resolution Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 2022.

[55] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11966–11976. IEEE, 2022.

[56] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R. Roth, and Daguang Xu. UNETR: Transformers for 3D Medical Image Segmentation. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1748–1758. IEEE, jan 2022.

[57] Xin He, Yong Zhou, Jiaqi Zhao, Di Zhang, Rui Yao, and Yong Xue. Swin Transformer Embedding UNet for Remote Sensing Image Semantic Segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 2022.

[58] Nicolae-Cătălin Ristea, Andrei Anghel, and Mihai Datcu. Sea ice segmentation from sar data by convolutional transformer networks. In *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, pages 168–171, 2023.

[59] Mohsen Ghanbari, David A. Clausi, and Linlin Xu. CP-IRGS: A Region-Based Segmentation of Multilook Complex Compact Polarimetric SAR Data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:6559–6571, 2021.

[60] C. E. Livingstone, I. Sikaneta, C. Gierull, S. Chiu, and P. Beaulne. Radarsat-2 system and mode description. Technical report, Defence Research and Development Canada, 2006.

[61] David A Clausi. An analysis of co-occurrence texture statistics as a function of grey level quantization. *Canadian Journal of Remote Sensing*, 28(1):45–62, 2002.

[62] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[63] Xinwei Chen, K. Andrea Scott, Linlin Xu, Mingzhe Jiang, Yuan Fang, and David A. Clausi. Uncertainty-Incorporated Ice and Open Water Detection on Dual-Polarized SAR Sea Ice Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023.

[64] Xiaoyu Sun, Xi Zhang, Weimin Huang, Zongjun Han, Xinrong Lyu, and Peng Ren. Sea Ice Classification Using Mutually Guided Contexts. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–19, 2023.

[65] Prabhishek Singh, Manoj Diwakar, Achyut Shankar, Raj Shree, and Manoj Kumar. A Review on SAR Image and its Despeckling. *Archives of Computational Methods in Engineering*, 28(7):4633–4653, 2021.

[66] Zihao Wang and Lei Wu. Theoretical analysis of the inductive biases in deep convolutional networks. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 74289–74338, Red Hook, NY, USA, 2023. Curran Associates, Inc.

[67] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, Red Hook, NY, USA, 2016. Curran Associates, Inc.

[68] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[69] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-Decem, pages 770–778, 2016.

[70] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[71] Pike Msonda, Sait Ali Uymaz, and Seda Soğukpinar Karaağaç. Spatial pyramid pooling in deep convolutional networks for automatic tuberculosis diagnosis. *Traitement du Signal*, 37(6):1075–1084, 2020.

[72] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4794–4803, June 2022.

[73] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1290–1299, June 2022.

[74] Rafael C. Gonzales and Richard E. Woods. *Digital Image Processing*. Prentice Hall, 2nd edition, 2002.

[75] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[76] Thorsten Markus, Julienne C. Stroeve, and Jeffrey Miller. Recent changes in Arctic sea ice melt onset, freezeup, and melt season length. *Journal of Geophysical Research: Oceans*, 114(12):1–14, 2009.

[77] Samira Alkaee Taleghan, Andrew P. Barrett, Walter N. Meier, and Farnoush Banaei-Kashani. Icebench: A benchmark for deep-learning-based sea-ice type classification. *Remote Sensing*, 17(9), 2025.

[78] X. Chen, M. Patel, F. J. Pena Cantu, J. Park, J. Noa Turnes, L. Xu, K. A. Scott, and D. A. Clausi. Mmseaice: a collection of techniques for improving sea ice mapping with a multi-task model. *The Cryosphere*, 18(4):1621–1632, 2024.

[79] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter

Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2022.

[80] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, October 2023.

[81] Le Wang, Chunyuan Diao, George Xian, Dameng Yin, Ying Lu, Shengyuan Zou, and Tyler A Erickson. A summary of the special issue on remote sensing of land change science with Google earth engine, 2020.

[82] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[83] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[84] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al.

Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[85] Zirui Xu, Raphael Tang, Mike Bianco, Qi Zhang, Rishi Madhok, Nikolaos Karianakis, and Fuxun Yu. Geospatial foundational embedder: Top-1 winning solution on earthvision embed2scale challenge (cvpr 2025). *arXiv preprint arXiv:2509.06993*, 2025.

