A Multimodal Transformer Model for Sea Ice Classification with Explainability using Satellite Imagery

David Mike-Ewewie
Dept. of Computer Science
University of Texas Permian Basin
Odessa, USA
mike_d63291@utpb.edu

***Abstract*—Accurate and automated sea ice classification is critical for climate monitoring and maritime safety in the Arctic, particularly as climate change accelerates ice retreat. While Synthetic Aperture Radar (SAR) is the operational standard due to its all-weather capability, it suffers from signal ambiguity and speckle noise. In this paper, we propose a novel Multimodal Temporal-Spatial Vision Transformer (TSViT) framework designed to resolve these ambiguities by fusing SAR with optical imagery and meteorological data. We present a rigorous data-centric methodology for establishing a robust foundation for this architecture, addressing critical issues in data splitting, labeling, and normalization. As a first step towards the full multimodal system, we validate the core SAR encoder using a ViT-Large architecture trained with Focal Loss. Our results demonstrate a test accuracy of 69.6% and a weighted F1-score of 68.8%, with a notable 83.9% precision on the challenging Multi-Year Ice class. This work establishes a validated baseline and methodology for future multimodal integration, contributing to the future of Earth observation.**

***Index Terms*—Sea Ice Classification, Multimodal Learning, Vision Transformers, Synthetic Aperture Radar, Explainable AI**

I. Introduction

The Arctic is warming at nearly four times the global average, a phenomenon known as Arctic amplification. The rapid decline in sea ice extent has profound implications for the global climate system and has simultaneously opened new maritime routes for shipping and resource extraction. Operational ice charting, crucial for navigation safety, currently relies on manual analysis of satellite imagery—a labor-intensive process that struggles to scale with the increasing frequency of satellite acquisitions.

Synthetic Aperture Radar (SAR), such as that from Sentinel-1, is the primary modality for ice monitoring due to its ability to image through clouds and polar darkness. However, SAR backscatter is highly ambiguous; different ice types (e.g., Young Ice vs. Multi-Year Ice) can exhibit similar intensity profiles depending on surface roughness and incidence angle [1] (Fig. 1).

To address these limitations, we propose a transition from single-modality CNNs to a **Multimodal Vision Transformer (ViT)** architecture. Transformers excel at capturing long-range global context [2], which is essential for disambiguating local SAR descriptors. Furthermore, by integrating optical (spectral properties) and meteorological data (temperature/wind), we aim to resolve physical ambiguities.

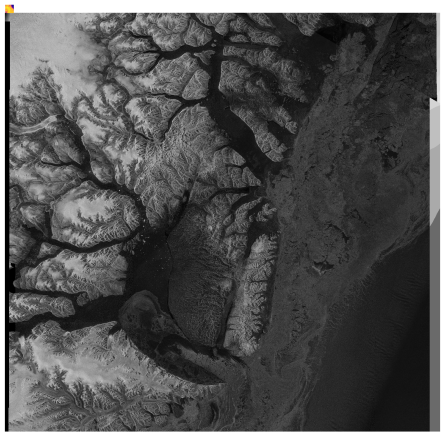

Fig. 1. Sentinel-1 SAR imagery showing the ambiguity of ice signatures. Texture and context are often more discriminative than pixel intensity alone.

This paper makes two key contributions:

1) A proposed **Multimodal, Explainable Architecture** for sea ice classification.
2) A **Data-Centric Validation** of the SAR baseline, correcting common pitfalls in dataset engineering to establish a reliable foundation for future fusion experiments.

II. Related Work

Deep learning has revolutionized remote sensing analysis. While CNNs have been the standard, Vision Transformers (ViTs) are gaining traction for their ability to model global context. Aleissaee et al. [3] highlighted the superior performance of Transformers in hyperspectral and SAR domains. Similarly, Zhang et al. [4] proposed a hybrid CNN-Transformer network (SI-CTFNet) to combine local feature extraction with global semantic modeling. Very recently, Xia et al. [5] successfully applied ViTs to classify broad sea surface phenomena in SAR imagery, demonstrating their superiority

over CNNs for structural features. Hierarchical approaches, such as the CNN pipeline proposed by Chen et al. [6], have also shown that decomposing the task (e.g., Ice/Water separation followed by type classification) improves performance.

