Use of Vision Transformer to Classify Sea Surface Phenomena in SAR Imagery

Junfei Xia, Roland Romeiser, *Senior Member, IEEE*, Wei Zhang, and Tamay Özgökmen

***Abstract*—The rapid advancement of satellite technology has led to a substantial increase in the volume of remote sensing data, particularly synthetic aperture radar (SAR) imagery, demanding efficient processing and analysis solutions. This study pioneers the application of vision transformers (ViTs) in classifying geophysical phenomena within SAR images, achieving higher recall (0.985), precision (0.984), and F1 scores (0.986) than existing methods and outperforming convolutional neural networks across multiple geophysical categories. We leveraged the single-polarization TenGeoP-SARwv dataset, comprising over 37 000 SAR vignettes, to train and evaluate the ViT models on a balanced subset, improving model generalizability by jointly using WV1 and WV2 images with advanced data augmentation techniques. In addition, our study is the first to apply a pretrained ViT model to a dataset with different polarizations and spatial resolutions—the AI4Arctic Sea Ice Challenge dataset—to rigorously assess model adaptability. Without extensive preprocessing, the ViT model achieved approximately 80% accuracy on HH-polarized images and 47% on HV-polarized images, underscoring the effect of polarization differences on classification performance. Notably, this study introduces the use of the attention mechanism in ViTs to elucidate model decision-making, providing interpretability by highlighting regions that influence predictions and revealing why the model succeeds or fails in specific cases. This comparative analysis also underscores the limitations of CNNs, which perform well on texture-based features but struggle with structural classifications, such as rain cells and oceanic fronts, where ViTs excel. These findings advance the application of deep learning in remote sensing, highlighting the robust adaptability of ViTs for diverse SAR imagery.**

***Index Terms*—Classification, deep learning (DL), synthetic aperture radar (SAR), sea surface events, vision transformer (ViT).**

I. Introduction

RECENT technological advancements have significantly expanded our capacity to collect data, particularly in the field of remote sensing [1]. One of the most vital tools in this domain is spaceborne synthetic aperture radar (SAR), which operates in the microwave spectral region and enables high-resolution sea surface backscatter data collection both day and night, under most weather conditions, unlike optical satellites, which are dependent on lighting conditions [2], [3]. SAR has been integral to numerous satellite missions, including ERS-1/2 (1991–2003) [4], [5], ENVISAT [6]/ASAR [7] (2002–2012), TerraSAR-X (2007–present) [8], RADARSAT-1/2 (1995–present) [9], [10], GaoFen-3 (2016–present) [11], and Sentinel-1 (2014–present) [12], which collectively generate hundreds of thousands of spatiotemporal images [13]. However, the large volume of data produced by these missions presents challenges in preprocessing, analysis, and interpretation. As remote sensing technologies continue to evolve, there is a growing need for robust, scalable processing techniques capable of handling this data's complexity and scale.

Classifying satellite images involves a multistep process, from feature extraction to categorization [14]. Traditional methods have included manual visual interpretation and spectral analysis [15], [16], [17]. However, recent advances in deep learning (DL) have shifted the focus toward automated image classification [18], [19], [20], [21], [22]. DL techniques offer powerful solutions for processing large datasets and extracting valuable insights from complex images.

Among DL methods, convolutional neural networks (CNNs) have become prominent in identifying sea surface phenomena from SAR imagery [23], [24]. Well-established CNN architectures, such as AlexNet, GoogLeNet, VGG, and ResNet, have significantly advanced oceanographic feature detection [25]. For instance, AlexNet, introduced in 2012, includes 60 million parameters across eight layers, while GoogLeNet, introduced in 2015, has 4 million parameters across 22 layers [26]. The VGGNet and ResNet models have also been instrumental in high-performance image classification [25], [27], but CNNs typically rely on localized pattern recognition, which limits their ability to capture global context within images [28], [29], [30], [31].

Recently, vision transformers (ViTs) have emerged as a promising alternative to CNNs [32], [33], [34], [35]. Unlike CNNs, ViTs leverage self-attention mechanisms, which capture long-range dependencies and contextual relationships within images [36]. This capability allows ViTs to offer a more comprehensive interpretation of image data, which is particularly valuable in SAR image classification, where contextual understanding is essential.

This study advances ViTs for SAR image classification through several key contributions as follows.

1) *Technical innovations in model customization:* We enhanced the standard ViT architecture by incorporating a

Received 2 December 2024; revised 6 March 2025; accepted 3 April 2025. Date of publication 8 April 2025; date of current version 30 April 2025. This work was supported in part by the Office of Naval Research under Grant N00014-20-1-2023 (MURI ML-SCOPE) to RSMAES, in part by the University of Miami, and in part by the Massachusetts Institute of Technology. *(Corresponding author: Junfei Xia; Wei Zhang.)*

The authors are with the Rosenstiel School of Marine, Atmospheric, and Earth Science, University of Miami, Miami, FL 33149 USA (e-mail: junfei.xia@earth.miami.edu; wei.zhang@earth.miami.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/JSTARS.2025.3558673, provided by the authors.

Digital Object Identifier 10.1109/JSTARS.2025.3558673

customized classifier layer tailored for SAR imagery. This architectural adjustment, combined with advanced data augmentation techniques (e.g., combining VV-polarized WV1 and WV2 images), optimized the model's ability to handle the unique characteristics of SAR data.

2) *Improved performance:* We achieved higher recall (0.985), precision (0.984), and F1 score (0.986) than previously reported methods in SAR classification. Our ViT model outperformed CNN benchmarks across various geophysical categories, particularly in categories that require complex pattern recognition, such as "Oceanic Fronts" and "Rain Cells."
3) *Insight into model decision-making:* By using the attention mechanism within ViTs, this study provides insights into how the model makes classification decisions. The attention maps reveal the critical regions influencing each prediction, offering interpretability in understanding the model's success and failure cases.
4) *Enhanced generalizability:* Unlike prior approaches, we trained the ViT model using both WV1 and WV2 data, combined with advanced data augmentation techniques. This approach enabled the model to generalize more effectively, which we validated by testing it on a different dataset—AI4Arctic Sea Ice Challenge—with different polarizations and spatial resolutions.
5) *Cross-polarization evaluation:* We are the first to evaluate a model trained on VV-polarized data against dual-polarization (HH/HV) SAR images. This raw evaluation demonstrates the model's adaptability, achieving approximately 80% accuracy on HH polarization and 47% on HV polarization. The reduced performance on HV images underscores the challenges posed by visual differences in polarization but also highlights the model's adaptability in handling complex data.
6) *Benchmarking against CNNs:* We used CNNs as a benchmark, comparing their performance to ViTs, particularly in texture-based versus structure-oriented classifications. The analysis revealed that CNNs performed adequately on texture-based phenomena but struggled significantly with structure-related categories, underscoring ViT's advantages in capturing structural complexity.

These contributions underscore the potential of ViTs for SAR image analysis, where they not only improve classification performance but also offer a more interpretable and generalizable framework for handling diverse SAR datasets. This work aims to advance the field of remote sensing by demonstrating how ViTs can bridge the gap between texture and structure-based image classification, facilitating better understanding and categorization of complex geophysical phenomena.

The rest of this article is organized as follows. *Section II (Materials and methods)* describes the datasets, including TenGeoP-SARwv and AI4Arctic Sea Ice Challenge datasets, their characteristics, and data preparation techniques. This section also details the customized ViT model and highlights the role of attention mechanisms in enhancing classification interpretability. *Sections III (Results) and IV (Discussion)* presents the experimental results, starting with overall model performance and moving into detailed analyses for texture-based and texture-structure-based geophysical categories. Furthermore, we discuss the model's generalization capabilities on the AI4Arctic dataset, analyzing performance on HH and HV polarizations, and providing insights into polarization effects. Finally, *Section V (Conclusion)* summarizes the key findings, emphasizing the contributions of ViTs to SAR image analysis and outlining future research directions to advance remote sensing applications.

II. Materials and Methods

A. Data Description and Selection

This study utilizes ocean SAR vignettes from the Sentinel-1 Wave Mode (S-1 WV) and sea ice data from the AI4Arctic Sea Ice Challenge Dataset to train and evaluate a ViT model. The TenGeoP-SARwv dataset serves as the primary training source, while the AI4Arctic dataset assesses generalization to dual-polarization SAR data.

1) TenGeoP-SARwv Dataset: The TenGeoP-SARwv dataset consists of over 37 000 labeled ocean SAR images from the Sentinel-1 A mission's Wave Mode, acquired in VV polarization. It classifies images into ten geophysical phenomena, including Pure Ocean Waves, Wind Streaks, Micro Convective Cells, Rain Cells, Biological Slicks, Sea Ice, Icebergs, Low Wind Areas, Atmospheric Fronts, and Oceanic Fronts (see Table S1 of the Supplementary Material) [37], ensuring a well-structured categorization for machine learning applications.

While this dataset provides a diverse representation of ocean surface conditions, the acquisition of SAR images across all possible sea states and radar configurations remains challenging. Many geophysical features are inherently limited to specific environmental conditions. For instance, low wind areas and natural surface films predominantly occur in calm sea states, making it unrealistic to obtain examples in high-wave conditions. Conversely, wind streaks and oceanic fronts require dynamic atmospheric or oceanic conditions, restricting their occurrence to specific sea states. Furthermore, although SAR backscatter properties vary with radar parameters such as incidence angle and polarization, the Sentinel-1 WV mode acquisitions used in this study are constrained to fixed imaging settings, limiting the range of radar configurations available for analysis.

Despite these constraints, the Sentinel-1 Wave (WV) Mode systematically captures small oceanic SAR vignettes every 100 km over the open ocean, ensuring consistent spatial and temporal coverage. This mission design ensures consistent spatial and temporal coverage, making it particularly suitable for visual inspection and data-driven analysis. The images were acquired at two incidence angles: WV1 (23.8°) and WV2 (36.8°), each covering an area of 20 km $\times$ 20 km with a spatial resolution of 5 m per pixel [37].

To accommodate different analytical needs, the dataset is available in GeoTIFF and PNG formats. GeoTIFF files preserve full radiometric detail, maintaining accurate brightness and contrast information. PNG files, optimized for visual interpretation, enhance contrast for better feature distinction but lose absolute brightness fidelity. Since most machine learning models are trained on standard image formats like PNG, this study adopts

contrast-enhanced PNG images to facilitate compatibility with DL frameworks. In addition, the TenGeoP-SARwv dataset has undergone speckle noise reduction, and its resolution has been downsampled to 50 m × 50 m per pixel [37], improving usability for machine learning applications.

An essential consideration in this dataset is the brightness variation between WV1 and WV2 images, caused by the different incidence angles at which the images were captured. WV1 data, taken at a lower incidence angle (23.8°), typically exhibits different brightness levels compared to WV2 data, captured at a higher incidence angle (36.8°). However, as detailed in Wang et al. (Section 3.1.1) [37], the TenGeoP dataset has been processed to account for these variations, converting the data into sea surface roughness values, which minimizes the effects of incidence angle across each image. This correction significantly reduces brightness discrepancies between the two angles. When using the PNG files, the mean brightness and contrast have been individually optimized for each image, making the differences between the two incidence angles less relevant. Therefore, the intensity differences across incidence angles play a minor role in our analysis, especially when using PNG files where contrast has been enhanced for visual clarity. This ensures that WV1 and WV2 data are normalized, allowing for more consistent image texture analysis.