[86] Lucas Prado Osco, Qiusheng Wu, Eduardo Lopes De Lemos, Wesley Nunes Gonçalves, Ana Paula Marques Ramos, Jonathan Li, and José Marcato Junior. The segment anything model (SAM) for remote sensing applications: From zero to one shot. *International Journal of Applied Earth Observation and Geoinformation*, 124:103540, 2023.

[87] Johannes Jakubik, Sujit Roy, CE Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Gabby Nyirjesy, Blair Edwards, et al. Foundation models for generalist geospatial artificial intelligence. *arXiv preprint arXiv:2310.18660*, 2023.

[88] Xiao Xiang Zhu, Zhitong Xiong, Yi Wang, Adam J Stewart, Konrad Heidler, Yuanyuan Wang, Zhenghang Yuan, Thomas Dujardin, Qingsong Xu, and Yilei Shi. On the foundations of earth foundation models. *Communications Earth & Environment*, 2026.

[89] Christopher F Brown, Michal R Kazmierski, Valerie J Pasquarella, William J Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Shelhamer, Estefania Lahera, Olivia Wiles, Simon Ilyushchenko, et al. Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *arXiv preprint arXiv:2507.22291*, 2025.

[90] Clay Foundation. Clay foundation model: Github repository. 2026. Accessed 2026-02-04.

[91] J. Seston, W. D. Harcourt, G. Leontidis, B. Rea, M. Spagnolo, and L. McWhinnie. Leveraging self-supervised learning for sea ice segmentation in the arctic to reduce on labelling. In *EGU General Assembly 2025*, pages EGU25–18799, Vienna, Austria, 2025. EGU.

[92] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal

Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc., 2020.

[93] Haipeng Qin, Xinwei Chen, Muhammad Patel, David A. Clausi, Linlin Xu, and Weimin Huang. Masked autoencoders for enhancing sar sea ice image segmentation. In *2025 IEEE 20th International Symposium on Antenna Technology and Applied Electromagnetics (ANTEM)*, pages 148–151, 2025.

[94] Samira Alkaee Taleghan, Morteza Karimzadeh, Andrew P. Barrett, Walter N Meier, and Farnoush Banaei-Kashani. Ice-fmbench a foundation model benchmark for sea ice type segmentation. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Polar Data Science*, PoIDS '25, New York, NY, USA, 2025. Association for Computing Machinery.

[95] Yi Wang, Hugo Hernández Hernández, Conrad M Albrecht, and Xiao Xiang Zhu. Feature guided masked autoencoder for self-supervised learning in remote sensing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18:321–336, 2025.

[96] Matt Allen, Francisco Dorr, Joseph A. Gallego-Mejia, Laura Martínez-Ferrer, Anna Jungbluth, Freddie Kalaitzis, and Raúl Ramos-Pollán. Large Scale Masked Autoencoding for Reducing Label Requirements on SAR Data. pages 1–12, sep 2024.

[97] Weijie Li, Wei Yang, Tianpeng Liu, Yuenan Hou, Yuxuan Li, Zhen Liu, Yongxiang Liu, and Li Liu. Predicting gradient is better: Exploring self-supervised learning for sar atr with a joint-embedding predictive architecture. *ISPRS Journal of Photogrammetry and Remote Sensing*, 218:326–338, 2024.

[98] Yi Wang, Conrad M Albrecht, and Xiao Xiang Zhu. Self-supervised vision transformers for joint sar-optical representation learning. In *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 139–142, 2022.

[99] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9912–9924. Curran Associates, Inc., 2020.

[100] Xian Sun, Peijin Wang, Wanxuan Lu, Zicong Zhu, Xiaonan Lu, Qibin He, Junxi Li, Xuee Rong, Zhujun Yang, Hao Chang, Qinglin He, Guang Yang, Ruiping Wang, Jiwen Lu, and Kun Fu. Ringmo: A remote sensing foundation model with masked image modeling. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–22, 2023.

[101] Di Wang, Qiming Zhang, Yufei Xu, Jing Zhang, Bo Du, Dacheng Tao, and Liangpei Zhang. Advancing plain vision transformer toward remote sensing foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023.

[102] Jørgen Buus-Hinkler, Tore Wulf, Andreas Rønne Stokholm, Anton Korosov, Roberto Saldo, Leif Toudal Pedersen, and Morten Brandt Kreiner. Ai4arctic sea ice challenge dataset (version 2). Technical University of Denmark, DTU, 2022. https://doi.org/10.11583/DTU.c.6244065.v2.