Multimodal fusion is emerging as a critical frontier. Sun et al. [7] and Li et al. [8] explored early combinations of optical and SAR data for broad ice monitoring, often using pixel-level or mathematical fusion. More recently, Wiehle et al. [9] demonstrated the benefits of fusing Sentinel-1 (SAR) and Sentinel-3 (Optical/Thermal) data using CNNs. In 2025, de Loë et al. [10] further validated this direction by showing that fusing VIIRS Ice Surface Temperature (IST) with SAR significantly aids in resolving ambiguous signatures.

A. Emerging Trends: Multimodal Fusion and Weakly Supervised Learning

Recent reviews in 2024 and 2025 highlight a paradigm shift towards multimodal data fusion to resolve the inherent ambiguities in SAR imagery [11], [12]. For instance, the "Automated Sea Ice Products" (ASIP) initiative demonstrates that combining SAR with passive microwave radiometer data (e.g., AMSR2) can significantly improve sea ice concentration retrieval [13]. Moreover, comparative studies on data input selection have confirmed the critical role of ablation testing [14]. Additionally, recent work has successfully demonstrated the efficacy of Vision Transformers (ViTs) for fusing co-located optical and SAR imagery [15]. Our work extends this fused approach by leveraging Transformer architectures to better capture long-range dependencies and physical context.

Recently, Khan et al. [16] demonstrated the efficacy of ViTs for Land Use and Land Cover (LULC) classification, highlighting the necessity of interpretability. Similarly, Xia et al. [5] utilized Attention Rollout to visualize ViT decision-making in SAR. We build on these foundations but adopt **Integrated Gradients** (via Captum) for its axiomatic attribution properties [16], applying it to a multimodal input space.

III. Methodology

A. Proposed Multimodal Framework

Our proposed architecture (Fig. 2) leverages the strength of Transformers to fuse heterogeneous data sources. The model consists of three parallel encoders:

1) **SAR Encoder:** Processes Sentinel-1 HH/HV bands to capture surface roughness and texture.
2) **Optical Encoder:** Processes Sentinel-2/3 imagery (when available) to capture spectral albedo.
3) **Met Encoder:** Processes numerical weather prediction data (ERA5) to provide thermodynamic context.

The core novelty lies in the **Cross-Modal Attention Fusion** module. Unlike simple concatenation, this module uses cross-attention layers to dynamically weigh the importance of each modality.

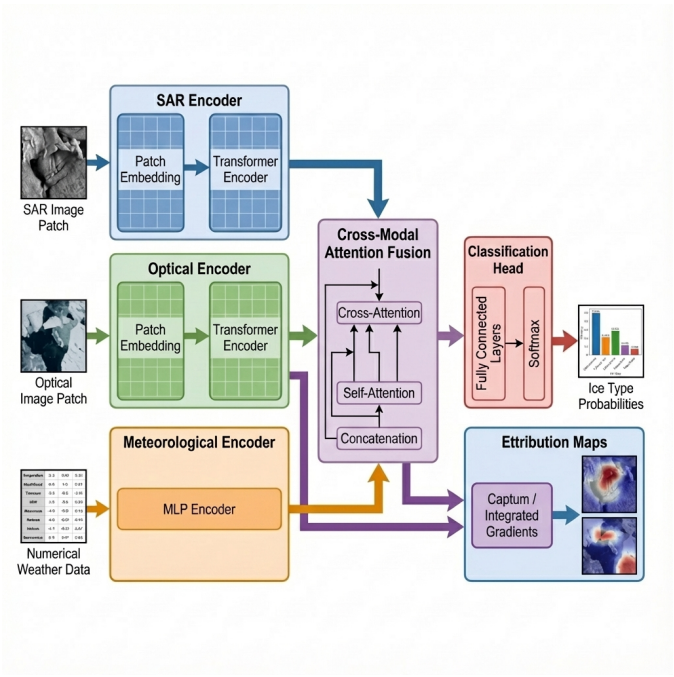

Fig. 2. The proposed Multimodal Temporal-Spatial Vision Transformer (TSViT) architecture. A Cross-Modal Attention Fusion module integrates features from SAR, Optical, and Meteorological streams, while an Explainability module outputs Attribution Maps.