While CNN-based studies previously had to train WV1 and WV2 separately due to their sensitivity to localized brightness and texture variations [13], [38], ViTs overcome this limitation by leveraging global self-attention mechanisms. Unlike CNNs, which primarily rely on localized feature extraction, ViTs can dynamically weigh the importance of different regions across an entire image, making them more robust to brightness differences and variations in texture. In addition, data augmentation techniques, including rotations, flips, scaling, and color adjustments, further enhance ViT model generalization by exposing the model to a broader range of conditions [39], [40], [41].

By integrating WV1 and WV2 data into a single training pipeline, this study ensures that ViT models learn a more comprehensive representation of geophysical phenomena, eliminating the need for separate CNN-based training pipelines. This unified approach improves model generalization across viewing angles and lighting conditions, ultimately enhancing classification accuracy for SAR imagery.

2) AI4Arctic Sea Ice Challenge Dataset: The AI4Arctic Sea Ice Challenge dataset was created for the AI4EO sea ice competition, organized by the European Space Agency, to develop DL models capable of generating detailed sea ice charts, including sea ice concentration, stage-of-development, and floe size information [42]. This dataset poses a distinct challenge compared to the TenGeoP-SARwv dataset due to the differences in SAR data polarizations.

The dataset includes Sentinel-1 SAR data in dual polarizations (HH and HV) and complementary passive microwave radiometer (MWR) data from the AMSR2 satellite sensor. While MWR data have been successfully used for sea ice concentration estimation and ice mapping, they have much lower spatial resolution, often resulting in gradual transitions between ice and water rather than sharp boundaries. In contrast, SAR imagery provides significantly higher spatial resolution—by orders of magnitude—allowing for the identification of fine-scale sea ice features. Given these fundamental differences in resolution and imaging principles, findings from SAR-based analyses cannot be directly transferred to MWR imagery. For this study, we focused exclusively on the SAR data and did not incorporate the MWR data into our model evaluations.

In addition, the AI4Arctic dataset provides label data produced by the Greenland Ice Service at the Danish Meteorological Institute and the Canadian Ice Service, ensuring that sea ice charts are generated. The dataset consists of 512 training scenes and 20 test scenes collected between 8 January 2018, and 21 December 2021. Each scene includes auxiliary information, such as the distance to land and numerical weather prediction model data, which are helpful for further classification accuracy, though these were not utilized in this study.

A major distinction between AI4Arctic and TenGeoP-SARwv lies in spatial resolution and contrast optimization. Unlike TenGeoP, which applies spatial smoothing and brightness normalization, the AI4Arctic dataset lacks such preprocessing, potentially increasing classification challenges. In addition, TenGeoP normalizes brightness variations across different incidence angles, whereas AI4Arctic does not apply similar corrections, further complicating direct model adaptation.

To evaluate ViT model generalization, we tested a model trained on VV-polarized TenGeoP-SARwv against the HH and HV-polarized AI4Arctic dataset. These differences in polarization, spatial resolution, and preprocessing assess the ViT model's ability to adapt across different SAR datasets. Each polarization (HH and HV) was evaluated separately to determine the model's effectiveness for individual channels, avoiding any unintended feature mixing.

3) Data Preparation: To train the ViT model, we used a balanced subset from the TenGeoP-SARwv dataset, which was divided into training (70%) and validation (30%) sets. A "balanced" subset ensures that each geophysical category contains an equal number of data points. The split ensured that each geophysical category was well-represented, mitigating potential category imbalance issues from the original dataset (see Table S1 of the Supplementary Material). In addition, a separate test set (5000 images, 500 per class) was extracted from previously unused data to provide an independent performance evaluation.

For evaluating the model's adaptability to different data sources, we employed the AI4Arctic Sea Ice Challenge dataset. This dataset was initially in NetCDF format, which required preprocessing to convert it into PNG files compatible with the input requirements of DL models. To maintain the integrity of HH and HV polarization channels, each polarization was evaluated separately, preventing feature mixing. This approach enabled a direct comparison of the ViT model's performance across different polarization datasets, helping assess its generalization capabilities.

B. Customized ViT

The ViT introduces a transformative approach to image classification by leveraging a transformer architecture initially developed for natural language processing tasks [36]. Unlike traditional CNNs, which utilize localized filters to extract

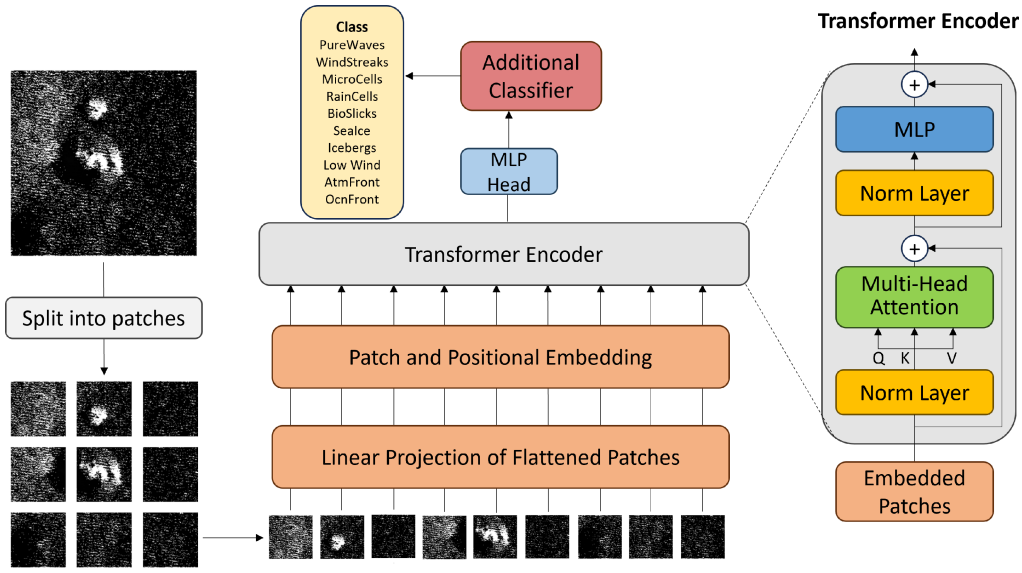

Fig. 1. *Customized ViT architecture for SAR image classification.* The model begins by dividing each input image into nonoverlapping patches, which are then flattened and linearly projected into embeddings. These patches undergo positional embedding to retain spatial information and are passed into the Transformer Encoder. The Transformer Encoder uses multihead self-attention layers to capture long-range dependencies between image patches, followed by normalization and multilayer perceptron (MLP) layers for further processing. The encoded information is then fed through the MLP head, followed by an additional classifier layer that assigns the image to one of ten geophysical categories.

features, ViTs divide images into patches and treat each patch as an independent token, similar to how words are handled in text processing. This allows ViTs to capture long-range dependencies and integrate information across the entire image, overcoming the limitations of CNNs, which can struggle with global context due to their focus on localized areas.

This study employed a customized ViT model explicitly tailored for remote sensing image classification, focusing on classifying geophysical phenomena from SAR images (see Fig. 1). The ViT architecture used here is based on the standard Transformer encoder, adapted to process sequences of image patches. The model divides each input image into nonoverlapping patches and represents each patch as a vector. These vectors are embedded and augmented with positional embeddings, preserving spatial information. The resulting sequence of embedded patches is fed into the Transformer encoder, which processes the sequence through multiple layers of multihead self-attention and feedforward networks. This architecture allows the model to develop a rich, context-aware representation of the image by capturing both local and global relationships.

Our customized ViT architecture consists of four key components as follows.

1) *Embedding layer:* Converts patches into vector representations with positional embeddings.
2) *Transformer encoder layers:* Applies self-attention mechanisms to learn feature dependencies across the entire image.
3) *Classification head:* Outputs the final classification decision.
4) *Additional classifier:* A set of fully connected layers refining predictions for enhanced classification accuracy (see Fig. 1).

To optimize the ViT model for SAR classification, we applied fine-tuning strategies on pre-trained ViT models, which included the following.

1) Replacing the classification head with a new task-specific layer.
2) Freezing early layers and gradually unfreezing deeper layers during training.
3) Using a lower learning rate for pre-trained layers to preserve learned representations.
4) Employing data augmentation (e.g., rotations, flips, scaling, and color adjustments) to improve generalization (see Supplementary Material for details).

We tested two pretrained ViT configurations [43], [44] as follows.

1) ViT-Base-16 and ViT-Large-32, both evaluated using input image sizes of 224×224 pixels and 384×384 pixels.

Performance was assessed using accuracy, precision, and F1 score to determine the best balance between computational efficiency and classification accuracy, especially for high-resolution SAR data.

A key improvement in our customized ViT was the additional classifier that refines predictions from the transformer encoder. By better differentiating between geophysical classes, this enhancement improved classification accuracy, ensuring reliable and consistent performance on complex SAR images.

1) Attention Mechanism in ViTs: This study uses the Attention Rollout technique to explore and interpret the self-attention mechanism within our customized ViT models [45]. The self-attention mechanism is a crucial feature of Transformer architectures, allowing them to focus on different parts of an image selectively. This enables the model to prioritize the most relevant regions when making classifications, which is especially

important in tasks like geophysical phenomena detection in remote sensing data.

The Attention Rollout method provides a systematic approach to tracking the flow of information from the input layer to the higher layers of the transformer. Specifically, it computes how attention at each position in a higher layer l_i relates to attention at preceding layers l_j (where $j < i$). This allows us to visualize and quantify how the model aggregates information across layers, providing insight into which regions of an image influence the model's final output.

In the attention graph of a Transformer, a path from node v at position k in layer l_i to node u at position m in layer l_j consists of a series of edges. Each edge has a weight that represents the proportion of information passed between nodes. To compute how much information at node v is propagated to node u, we multiply the attention weights of all edges along the path. Since there are often multiple paths between any two nodes, we sum the contributions from all possible paths to determine the total flow of information from v o u.

At the implementation level, Attention Rollout is computed by recursively multiplying the attention weight matrices from successive layers. The following formula represents the process [45]:

$$\tilde{A}(l_i) = \begin{cases} A(l_i)\tilde{A}(l_{i-1}), & \text{if } i > j \\ A(l_i), & \text{if } i = j. \end{cases} \quad (1)$$

Here, $\tilde{A}$ is the attention rollout matrix, A is the raw attention weight matrix, and the multiplication operation is used to propagate attention values across layers. To compute the input attention, we set $j = 0$, starting the rollout process from the bottom of the model and progressing upward.

By using this technique, we can visualize how different regions of the input image contribute to the final prediction, offering a transparent and interpretable view of the attention process in the ViT.

III. Results

The results from the dataset creator and a comprehensive paper [46] on CNNs provide a baseline for evaluating our work. One of the limitations of the comprehensive study was the use of all available data, leading to an imbalance in different categories. This imbalance can negatively impact the performance and generalizability of the models.

To address this issue, we created a balanced subset from the original dataset, ensuring each category had the same amount of training, validation, and testing data. This approach, although utilizing less overall data, allowed for a more balanced and fair training and testing process. Our method involved using a different DL architecture, the ViT, which is known for its superior performance compared to traditional CNNs.

To explore the potential of our ViT models further, we tested them on images from other sources, such as the AI4Arctic Sea Ice dataset. These additional tests aimed to assess the model's versatility and performance in recognizing phenomena outside the training image frame. We also discussed whether each category is texture-oriented or structure-oriented. ViT demonstrated significantly better performance than CNNs in structure-oriented categories.

A. Overall Results

We evaluated the performance of different ViT models, specifically ViT-b16-224, ViT-b16-384, ViT-l32-224, and ViT-l32-384, across ten geophysical categories. The metrics used for evaluation were Recall, Precision, and F1 score. Detailed performance visualizations, including confusion matrices (see Figs. S1 and S2 of the Supplementary Material), training versus validation loss curves (see Fig. S3 of the Supplementary Material), and precision–recall curves (see Fig. S4 of the Supplementary Material), are provided in the Supplementary Material.