[103] Juha Karvonen, Jouni Vainio, Marika Marnela, Patrick Eriksson, and Tuomas Niskanen. A comparison between high-resolution eo-based and ice analyst-assigned sea ice concentrations. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(4):1799–1807, 2015.

[104] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[105] Andrzej Kucik and Andreas Stokholm. AI4SeaIce: selecting loss functions for automated SAR sea ice concentration charting. *Scientific Reports*, 13(1):1–10, 2023.

[106] Caleb Robinson, Le Hou, Kolya Malkin, Rachel Soobitsky, Jacob Czawlytko, Bistra Dilkina, and Nebojsa Jojic. Large scale high-resolution land cover mapping with multi-resolution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[107] Donald K. Perovich and Jacqueline A. Richter-Menge. Loss of sea ice in the arctic. *Annual Review of Marine Science*, 1:417–441, 2009.

[108] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[109] David H Hubel and Torsten N Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243, 1968.

[110] Kunihiko Fukushima, Sei Miyake, and Takayuki Ito. Neocognitron: A neural network model for a mechanism of visual pattern recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13(5):826–834, sep 1983.

[111] John Paul Mueller and Luca Massaron. *Data Science Programming All-in-One For Dummies*, volume 1. John Wiley I& Sons, 2019.

[112] G. Bebis and M. Georgiopoulos. Feed-forward neural networks. *IEEE Potentials*, 13(4):27–31, oct 1994.

[113] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *International Conference on Learning Representations*, 2020.

[114] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. In *ICLR*, 2023.

[115] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurélien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels, 2010.

[116] Stan Z Li. *Markov random field modeling in image analysis*. Springer Science & Business Media, 2009.

[117] Devis Tuia, Claudio Persello, and Lorenzo Bruzzone. Domain Adaptation for the Classification of Remote Sensing Data: An Overview of Recent Advances. *IEEE Geoscience and Remote Sensing Magazine*, 4(2):41–57, jun 2016.

[118] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional sequence to sequence learning. *34th International Conference on Machine Learning, ICML 2017*, 70:1243–1252, 2017.

[119] Yifang Yin, Zhenguang Liu, Ying Zhang, Sheng Wang, Rajiv Ratn Shah, and Roger Zimmermann. GPS2Vec: Towards generating worldwide GPS embeddings. *GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems*, pages 416–419, 2019.

[120] Konstantin Klemmer, Nathan S. Safir, and Daniel B. Neill. Positional encoder graph neural networks for geographic data. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 1379–1389. PMLR, 25–27 Apr 2023.

[121] Günter Rote. The degree of convexity. In *29th European Workshop on Computational Geometry*, pages 69–72, 2013.

[122] GH Born, JA Dunne, and DB Lame. Seasat mission overview. *Science*, 204(4400):1405–1406, 1979.

[123] B. Ramsay, M. Manore, L. Weir, K. Wilson, and D. Bradley. Use of radarsat data in the canadian ice service. *Canadian Journal of Remote Sensing*, 24(1):36–42, 1998.

[124] Robert M. Haralick, K. Shanmugam, and Its'Hak Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3(6):610–621, 1973.

[125] Natalia Zakhvatkina, Vladimir Smirnov, and Irina Bychkova. Satellite SAR data-based sea ice classification: An overview. *Geosciences (Switzerland)*, 9(4):3–5, 2019.

[126] J. Louet and S. Bruzzi. Envisat mission and system. In *IEEE 1999 International Geoscience and Remote Sensing Symposium. IGARSS'99 (Cat. No.99CH36293)*, volume 3, pages 1680–1682 vol.3, 1999.

[127] Paul J Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990.

[128] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, page 144–152, New York, NY, USA, 1992. Association for Computing Machinery.

[129] Raghu G. Raj, Alan C. Bovik, and Wilson S. Geisler. Non-Stationarity Detection in Natural Images. In *2007 IEEE International Conference on Image Processing*, pages III – 305–III – 308. IEEE, 2007.

[130] Kun Yang, Haiyan Li, William Perrie, Randall Kenneth Scharien, Jin Wu, Menghao Zhang, and Fan Xu. Fine Resolution Classification of New Ice, Young Ice, and First-Year Ice Based on Feature Selection from Gaofen-3 Quad-Polarization SAR. *Remote Sensing*, 15(9):2399, may 2023.