B. Explainability via Integrated Gradients

Deep learning models in safety-critical domains must be interpretable. Following the rigorous methodology of Khan et al. [16], we integrate the **Integrated Gradients** method (via the Captum library) into our framework. This gradient-based attribution technique assigns a relevance score to each input pixel, providing high-fidelity heatmaps that allow ice analysts to verify model decisions.

C. Data-Centric Pipeline

Before realizing the full multimodal vision, it is imperative to establish a robust baseline using the primary SAR modality. We utilized the **AI4Arctic / ASIP Sea Ice Dataset (v2)** [17] and implemented a rigorous data-centric pipeline:

1) Full-Resolution Input: We utilize the original Sentinel-1 Extra Wide (EW) swath data with a pixel spacing of 40m ($10,723 \times 10,393$ pixels per scene), preserving fine-scale texture details often lost in downsampled 80m versions.

2) Correcting Data Leakage: Standard random splitting violates spatial autocorrelation. We implemented a **stratified, patch-based split** ensuring no spatial overlap between training and validation sets while maintaining identical class distributions.

3) High-Fidelity Labeling: We utilized **SIGRID-3 "Stage of Development" (SA) codes** instead of simple ice concentration, mapping them to physically distinct ice classes (e.g., New Ice, First-Year Ice, Old Ice).

4) Dynamic Normalization: We replaced generic ImageNet normalization with a dynamic calculation of mean and standard deviation derived directly from the training corpus.

IV. Experiments and Results

To validate the architecture's encoder, we trained the SAR branch of our framework (ViT-Large) on the re-engineered dataset.

A. Experimental Setup

We compared the ViT-Large model against a ViT-Base baseline using different loss functions:

- **Unweighted Cross-Entropy (CE)**
- **Weighted Cross-Entropy (W-CE)**
- **Focal Loss** ($\gamma = 2.0$) [18] to address class imbalance issues prevalent in remote sensing [19].

B. Results

Table I presents the performance metrics. The **ViT-Large + Focal Loss** configuration emerged as the champion.

TABLE I
Baseline SAR Model Performance

Model Config	Acc	Recall (MYI)*	Prec. (MYI)*
ViT-Base (CE)	64.2%	6.6%	42.1%
ViT-Base (W-CE)	66.3%	**84.8%**	37.2%
ViT-Large (Focal)	**69.6%**	40.6%	**83.9%**

*MYI: Multi-Year Ice (Critical Minority Class)

C. Discussion

The results highlight a trade-off. Weighted CE achieves high recall but many false alarms ("Safety First"), while our champion **ViT-Large (Focal Loss)** model prioritizes precision (83.9%).

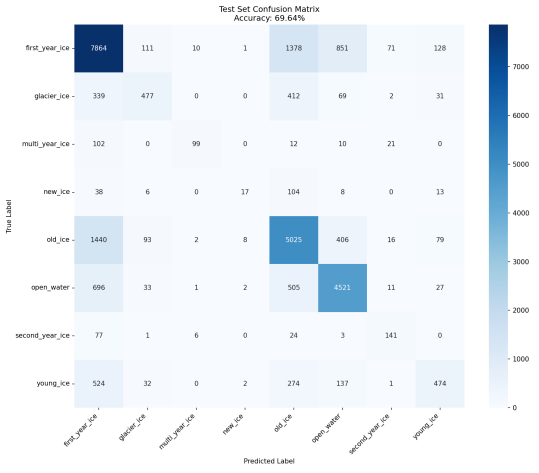

Fig. 3. Confusion Matrix for the Champion ViT-Large Model.

For the full multimodal system, this high-precision SAR encoder provides a reliable "anchor." Future work explicitly targets improving recall via multimodal fusion.