As shown in Table I, the ViT-b16-224 model showed a high level of accuracy, with Recall, Precision, and F1 scores averaging around 0.951. The ViT-b16-384 model demonstrated even better performance, with an average of 0.967 for all three metrics. Similarly, the ViT-l32-224 model maintained a strong performance with an average of 0.955, while the ViT-l32-384 model outperformed the rest with an average of 0.973.

These results indicate that the higher input-size models (384) generally outperformed their lower input-size counterparts (224), and notably, the smaller pretrained model (b16) with higher input size surpassed the larger pretrained model (l32) with lower input size. This underscores the critical role that input size plays in enhancing model accuracy. The consistent performance across various models further highlights the robustness of the ViT architecture in effectively handling SAR imagery.

We also compared the ViT models (ViT-l32) with traditional CNNs (Inception and AlexNet) using different sample sizes per category, as presented in Table II. The CNN-Inception model's performance is reported by the TenGeoP-SARwv dataset creators as a benchmark [38], while the results for AlexNet are drawn from a recently published study [13]. The comparison was made using Recall, Precision, and F1 scores.

According to Wang et al. [38], CNNs do not show significant performance improvements beyond 320 training samples per category. In contrast, our results indicate that ViTs can continue to benefit from additional training data. This aligns with findings in DL literature [47], where transformer-based models, due to their self-attention mechanisms and lack of inductive biases present in CNNs, generally require larger datasets for optimal performance. Specifically, our analysis shows that even at 84 samples per category, ViTs demonstrate noticeable improvements in classification accuracy compared to CNNs (see Table II).

For the category of Pure Waves, the CNN-Inception model achieved a recall of 0.43, precision of 0.99, and F1 score of 0.60 when trained with 320 samples per category. In contrast, the ViT-l32 model achieved a recall of 0.96, precision of 0.99, and F1 score of 0.98, demonstrating significant improvement even with fewer training samples.

Across all categories, the average performance of the ViT-l32 model with 699 samples per category was superior, with recall, precision, and F1 scores of 0.975, 0.976, and 0.970, respectively. The CNN model, while competent, lagged behind with average scores of 0.876, 0.681, and 0.679, respectively.

TABLE I
COMPARISON OF VIT MODELS ACROSS GEOPHYSICAL CATEGORIES

Category	ViT-b16-224			ViT-b16-384			ViT-l32-224			ViT-l32-384		
	Recall	Precision	F1 score	Recall	Precision	F1 score	Recall	Precision	F1 score	Recall	Precision	F1 score
PureWaves	0.97	0.98	0.98	1	0.98	0.99	0.97	0.97	0.97	0.97	0.99	0.98
WindStreaks	0.97	1	0.98	0.97	1	0.99	0.99	1	0.99	1	1	1
MicroCells	1	0.89	0.94	0.97	0.94	0.96	0.96	0.98	0.99	0.99	1	1
RainCells	0.98	0.99	0.99	1	0.99	0.99	0.97	0.98	0.97	0.99	0.99	0.99
BioSlicks	0.94	0.92	0.93	0.99	0.94	0.96	0.96	0.91	0.93	0.98	1	0.99
SeaIce	0.86	1	0.93	0.93	1	0.97	0.9	1	0.94	1	0.99	1
Icebergs	0.9	0.97	0.93	0.93	1	0.96	0.94	0.96	0.95	0.98	0.97	0.98
LowWind	1	0.93	0.96	1	0.97	0.98	1	0.9	0.95	1	0.98	0.99
AtmFront	0.94	0.98	0.96	0.97	0.94	0.95	0.94	0.97	0.95	0.99	0.95	0.97
OcnFront	0.95	0.87	0.91	0.91	0.94	0.92	0.93	0.93	0.93	0.95	0.97	0.96
Average	0.951	0.953	0.951	0.967	0.97	0.967	0.956	0.96	0.955	0.985	0.984	0.986

TABLE II
PERFORMANCE COMPARISON BETWEEN CNN AND VIT MODELS WITH DIFFERENT SAMPLE SIZES

Category	Recall				Precision				F Score			
	Inception (320)	AlexNet (320)	ViT (699)	ViT (84)	Inception (320)	AlexNet (320)	ViT (699)	ViT (84)	Inception (320)	AlexNet (320)	ViT (699)	ViT (84)
Texture-based Categories												
PureWaves	0.43	0.515	**0.97**	0.96	0.99	0.88	0.99	**1**	0.6	0.775	**0.98**	**0.98**
SeaIce	0.93	0.8	**1**	0.87	0.96	0.835	**0.99**	0.98	0.945	0.81	**1**	0.93
LowWind	**1**	0.75	**1**	**1**	0.83	0.88	**0.98**	0.96	0.905	0.81	**0.99**	0.97
Texture-Structurebased Categories												
WindStreaks	0.83	0.66	**1**	0.97	0.865	0.65	**1**	0.94	0.845	0.64	**1**	0.95
MicroCells	0.825	0.395	**0.99**	0.93	0.85	0.85	**1**	0.97	0.835	0.54	**1**	0.95
RainCells	0.93	0.695	0.99	**1**	0.84	0.88	**0.99**	0.95	0.88	0.775	**0.99**	0.97
BioSlicks	0.92	0.8	**0.98**	0.95	0.895	0.685	**1**	0.94	0.905	0.725	**0.99**	0.94
Icebergs	0.945	0.26	**0.98**	0.94	0.17	0.74	**0.97**	0.93	0.285	0.32	**0.98**	0.94
AtmFront	0.945	0.295	**0.99**	**0.99**	0.385	0.765	**0.95**	0.75	0.55	0.39	**0.97**	0.85
OcnFront	**1**	0.54	0.95	0.72	0.02	0.65	0.97	**0.99**	0.04	0.64	**0.96**	0.83
Average	0.876	0.571	**0.985**	0.933	0.681	0.7815	**0.984**	0.941	0.679	0.6425	**0.986**	0.931

The number of input samples per category is indicated below each model name.
Bold values indicate the best performance for each category across all models.

These comparisons underline the ViT model's enhanced capability to recognize and classify different geophysical phenomena in SAR imagery accurately. The superior performance of ViT, especially in structure-oriented categories, suggests that the self-attention mechanism in the Transformer architecture provides a more nuanced understanding of complex patterns and textures than CNNs.

B. Performance on Texture-Based Categories

This section discusses the performance of our customized ViT and CNNs on texture-based categories, including pure ocean waves, sea ice, and low wind areas. These categories primarily rely on the textural features of SAR images, which are less complex in terms of structural elements. Both ViT and CNN models demonstrate strong performance in these texture-oriented categories, indicating their ability to accurately capture and interpret the repetitive patterns and homogeneous textures present in these images.

1) Successful classifications: In all three texture-based categories, the models showed high accuracy in predictions, as demonstrated by the examples provided.

For the pure ocean waves category, the defining characteristics include periodic wave signatures, wavelengths ranging from 0.1 to 0.8 km, and homogeneous intensity modulation across the image [37]. The attention map (see Fig. 2, top row, right panel) reveals that the ViT model focuses on the uniform, repetitive patterns in the SAR image, which are indicative of ocean waves. The model's ability to capture these periodic signatures across the image is reflected in the high-confidence prediction of 84% for pure ocean waves. The attention is evenly distributed across the image boundaries, aligning with the expectation that no other competing geophysical features are present in this category. This demonstrates the model's capacity to discern the uniform textural patterns characteristic of ocean waves.

Sea ice is characterized by complex textural contexts, including web-shaped, wiggly fractures, pebble-like patterns, and strong intensity contrasts between patches [37]. In the successful prediction for sea ice (see Fig. 2, middle row, right panel), the attention map highlights the intricate, fragmented structures that are typical of sea ice formations. The ViT model successfully identifies the key textural features, highlighting the boundaries and varied textures within the sea ice patches, which are essential for distinguishing this category from others. The focus on these critical areas leads to a confident prediction of 86% for sea ice, showcasing the model's effectiveness in processing complex textures and contrasts in SAR imagery.

Low wind areas are marked by a dominant dark patch, heterogeneous intensity modulation, and the absence of periodic wave signatures [37]. In the case of Low Wind Areas (see Fig. 2, bottom row, right panel), the attention map highlights the dark, smooth regions that signify calm sea conditions with minimal wind activity. The ViT model correctly identifies these dark patches as the most significant features, leading to a high-confidence prediction of 81%. The attention is concentrated on the homogeneous areas devoid of wave patterns, which are

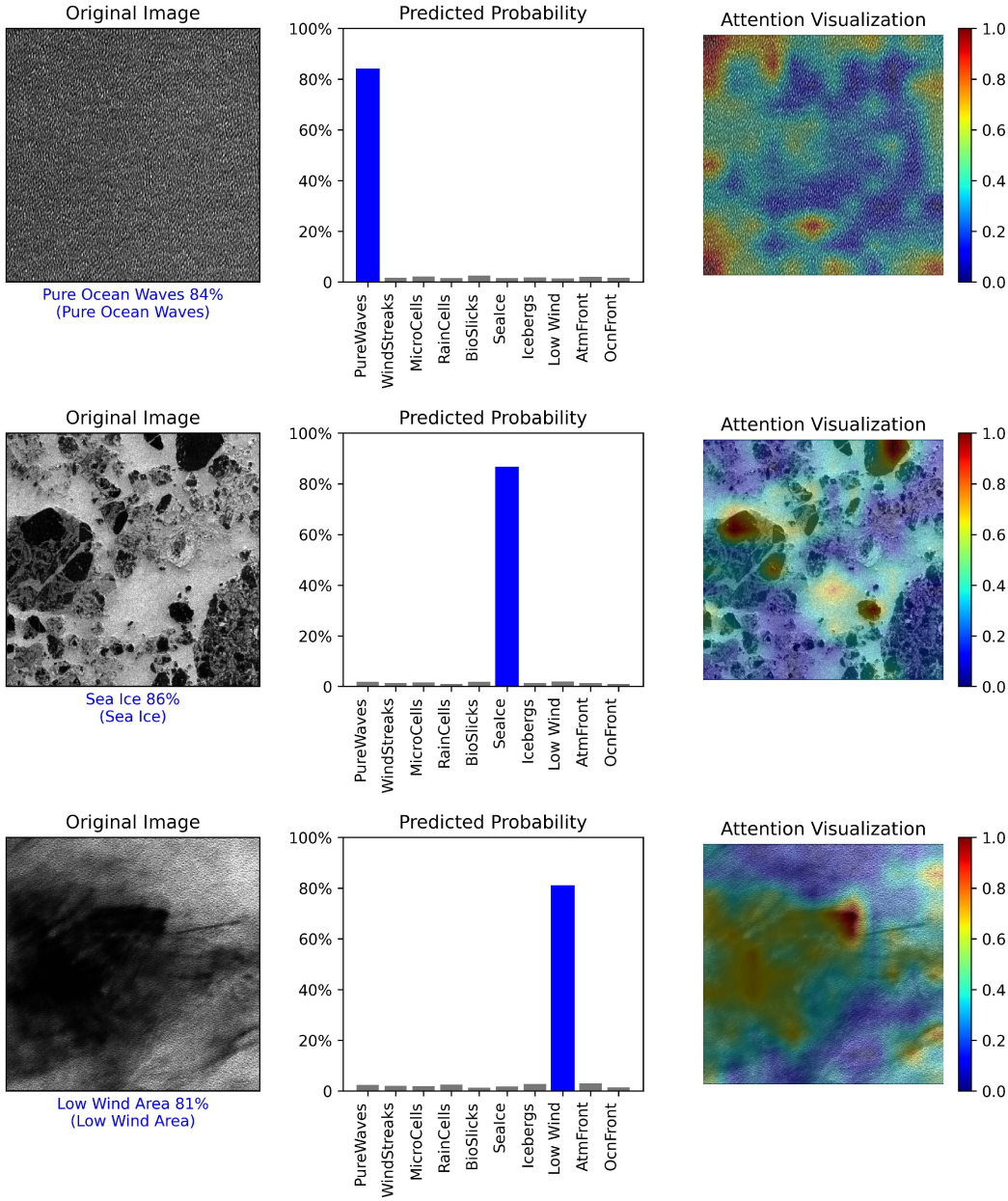

Fig. 2. *Examples of successful classifications of texture-based categories using the ViT model. Left column:* The original SAR image depicting specific ocean phenomena. The first line below each image indicates the predicted category along with the model's confidence probability, while the second line in parentheses indicates the true category. *Middle column:* The predicted probability distribution for various geophysical categories, showing the overwhelming confidence for the correct category. *Right column:* The attention map generated by the ViT model highlights the regions that contributed most to the model's prediction. These figures focus on texture-based geophysical categories, where the ViT model performed well, capturing the relevant texture features.

crucial for accurately classifying low wind areas. This indicates that the model is well-tuned to recognize the subtle textural cues associated with low wind conditions.