[131] Yiru Lu, Biao Zhang, and William Perrie. Arctic Sea Ice and Open Water Classification from Spaceborne Fully Polarimetric Synthetic Aperture Radar. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023.

[132] Danyang Tu, Xiongkuo Min, Huiyu Duan, Guodong Guo, Guangtao Zhai, and Wei Shen. Iwin: Human-object interaction detection via transformer with irregular windows. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 87–103, Cham, 2022. Springer Nature Switzerland.

[133] Thorsten Markus, Julienne C. Stroeve, and Jeffrey Miller. Recent changes in arctic sea ice melt onset, freezeup, and melt season length. *Journal of Geophysical Research: Oceans*, 114(C12), 2009.

[134] M. Anderson Bliss, A. C. and S. Drobot. Snow melt onset over arctic sea ice from smmr and ssm/i-ssmis brightness temperatures, version 5, 2022.

[135] NASA Earth Sciences. Arctic sea ice melt, 2025. Accessed: 2025-02-25.

APPENDICES

Appendix A

ArcticScope: Technology Transfer and Operational Deployment of ITT at the Canadian Ice Service

Motivation

Following the development of the Irregular Tokens on Transformers (ITT) framework in Chapter 2, a collaborative technology transfer initiative was launched with the Canadian Ice Service (CIS) to evaluate its operational integration for ice-water classification using RADARSAT Constellation Mission (RCM) imagery. The objective of this project is to augment analyst workflows by generating high-resolution, consistent ice-water maps that can be efficiently refined through expert interaction. The initiative aims to establish an operational pipeline at CIS capable of (i) executing advanced SAR-based deep learning models, (ii) enabling expert-driven correction, and (iii) generating large-scale annotated datasets for iterative model refinement.

Human-in-the-Loop Active Learning Framework

The operational deployment follows an implementation plan centered on a human-in-the-loop active learning process (see Figure A.1). In this workflow, SAR scenes are first pre-processed and subsequently passed to trained deep learning models, including ITT variants. The resulting predictions are then visualized and overlaid on the SAR imagery within ArcticScope, enabling CIS analysts to perform quality assessment and region-level corrections. Once validated or modified, the corrected annotations are systematically stored and incorporated into subsequent retraining cycles. The updated models are thereafter

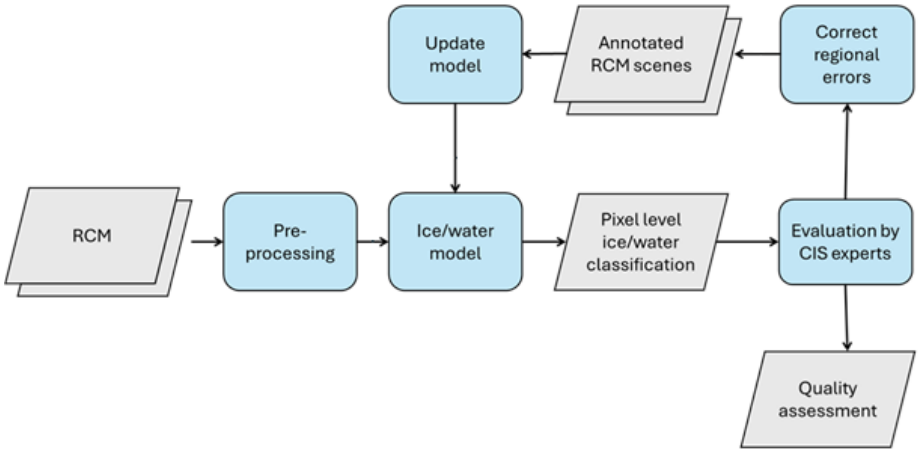

Figure A.1: Human-in-the-loop active learning workflow for operational ice-water classification at CIS.

redeployed on new scenes, closing the feedback loop and progressively enhancing model performance over time.

ArcticScope

ArcticScope is a desktop application developed as an operational extension of the research presented in this thesis, enabling practical evaluation and refinement of SAR-based sea ice classification models. A screenshot of the ArcticScope interface is provided in Figure A.2, illustrating the visualization panel, prediction overlay, and annotation controls. It currently provides the following capabilities:

- **Scene visualization:** Loading RCM products in both Earth and sensor geometry, image enhancement controls, zooming and panning functionality, and a minimap with geolocation readout.
- **Model inference:** Automatic generation of ice-water predictions upon scene loading, with support for multiple model configurations.
- **Annotation:** Polygon-based region selection for correction, label assignment (ice, water, and auxiliary classes), and integrated unsupervised local segmentation (IRGS) with adjustable granularity to assist boundary refinement.
- **Evaluation and logging:** Per-scene qualitative and quantitative evaluation, along with structured saving of annotations and analyst notes (e.g., JSON format).