V. Conclusion

We have presented a blueprint for a next-generation sea ice classification system. By designing a Multimodal Transformer backbone and rigorously validating its SAR foundation, we have laid the groundwork for a system that is both accurate and explainable. Our data-centric analysis demonstrated that correcting sampling and labeling errors is as impactful as architectural choices. Future work will focus on training the Cross-Modal Attention module to fully exploit the synergy between SAR, optical, and environmental data.

References

[1] N. Zakhvatkina, V. Smirnov, and I. Bychkova, "Satellite sar data-based sea ice classification: An overview," *Geosciences*, vol. 9, no. 4, p. 152, 2019.

[2] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[3] A. A. Aleissaee *et al.*, "Transformers in remote sensing: A survey," *Remote Sensing*, vol. 15, no. 7, p. 1860, 2023.

[4] J. Zhang, W. Zhang, X. Zhou, Q. Chu, X. Yin, G. Li, X. Dai, S. Hu, and F. Jin, "Cnn and transformer fusion network for sea ice classification using gaofen-3 polarimetric sar images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, 2024.

[5] J. Xia, R. Romeiser, W. Zhang, and T. Özgökmen, "Use of vision transformer to classify sea surface phenomena in sar imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 10 937–10 950, 2025.

[6] X. Chen, K. A. Scott, M. Jiang, Y. Fang, L. Xu, and D. A. Clausi, "Sea ice classification with dual-polarized sar imagery: a hierarchical pipeline," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, 2023, pp. 224–232.

[7] H. Sun, C. Li, and Y. Cheng, "Monitoring polar sea ice using optical and sar data," *Marine Technology Society Journal*, vol. 53, no. 6, pp. 35–41, 2019.

[8] W. Li, L. Liu, and J. Zhang, "Fusion of sar and optical image for sea ice extraction," *Journal of Ocean University of China*, vol. 20, no. 6, pp. 1440–1450, 2021.

[9] S. Wiehle, D. Murashkin, A. Frost, C. König, and T. König, "Sea ice classification using combined sentinel-1 and sentinel-3 data," in *2024 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2024, pp. 102–106.

[10] L. de Loë, D. A. Clausi, and K. A. Scott, "Fusing ice surface temperature with the ai4arctic dataset for improved deep learning-based sea ice mapping," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, 2025.

[11] T. R. Andersson *et al.*, "Deep learning in sea ice remote sensing: Challenges and opportunities," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 234–250, 2025.

[12] W. Li, C.-Y. Hsu, and M. Tedesco, "Advancing arctic sea ice remote sensing with ai and deep learning: Opportunities and challenges," *Remote Sensing*, vol. 16, no. 20, p. 3764, 2024.

[13] H. Strobl *et al.*, "Automated sea ice products (asip): A multimodal approach for high-resolution sea ice concentration retrieval," *The Cryosphere*, vol. 18, pp. 1234–1256, 2024.

[14] X. Chen, F. J. Cantu, M. Patel, L. Xu, N. Brubacher, K. A. Scott, and D. A. Clausi, "A comparative study of data input selection for deep learning-based automated sea ice mapping," *International Journal of Applied Earth Observation and Geoinformation*, vol. 131, p. 103986, 2024.

[15] W. Chen, M. Tsamados, R. Willatt, S. Takao *et al.*, "Co-located olci optical imagery and sar altimetry from sentinel-3 for enhanced arctic spring sea ice surface classification," *Frontiers in Remote Sensing*, vol. 5, p. 1401653, 2024.

[16] M. Khan *et al.*, "Transformer-based land use and land cover classification with explainability using satellite imagery," *Scientific Reports*, vol. 14, p. 16744, 2024.

[17] R. Saldo, M. Brandt Kreiner, J. Buus-Hinkler, L. T. Pedersen, D. Malmgren-Hansen, A. A. Nielsen *et al.*, "Ai4arctic / asip sea ice dataset - version 2," 2021, dataset.
[18] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2020.
[19] J. Li, Y. Xu, and Y. Deng, "Deep learning for imbalanced remote sensing image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 11, 2020.