2) Inaccurate Classifications: In contrast to the successful predictions, Fig. 3 highlights instances where the ViT model inaccurately predicted the geophysical phenomena depicted in SAR images. The examples illustrate misclassifications in the texture-based categories, specifically for pure ocean waves and sea ice.

The upper section of Fig. 3 shows a SAR image that the model incorrectly classified as "Icebergs" with a confidence probability of 39%. The true category of the image is "pure ocean waves," which was misidentified by the model. The attention visualization on the right highlights the regions that contributed most to the ViT model's prediction. Instead of focusing on the uniform, periodic texture of ocean waves, the model erroneously emphasized areas with higher contrast and isolated patterns. According to the criteria for pure ocean waves, the image should have been identified by its homogeneous intensity modulation and periodic wave signatures. However, the attention map shows a dispersed focus, suggesting that the model was confused by subtle textural variations that it incorrectly associated with

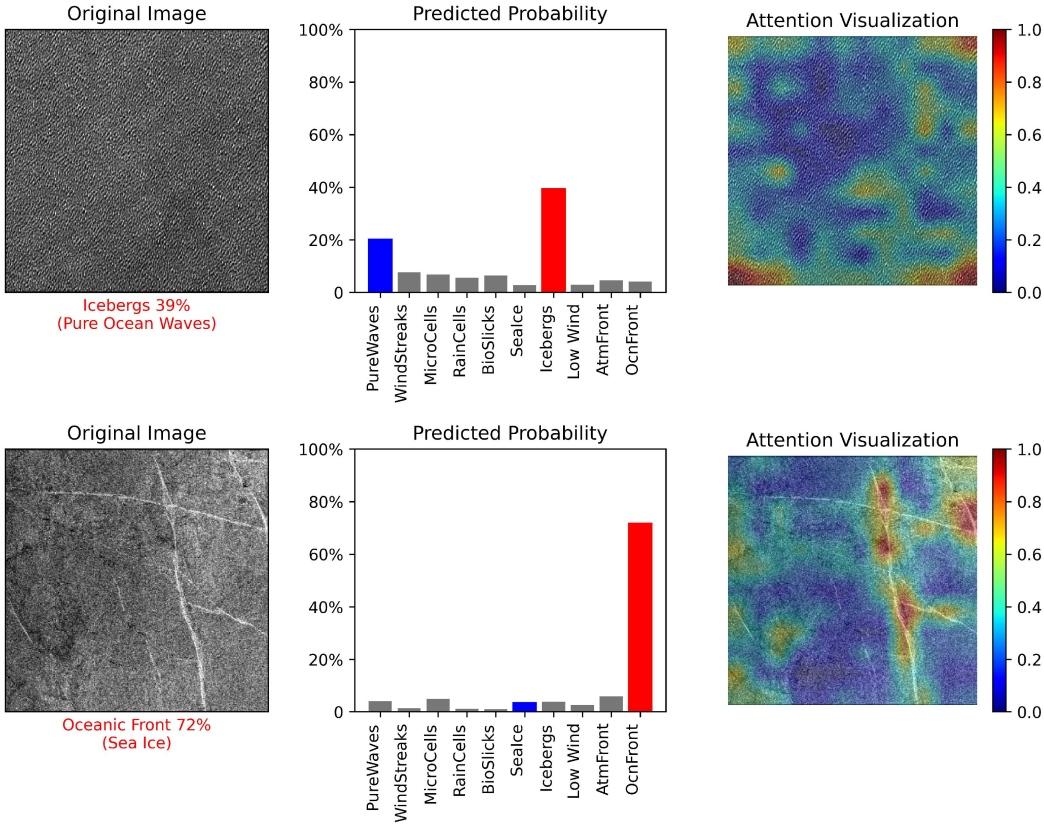

Fig. 3. *Examples of inaccurate classifications for texture-based categories. Left column:* The original SAR image depicting specific ocean phenomena. The first line below the image indicates the predicted category along with the model's confidence probability, and the second line in parentheses indicates the true category. *Middle column:* The predicted probability distribution for different geophysical categories, illustrating the model's uncertainty or misclassification. *Right column:* The attention map generated by the ViT model, highlighting the regions the model focused on that led to the incorrect classification.

Icebergs, which typically feature bright or dark targets with related shadows.

The lower section of Fig. 3 illustrates a misclassification where an image of "sea ice" was predicted as "oceanic front" with a confidence probability of 72%. While sea ice is typically characterized by complex textural patterns, such as web-shaped, wiggly fractures, or pebble-like structures, this particular image depicted a continuous ice cover with bright ridges. These linear ridges visually resemble the thin, crossing features that are often indicative of oceanic fronts, which likely influenced the ViT model's attention and led to the misclassification.

This example underscores a challenge in the dataset's categorization. Unlike many sea ice images that show distinct ice floes with clear boundaries and strong intensity contrasts, this case presents a unique, complete ice cover with intersecting bright ridges. The model's inability to differentiate this from multiple, intersecting oceanic fronts points to a gap in the learning process—it has not learned that several crossing oceanic fronts do not typically occur in this manner. In addition, this type of continuous ice cover is visually distinct from typical sea ice images, suggesting that more granular subcategories, such as "sea ice—distinct ice floes" and "sea ice—complete cover," could help mitigate confusion. The input dataset could benefit from more specific subcategorization to address these variations, ultimately improving classification accuracy and model performance.

There were no instances of inaccurate predictions for the low wind areas category. According to Table I, the recall for low wind areas was consistently 1, regardless of the pretrained model size or input image size. This indicates that the ViT model was exceptionally reliable in recognizing the distinct dark patches and heterogeneous intensity modulation that characterize low wind areas.

C. Performance on Texture-Structure-Based Categories

In texture-structure-based categories, which include wind streaks, micro convective cells, rain cells, biological slicks, icebergs, atmospheric fronts, and oceanic fronts, the ViT model demonstrates markedly superior performance compared to traditional CNNs. This section discusses both the successful and inaccurate predictions made by the ViT model, underscoring the areas where ViT outperforms CNNs and highlighting the challenges that remain.

1) Successful Classifications: The ViT model excels in accurately predicting categories that combine both texture and structure features. Figs. 4 and 5 illustrate examples of successful predictions across the various texture-structure-based categories.

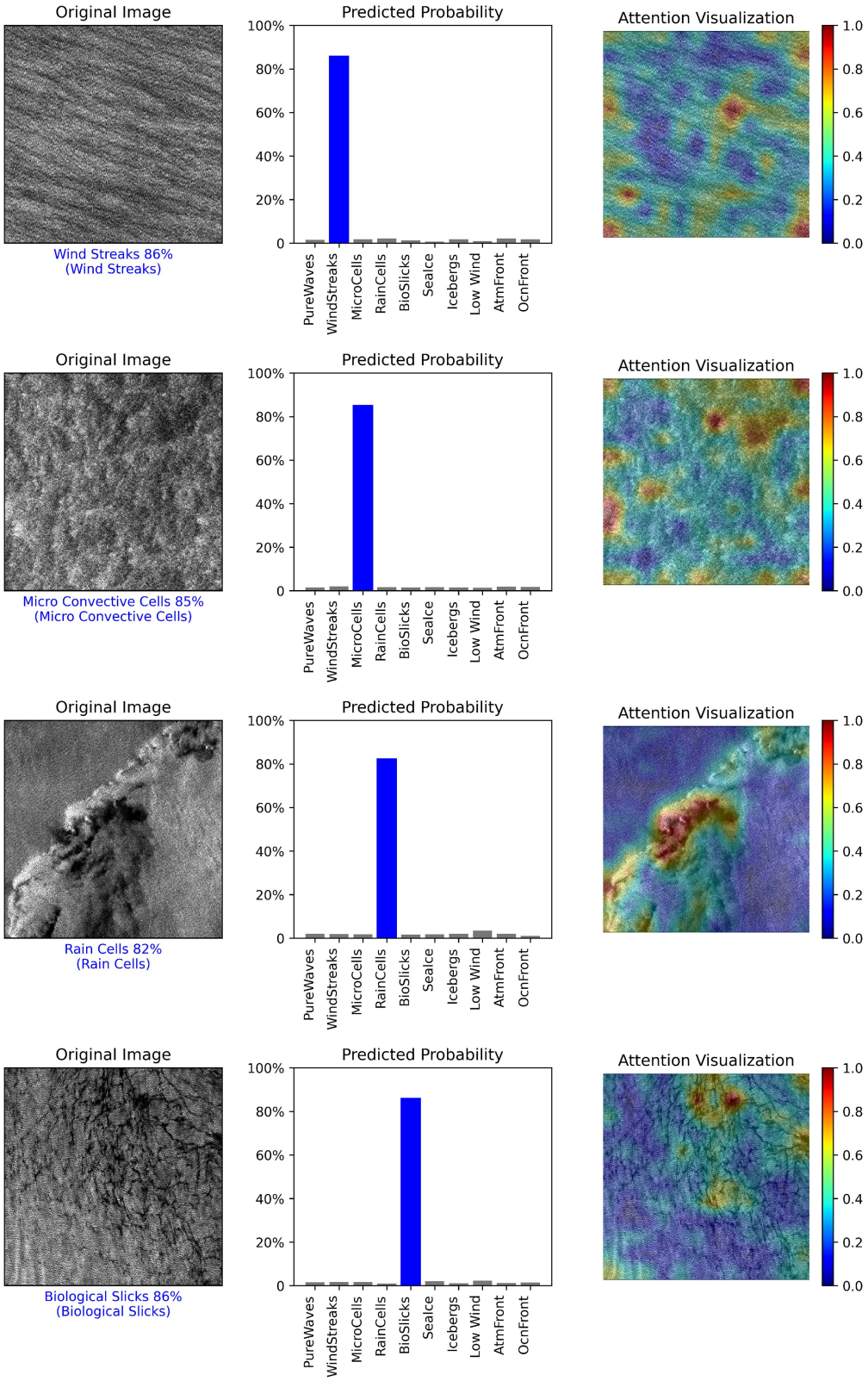

Fig. 4. *Examples of successful classifications for texture-structure-based categories. Left column:* The original SAR image depicting specific ocean phenomena. The first line below the image shows the predicted category with the model's confidence probability, and the second line in parentheses indicates the true category. *Middle column:* The predicted probability distribution for various geophysical categories demonstrates a high confidence in the correct classification. *Right column:* The attention map generated by the ViT model showcases the regions that were most influential in the accurate classification. This set of figures emphasizes the performance of texture-structure-based categories, in contrast to the earlier focus on texture-based categories.