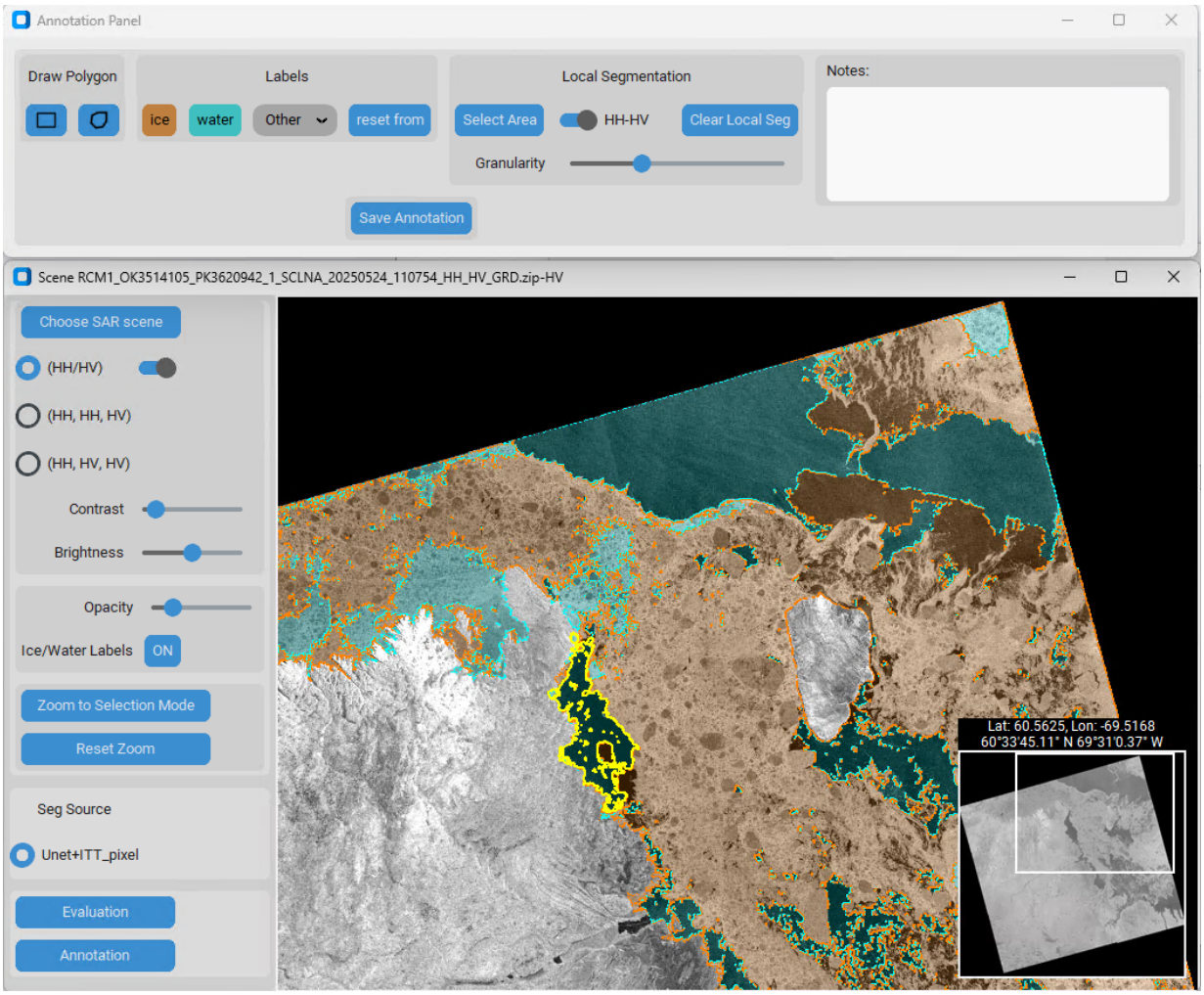

Figure A.2: ArcticScope interface showing SAR visualization, prediction overlay, and annotation tools.