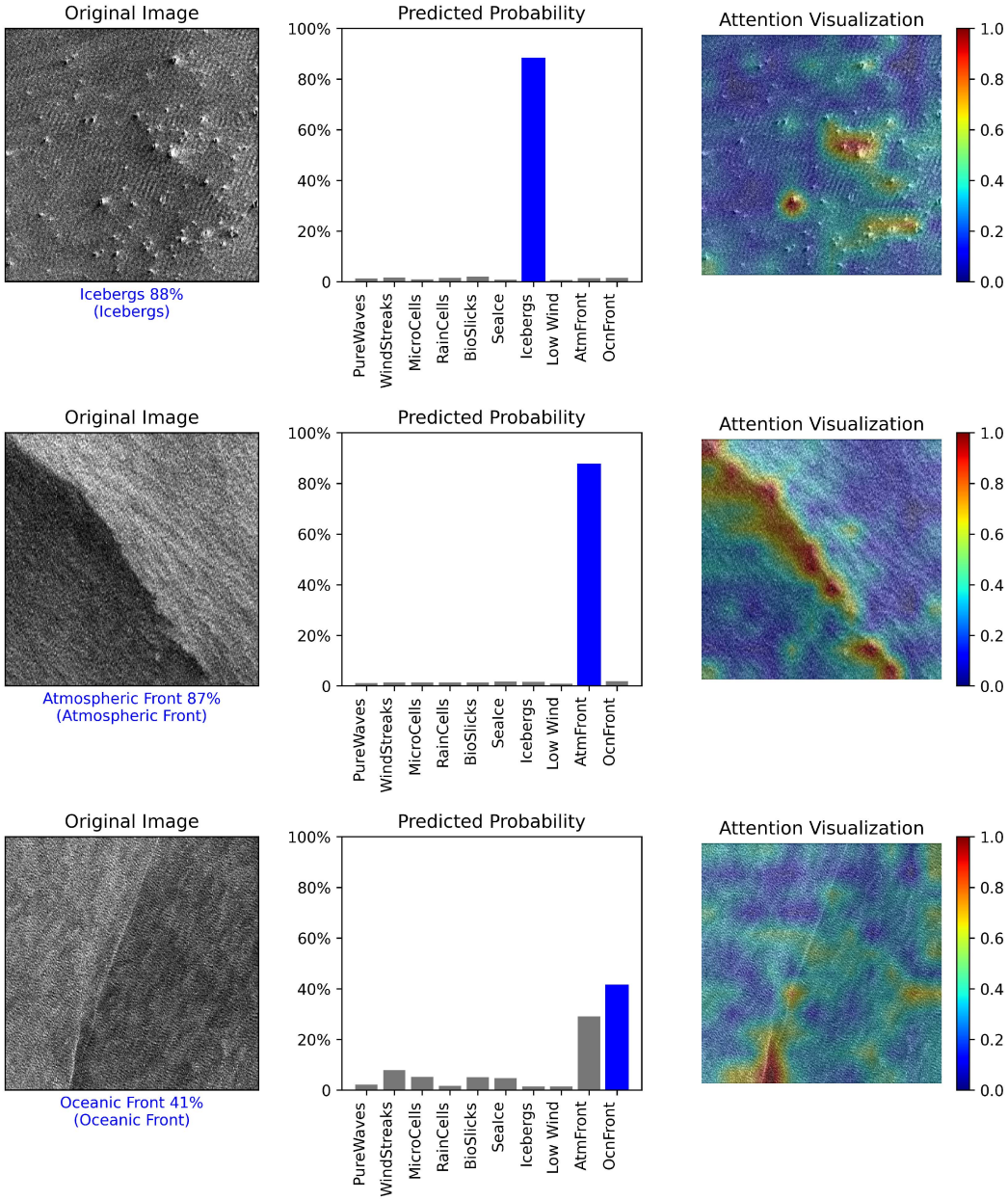

Fig. 5. *Examples of successful classifications for texture-structure-based categories* (continued).

For wind streaks, ViT effectively identifies the periodic linear features and homogeneous intensity modulation that are characteristic of this category. The attention map highlights these linear structures, demonstrating the model's ability to focus on relevant features that distinguish wind streaks from other phenomena.

In the case of micro convective cells, the model accurately captures the coherent, periodic, and cell-shaped features that dominate the image. The attention visualization shows that the model zeroes in on the specific areas of the image that correspond to these features, leading to a high-confidence prediction.

Rain cells are also correctly identified by the ViT model, which effectively recognizes the circular or semi-circular shapes and the heterogeneous intensity modulation within the scene. The attention map underscores the model's focus on these defining features, ensuring an accurate classification.

For biological slicks, although the ViT model successfully identifies the category, the attention map reveals that the model sometimes focuses on areas other than the slicks themselves. Despite this, the overall prediction accuracy remains high, reflecting the model's robustness in handling complex scenes.

Icebergs are another category where ViT excels, particularly in identifying bright or dark targets associated with shadows. The attention map clearly highlights these structures, which are key to distinguishing Icebergs from other categories.

Atmospheric fronts and oceanic fronts are both characterized by linear features and intensity gradients. The ViT model effectively identifies these characteristics, as evidenced by the

attention maps that highlight the relevant front structures and surrounding areas. These successful predictions underscore ViT's ability to handle the complex interplay of texture and structure in these categories, outperforming CNNs in both accuracy and confidence.

2) Inaccurate Classifications: Despite the overall superior performance of the ViT model, there are instances where misclassifications occur. Figs. 6 and 7 illustrate these inaccurate predictions.

For wind streaks, the ViT model occasionally misclassified these images as micro convective cells, particularly when the image displayed convection cells of different length scales that produced a streaky pattern. The attention map in these cases indicated that the model focused on the cell structure of the features rather than the linear patterns typical of wind streaks. This suggests that the model detected underlying convection-like behavior, which aligns with the complex nature of the image. The original dataset classification might not have been definitive, as some human interpreters could classify such features as either wind streaks or convection cells due to their dual characteristics. This result points to a limitation in the dataset labeling, where ambiguous cases were not eliminated or clearly categorized. Importantly, the model's identification of these mixed features is not due to a fundamental misinterpretation but rather an emphasis on the convection cell structure over the convective roll or streak pattern. This demonstrates the ViT model's nuanced capability to recognize complex image features and suggests its potential for refining and enhancing geophysical classifications beyond initial human labeling.

In the case of micro convective cells, the ViT model occasionally misclassified these images as atmospheric fronts, particularly when there was noticeable variation in surface roughness across the image, such as a distinct contrast between the lower left and upper right areas. The attention map revealed that the model concentrated on the linear features present in the image, which might have contributed to the misclassification. This misidentification suggests that while the model recognized significant features associated with surface roughness, it focused more on elements resembling atmospheric fronts rather than the cell-shaped structures characteristic of MCC. However, given the image's mixed characteristics, the model's classification is not entirely incorrect, as it picked up on relevant patterns that could reasonably align with multiple geophysical categories. This outcome emphasizes the complexity of accurately distinguishing between such categories and underscores the need for more nuanced labeling practices in the dataset.

Rain Cells are occasionally misclassified as Icebergs when the circular or semi-circular features in the image resemble cell-like structures rather than the characteristic bright or dark patches of rain cells. In these instances, the ViT model's attention map focuses on regions that lead to misclassification, prioritizing structural features over the distinctive texture of rain cells. This challenge underscores the difficulty of interpreting SAR images without supplementary context, as human operators often rely on additional information—such as latitude, longitude, and temperature data—to determine whether a feature is more likely an Iceberg or a rain cell. For example, according to the dataset creator, Icebergs are primarily found in the Southern ocean near Antarctica [37], meaning spatial and environmental context is crucial for accurate classification. This limitation highlights the importance of external geographical and climatic cues, which are routinely considered by human experts but are absent in automated models.

For biological slicks, the ViT model occasionally misclassifies the category as Icebergs. This misclassification likely occurs because the slicks in the image are too thin. When the image is divided into patches, the dark areas characteristic of Biological Slicks do not dominate each patch. Only in the lower region of the image, highlighted by the attention map, do the dark areas account for the majority, leading the model to classify it as a dark, cell-shaped feature typical of Icebergs. While this results in a misclassification, it is notable that the second-highest prediction was the correct category, Biological Slicks. This indicates that the model is not entirely off-track, but instead, its focus on patch-level details sometimes leads it astray.

Icebergs are occasionally misclassified as rain cells when the ViT model places undue emphasis on circular structures in the image, rather than focusing on the bright or dark targets with associated shadows that are characteristic of Icebergs. In these instances, the attention map reveals that the model's focus is directed toward circular regions, which more closely resemble the cell-shaped features typical of rain cells. This misprioritization leads to inaccurate classification, highlighting a limitation of the model when interpreting structural cues in isolation. Similar to the rain cell-iceberg confusion, this demonstrates the need for additional geographical or environmental context—such as latitude, longitude, or temperature data—to guide more accurate predictions in cases where visual features overlap between categories.

Atmospheric fronts are occasionally misclassified as low wind areas when certain regions of the image exhibit low wind characteristics, even though they do not dominate the entire image. In these cases, the attention map reveals that the ViT model correctly identified parts of the image as having low wind activity, but this focus led to misclassification due to the model not fully capturing the broader features of the atmospheric front. Notably, the second-highest prediction probability was for Atmospheric Front, indicating that the model recognized elements of the correct category but ultimately prioritized the wind features. While the overall classification may seem incorrect based on probabilities, the model's recognition of low wind patterns within the image demonstrates a nuanced understanding, making this result less problematic than it initially appears.

Finally, Oceanic Fronts can sometimes be misclassified as atmospheric fronts, particularly when the linear features in the image resemble those typical of atmospheric fronts. Determining whether such a front is oceanic or atmospheric can be challenging, especially when only a single polarization is available, as it limits the depth of information needed for clear differentiation. The attention map indicates that the model accurately highlighted the linear structures but ultimately assigned them to the wrong category. Importantly, the second-highest prediction probability was for oceanic front, suggesting that the model was close to the correct classification but leaned toward interpreting

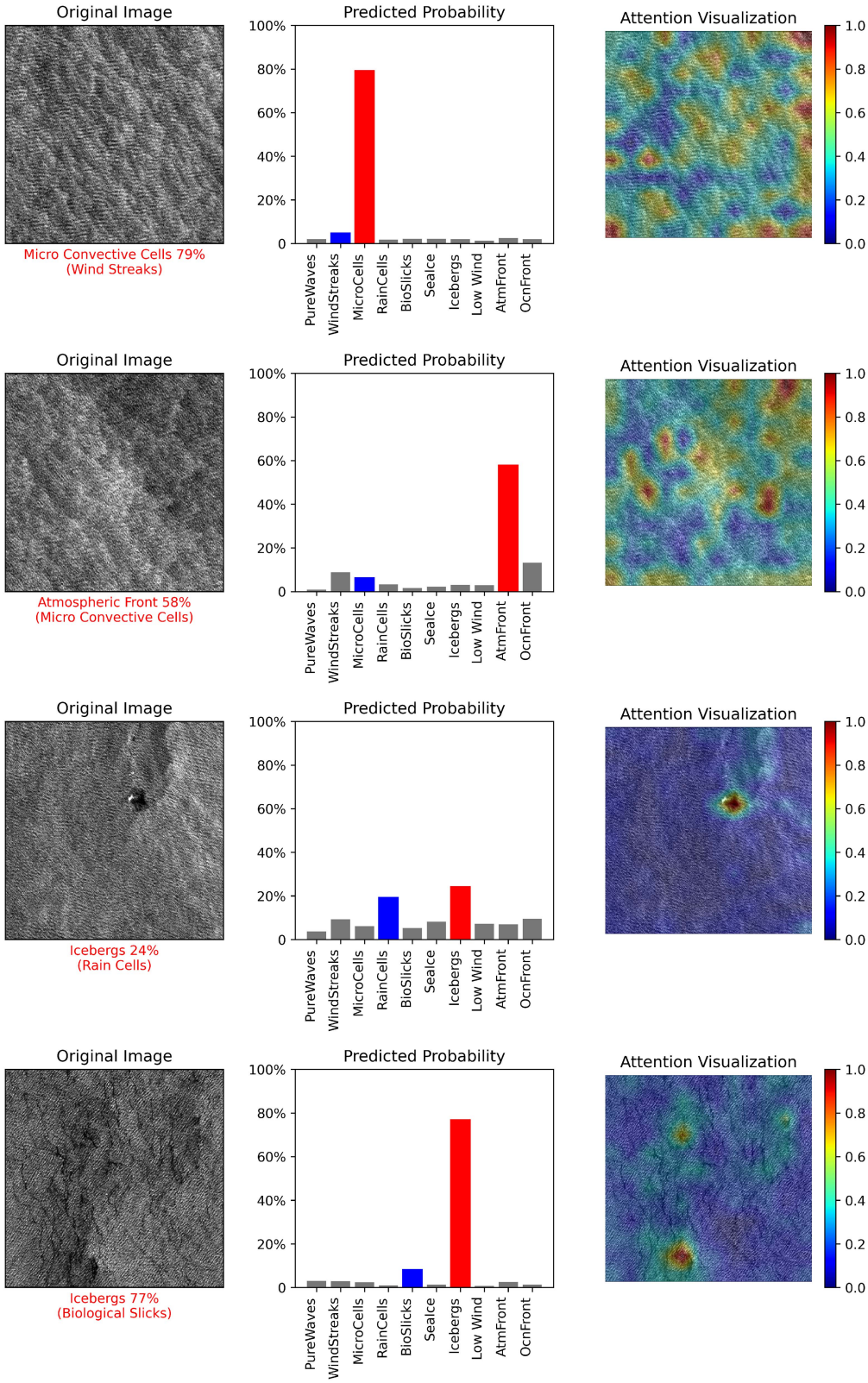

Fig. 6. *Examples of inaccurate classifications for texture-structure-based categories. Left column:* The original SAR image depicting specific ocean phenomena. The first line below the image indicates the predicted category along with the model's confidence probability, and the second line in parentheses indicates the true category. *Middle Column:* The predicted probability distribution for different geophysical categories, highlighting the model's misclassification. *Right column:* The attention map generated by the ViT model reveals the regions of the image that the model misinterpreted. These figures specifically target texture-structure-based categories and how the model struggled with distinguishing between structural and textural features.

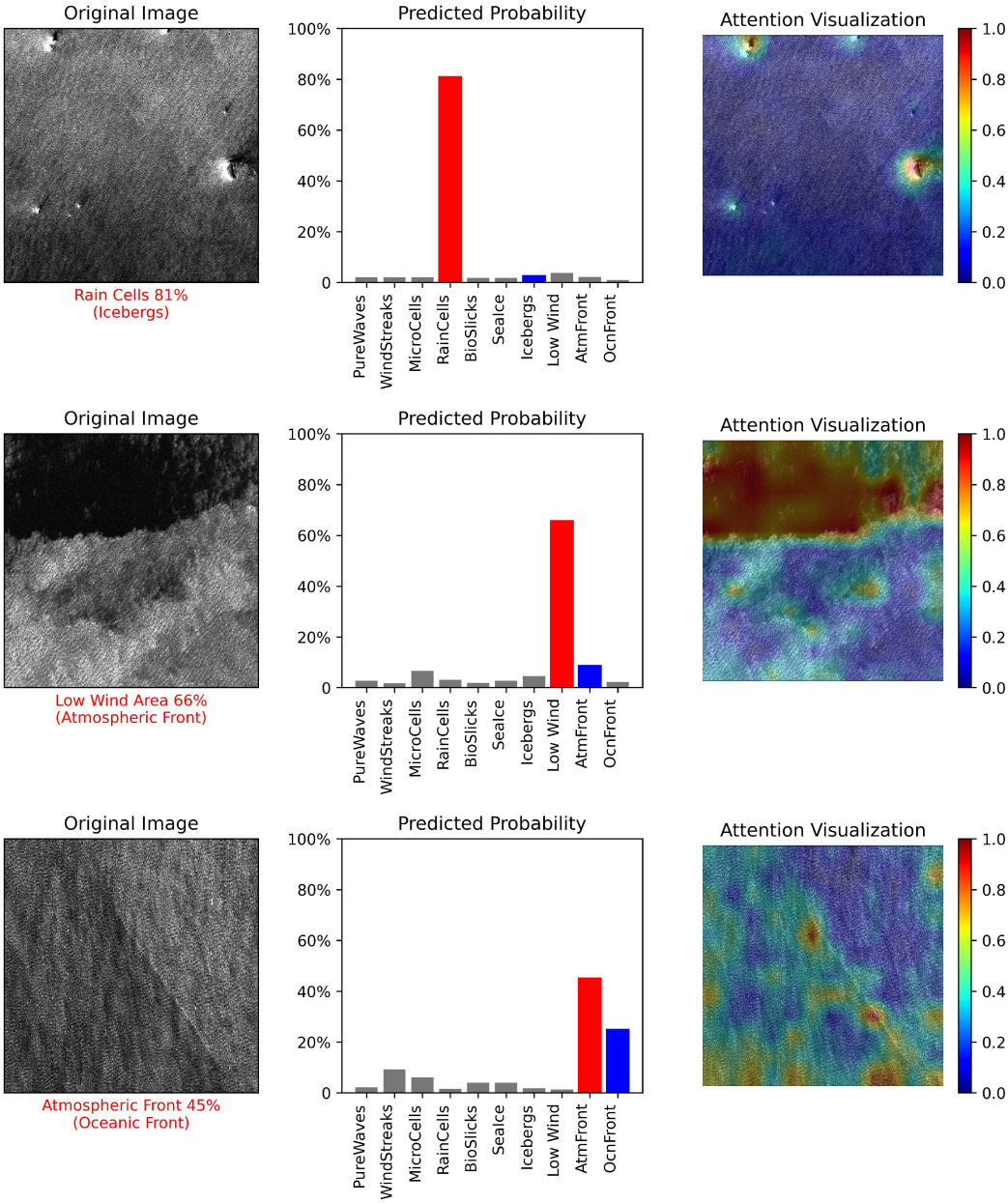

Fig. 7. *Examples of inaccurate classifications for texture-structure-based categories* (continued).

the linear patterns as atmospheric rather than oceanic. This result underscores the inherent difficulty of distinguishing between these two types of fronts using limited polarization data and highlights the nuanced nature of SAR image interpretation in such cases.

D. Performance on AI4Arctic Sea Ice Dataset

In this section, we evaluate the ViT model's ability to generalize to unseen data from dual-polarization sensors using the AI4Arctic sea ice challenge dataset. Unlike the TenGeoP-SARwv dataset, which contains VV polarization data used for training, the AI4Arctic dataset features SAR images with HH and HV polarizations. This distinction provides a unique opportunity to test the model's adaptability to different polarizations and sensor configurations. Importantly, all data in the AI4Arctic dataset are classified as "sea ice."

1) Overall Evaluation on HH Polarizations: For HH polarization data (see Fig. 8), the ViT model achieved an accuracy of approximately 80%, correctly classifying 405 out of 512 samples as "sea ice." This performance demonstrates the model's robustness in identifying sea ice patterns similar to those in the training data. However, 66 samples were misclassified as "low wind," 18 as "atmospheric front," 14 as "Rain Cells," 3 as "Biological Slicks," and 6 as "Icebergs." These errors suggest occasional misinterpretations of "sea ice" features, which could be attributed to shared structural or textural elements across categories.

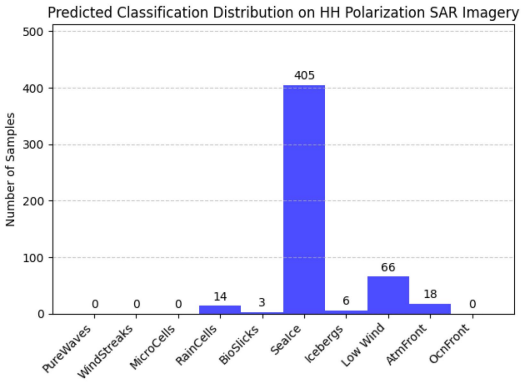

Fig. 8. Predicted classification distribution on HH polarization SAR imagery from the AI4Arctic sea ice challenge dataset.

The successful classification of HH-polarized images demonstrates the ViT model's robustness in recognizing sea ice formations, even when applied to a dataset with a different polarization from the training data. The attention visualizations indicate that the model focuses on relevant ice structures, which enhances interpretability and confirms the model's ability to generalize from VV to HH polarizations. However, a notable observation is the attention directed toward areas of missing data, suggesting a potential vulnerability in the model's classification process. This focus implies that, although the model effectively manages the shift between polarizations, the presence of large missing data regions could reduce its classification accuracy. This finding highlights the importance of preprocessing and ensuring complete datasets to optimize model performance.

The examples in Fig. 9 illustrate both the strengths and limitations of the ViT model when applied to HH-polarized data. The top two images demonstrate successful classifications, with the model identifying "sea ice" confidently at probabilities of 49% and 84%. The attention maps for these cases reveal a strong focus on ice features, such as distinct textures and structures, showcasing the model's ability to generalize to HH polarization, even when trained on VV-polarized data.

However, the lack of preprocessing in the raw images presents challenges that are evident in the bottom example. The grayscale contrasts among the images differ significantly, as seen in the much darker first image compared to the others. These inconsistencies arise from using raw data without standardizing brightness or contrast levels, making it harder for the model to adapt uniformly across samples. In addition, the presence of missing data further complicates classification. In the misclassified case, the model predicted "low wind area" instead of "sea ice," with a confidence probability of 41%. The attention map reveals that the model focused on regions of missing data, which it mistakenly associated with patterns typical of "low wind area."

2) Overall Evaluation on HV Polarizations: For HV polarization data (see Fig. 10), the ViT model's performance declined, achieving an accuracy of approximately 47%. Only 240 out of 512 samples were correctly identified as "sea ice," while 235 samples were misclassified as "low wind," and additional errors included 25 samples as "Atmospheric front," 10 as "icebergs," and 2 as "biological slicks." This elevated misclassification rate reflects the model's struggle with the reduced surface detail and lower contrast inherent to HV-polarized images.

The examples presented in Fig. 11 illustrate the ViT model's struggles with HV-polarized data. The second image from the top showcases a successful classification, where the model identified "sea ice" with 76% confidence. This result demonstrates the model's ability to process HV-polarized data under certain conditions when distinct features are present.

However, the other examples reveal significant challenges. In these cases, the true category "sea ice" was misclassified as "low wind area," with confidence probabilities of 81% and 61%. A key factor contributing to these errors is the lack of preprocessing in the raw images, leading to varying grayscale contrasts and the presence of missing data regions. These inconsistencies are evident in the images, where some appear darker or less defined compared to others. The attention maps for the misclassified examples show that the model focused heavily on regions of missing data, misinterpreting them as features typical of "low wind area."