Appendix B

Visualizations

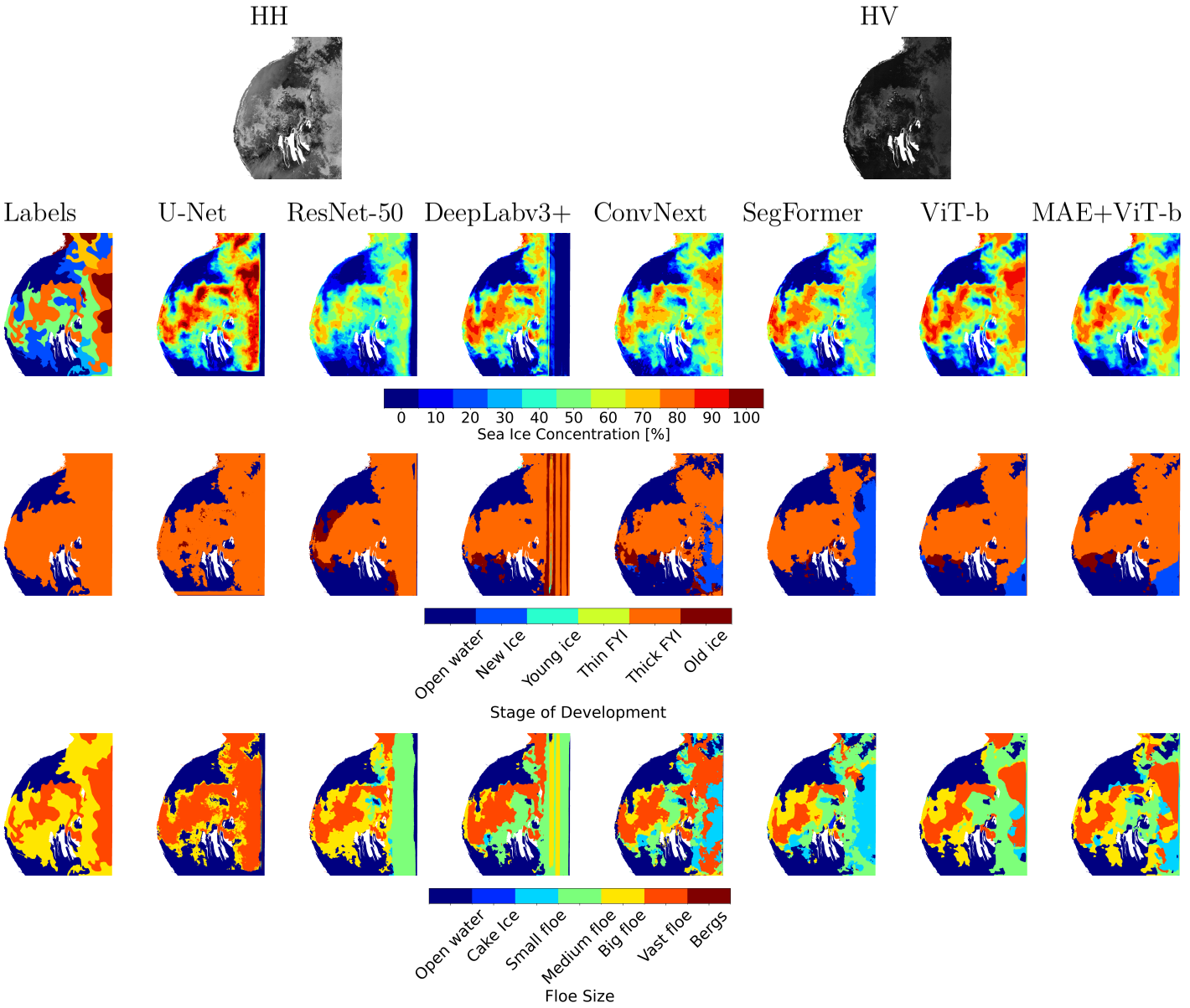

Figure B.1: Qualitative comparison of SAR-based sea ice predictions across multiple architectures, including supervised models trained on 20% of the data and MAE-pretrained models, which are pre-trained on 80% and fine-tuned on the remaining 20%. From top to bottom, the rows show the SAR input image (July 07, 2018, Foxe Basin), SIC, SOD, and FLOE.