IV. Discussion

A. Key Findings and Their Implications

This study demonstrates the effectiveness of ViTs in classifying geophysical phenomena in SAR imagery, significantly outperforming traditional CNN-based approaches. By leveraging self-attention mechanisms, ViTs capture long-range dependencies and structural patterns within images, leading to improved classification accuracy across multiple datasets. The results show that ViTs achieve higher recall, precision, and F1 scores compared to CNNs, particularly in categories that require complex pattern recognition, such as rain cells, oceanic fronts, and wind streaks.

A key contribution of this work is the evaluation of model generalization to different polarization datasets. While the ViT model was trained on VV-polarized SAR images, its performance on the AI4Arctic dataset, which includes HH- and HV-polarized images, revealed significant differences in classification accuracy. The model performed well on HH-polarized images (80% accuracy) but struggled with HV-polarized images (47% accuracy), indicating that surface scattering characteristics play a crucial role in SAR image classification.

B. Discussion of Polarization Performance

The performance of the ViT model across different polarizations highlights the significant role of scattering mechanisms in classification accuracy, especially when the training data and evaluation datasets differ in polarizations. The model demonstrated superior performance on HH-polarized images compared to HV-polarized images, underscoring the importance of texture and structural information in accurate classification. HH polarization emphasizes surface scattering, providing detailed texture and structural cues that make surface features like "sea ice" formations and roughness patterns more discernible (see Fig. S5 of the Supplementary Material). These similarities between VV

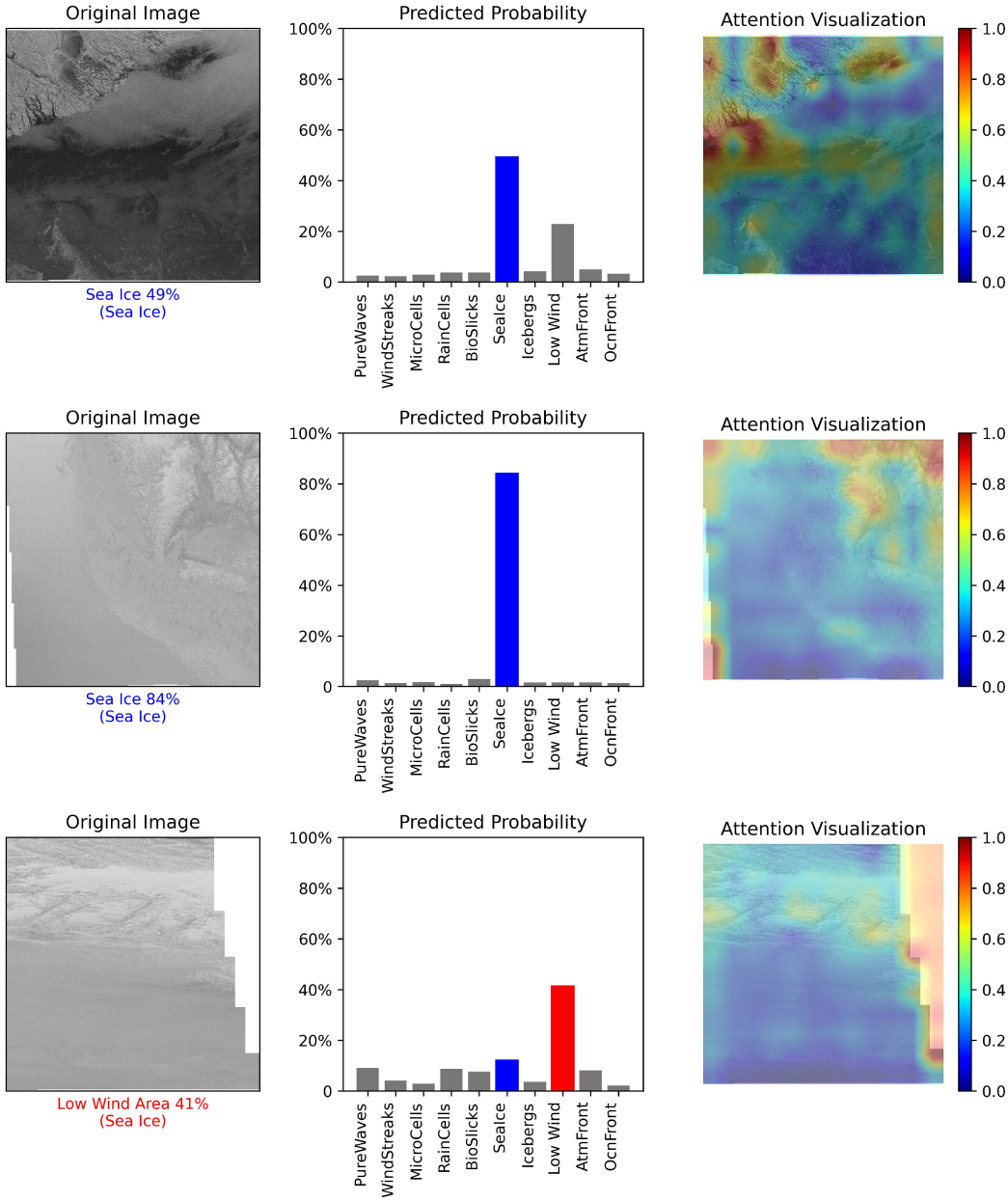

Fig. 9. *Examples of successful and inaccurate classifications from the HH-Polarized AI4Arctic Dataset Using the ViT Model.* This figure illustrates both correctly classified and misclassified examples, featuring original images, predicted probability distributions, and attention visualizations. The top two images represent successful classifications where the model accurately identified "sea ice," with confidence probabilities of 49% and 84%. The attention maps for these cases show the model's ability to effectively focus on relevant sea ice features, demonstrating strong generalization to HH-polarized data. However, the bottom image depicts a misclassification where the true category was "sea ice," but the model predicted "low wind area" with a confidence probability of 41%. Notably, differences in grayscale contrast, such as the significantly darker first image, highlight the lack of preprocessing in the raw data. In addition, the attention maps reveal that the model focused on regions with missing data, which were misinterpreted as "low wind area" features.

(used for training) and HH polarizations enable better generalization and classification accuracy for HH images. Typically, VV and HH polarizations differ only in contrast rather than overall structure, making this transition manageable for the model.

In contrast, HV polarization is more sensitive to volume scattering, resulting in images with less-defined surface details and lower contrast. This leads to a more uniform appearance, which limits the sharp textures and contrasts that the model relies on for accurate classification [48]. Consequently, the ViT model struggled with HV-polarized data, achieving lower accuracy due to the reduced visibility of surface features essential for distinguishing "sea ice" from other geophysical categories. Research by Turkar et al. [49] support this finding, showing that (HH, VV) polarization combinations outperform (HH, HV) and (VV, HV) combinations in land feature classification, indicating that HV channels generally lower classification accuracy.

A notable factor influencing misclassifications in both HH and HV datasets is the presence of regions without data in the

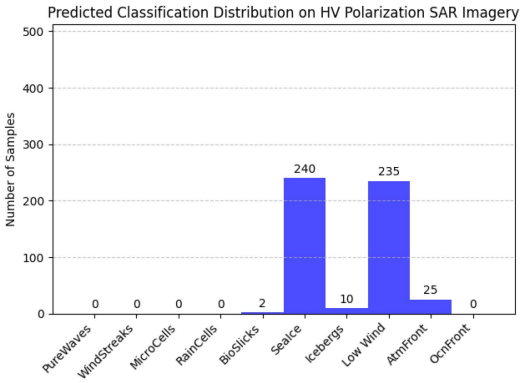

Fig. 10. Predicted classification distribution on HV polarization SAR imagery from the AI4Arctic sea ice challenge dataset.

raw images, which typically correspond to low-SNR areas where backscatter values were filtered or interpolated due to preprocessing steps. Since the model was trained on well-preprocessed VV-polarized data, which lacks such missing regions, it interprets these areas as visually similar to "low wind area" features. Attention maps reveal that the model disproportionately focuses on these data-deficient regions, contributing to incorrect predictions (see Figs. 9 and 11). This finding underscores the utility of attention mechanisms in identifying potential vulnerabilities and providing interpretable insights into the model's decision-making process.

The successful classification of HH-polarized images demonstrates the robustness of the ViT model in recognizing "sea ice" formations, even when applied to a dataset with a different polarization from the training data. The attention maps indicate that the model correctly focuses on relevant ice structures, confirming its ability to generalize from VV to HH polarizations. However, the maps also highlight the model's focus on regions with missing data, suggesting a potential vulnerability. This observation emphasizes the importance of preprocessing to ensure complete datasets for optimal model performance.

In contrast, the HV-polarized examples highlight the challenges the ViT model faces with this type of data. While some instances were successfully classified as "sea ice," many images were misclassified as "low wind area." The attention maps for these cases reveal that the model often focuses on regions with missing data, interpreting them as patterns similar to "low wind area." This behavior suggests that training on VV data, which lacks such ambiguous features, contributes to the misclassification. Incorporating similar ambiguous features into the training process could enhance robustness and reduce errors.

Despite the performance gap between HH and HV polarizations, the ViT model demonstrated its potential for generalization across different polarizations. However, the findings underscore the importance of aligning training data characteristics—such as polarization type and image quality—with the target application data. Models trained on a baseline polarization (e.g., VV) are more likely to excel with like-polarization (e.g., HH) but face challenges with cross-polarization (e.g., HV).

These results emphasize the inherent differences in the scattering mechanisms of VV/HH versus HV polarizations. VV and HH polarizations provide clearer texture and contrast due to surface scattering, enabling the model to leverage these features effectively. By contrast, HV polarization's sensitivity to volume scattering reduces detail and contrast, increasing the likelihood of misclassification, particularly in the presence of missing data. Attention maps reinforce the importance of interpretability tools in understanding model behavior, identifying weaknesses, and guiding improvements. Addressing these challenges through enhanced data preprocessing and incorporating training samples with regions without data could significantly improve classification accuracy and robustness in real-world SAR remote sensing applications.

C. Comparison With Previous Work

Compared to previous studies that utilized CNNs for SAR image classification, our results highlight ViTs' superior capability in structural classification tasks. While CNNs performed adequately on texture-based features, they struggled with identifying complex structures (see Tables S2 and S3 of the Supplementary Material), which ViTs successfully captured. Similar studies have reported that CNNs excel at localized pattern recognition but often fail to generalize to new datasets with different spatial resolutions and polarizations [50], [51]. Our findings reinforce the notion that ViTs provide a more robust and interpretable framework for SAR image analysis.

The cross-polarization evaluation also aligns with prior research demonstrating that SAR classification models trained on VV polarization tend to generalize better to HH than HV due to similar scattering mechanisms. Studies by Turkar et al. [49] suggest that VV and HH polarizations retain strong structural features, whereas HV polarization, dominated by volume scattering, lacks the necessary contrast for distinguishing geophysical categories.

D. Challenges and Limitations

Despite the strong performance of ViTs, several challenges remain as follows.

1) *Misclassifications*—Attention maps revealed that the model often focused on regions lacking valid data, leading to frequent misclassifications, particularly in HV-polarized images. These regions correspond to areas where backscatter values were either filtered out or interpolated during preprocessing, reducing the availability of reliable features for classification. In addition, dark areas in SAR imagery can arise from various environmental factors, including low wind conditions, biogenic slicks, or even oil spills, making accurate differentiation challenging. Properly distinguishing between these phenomena requires additional contextual information, such as ancillary meteorological data or multitemporal observations, which were not incorporated in this study. Furthermore, while some class assignments within the dataset may be ambiguous, we adhered to the provided labels to ensure dataset integrity and maintain comparability with existing benchmarks and prior research.

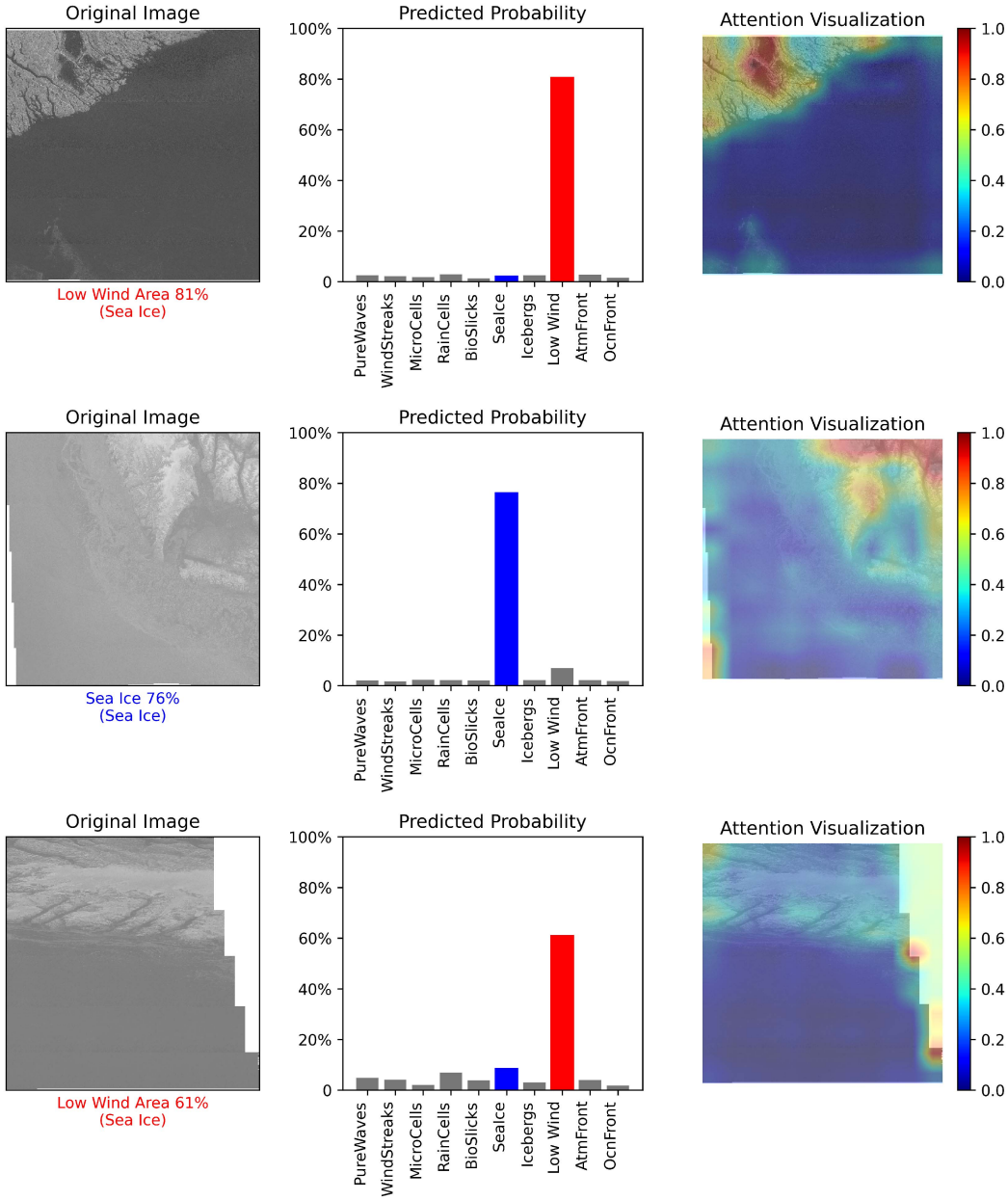

Fig. 11. *Examples of successful and inaccurate classifications from the HV-Polarized AI4Arctic Dataset Using the ViT Model.* This figure illustrates both correctly classified and misclassified examples of HV-polarized SAR images, including original images, predicted probability distributions, and attention visualizations. The second image from the top represents a successful prediction where the model accurately classified the image as "sea ice" with 76% confidence. However, the other images show misclassifications where the true category "sea ice" was incorrectly predicted as "low wind area," with confidence probabilities of 81% and 61%. Differences in grayscale contrast, evident across the images due to the lack of preprocessing, and the presence of missing data regions contributed to these misclassifications. The attention maps highlight that the model focused on these regions without data, interpreting them as patterns associated with "low wind area."

2) *Polarization-based limitations*—The significant drop in accuracy for HV-polarized images suggests that models trained on VV-polarized data may require additional fine-tuning or retraining on multipolarized data to handle cross-polarization discrepancies effectively. Since HV backscatter is dominated by volume scattering rather than surface roughness, the model may struggle with feature transfer between polarization modes. Furthermore, under high wind speed conditions, the radar backscatter from open water can be comparable to that from first-year ice (FYI), especially at near-range incidence angles in SAR imagery. In these cases, distinguishing FYI from open water can be challenging based on backscatter intensity alone.
3) *Dataset preprocessing and variability*—The AI4Arctic dataset was used in raw format, meaning that brightness and contrast variations were not normalized. This lack of preprocessing introduced inconsistencies that likely

contributed to misclassification errors. While contrast enhancement can aid feature extraction, excessive preprocessing may alter the statistical properties of the dataset, necessitating a careful balance between normalization and preserving raw SAR characteristics.

E. Future Directions

To further enhance classification performance and generalization ability, several future research directions are proposed as follows.

1) *Multipolarization training:* Incorporating multipolarization datasets (VV, HH, and HV) during training could improve model robustness across different SAR sensors.
2) *Self-Supervised learning:* Recent advancements in self-supervised ViTs have demonstrated strong performance in generalization tasks. Applying contrastive learning techniques could reduce the dependence on large labeled datasets.
3) *Enhanced preprocessing techniques:* Normalizing brightness and contrast variations in raw datasets may help reduce misclassification errors caused by variable image quality.
4) *Integration with auxiliary data:* Incorporating additional metadata (such as temperature, wind speed, and distance to land) could improve classification accuracy by providing context for ambiguous image regions.

V. Conclusion

This study demonstrates the effectiveness of ViT for classifying geophysical phenomena in SAR imagery, offering superior performance compared to CNNs. By leveraging self-attention mechanisms, ViTs captured long-range dependencies and structural features, improving classification accuracy and interpretability. The TenGeoP-SARwv dataset served as the primary training and evaluation source, while the AI4Arctic sea ice challenge dataset tested the model's generalization to different polarizations and spatial resolutions.

Results show that ViTs significantly outperform CNNs, achieving average recall, precision, and F1 scores of 0.985, 0.984, and 0.986, respectively. The polarization evaluation highlighted that the model, trained on VV-polarized SAR data, generalized well to HH-polarized images (80% accuracy) but struggled with HV-polarized images (47% accuracy) due to differences in scattering mechanisms. Misclassifications often occurred in regions with missing data, where low-SNR areas were incorrectly interpreted as low-wind zones, emphasizing the need for preprocessing improvements to enhance robustness.

Despite strong performance, challenges remain. Multipolarization training (VV, HH, HV) could improve adaptation across SAR sensors, while self-supervised learning may reduce dependence on labeled datasets. Addressing low-SNR misclassifications and standardizing preprocessing techniques could further enhance reliability. Integrating auxiliary metadata, such as wind speed and temperature, may provide additional contextual cues to resolve ambiguous cases.

This work establishes ViTs as a powerful tool for SAR-based geophysical classification, with potential applications in oceanography, climate monitoring, and remote sensing. Future research should focus on expanding ViT architectures, improving generalization across diverse SAR datasets, and refining attention-based learning frameworks to advance automated geophysical analysis.

Code and Data Availability

The scripts used in this research are available on GitHub: https://github.com/JunfeiXia/SARimage_Classification. The data used in this research are available on SEANOE and DTU: https://www.seanoe.org/data/00456/56796/ https://data.dtu.dk/collections/AI4Arctic_Sea_Ice_Challenge_Dataset/6244065.

References

[1] A. P. Cracknell, "The development of remote sensing in the last 40 years," *Int. J. Remote Sens.*, vol. 39, no. 23, pp. 8387–8427, 2018.
[2] L. Alparone, B. Aiazzi, S. Baronti, and A. Garzelli, *Remote Sensing Image Fusion*. Boca Raton, FL, USA: CRC Press, 2015.
[3] X. Li, Z. Du, Y. Huang, and Z. Tan, "A deep translation (GAN) based change detection network for optical and SAR remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 179, pp. 14–34, 2021.
[4] G. Duchossois, "The ERS-1 mission objectives," *ESA Bull.*, no. 65, pp. 16–25, 1991.
[5] C. Francis et al., "The ERS-2 spacecraft and its payload," *ESA Bull.*, vol. 83, pp. 13–31, 1995.
[6] J. Louet and S. Bruzzi, "ENVISAT mission and system," in *Proc. 1999 IEEE Int. Geosci. Remote Sens. Symp.*, 1999, pp. 1680–1682.
[7] A. Arnaud et al., "ASAR ERS interferometric phase continuity," in *Proc. 2003 IEEE Int. Geosci. Remote Sens. Symp.*, 2003, pp. 1133–1135.
[8] R. Werninghaus and S. Buckreuss, "The TerraSAR-X mission and system design," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 2, pp. 606–614, Feb. 2010.
[9] C. C. Wackerman, K. S. Friedman, W. G. Pichel, P. Clemente-Colón, and X. Li, "Automatic detection of ships in RADARSAT-1 SAR imagery," *Can. J. Remote Sens.*, vol. 27, no. 5, pp. 568–577, 2001.
[10] P. Cheng and T. Toutin, "RADARSAT-2 data," *GeoInformatics*, vol. 13, no. 5, 2010, Art. no. 22.
[11] L. Zhao et al., "China's Gaofen-3 satellite system and its application and prospect," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 11019–11028, 2021.
[12] P. Potin et al., "Sentinel-1 mission status," in *Proc. 11th Eur. Conf. Synthetic Aperture Radar*, 2016, pp. 1–6.
[13] C. Bai, S. Zhang, X. Wang, J. Wen, and C. Li, "A multichannel-based deep learning framework for ocean SAR scene classification," *Appl. Sci.*, vol. 14, no. 4, 2024, Art. no. 1489.
[14] N. Hashim and J. R. A. Hamid, "Multi-level image segmentation for urban land-cover classifications," in *Proc. IOP Conf. Series: Earth Environ. Sci.*, 2021, Art. no. 012024.
[15] K. A. Call, J. T. Hardy, and D. O. Wallin, "Coral reef habitat discrimination using multivariate spectral analysis and satellite remote sensing," *Int. J. Remote Sens.*, vol. 24, no. 13, pp. 2627–2639, 2003.
[16] H. Ye et al., "Spectral classification of the yellow sea and implications for coastal ocean color remote sensing," *Remote Sens.*, vol. 8, no. 4, 2016, Art. no. 321.
[17] P. M. Atkinson and P. Lewis, "Geostatistical classification for remote sensing: An introduction," *Comput. Geosci.*, vol. 26, no. 4, pp. 361–371, 2000.
[18] N. Fiorentini, M. Maboudi, P. Leandri, M. Losa, and M. Gerke, "Surface motion prediction and mapping for road infrastructures management by PS-InSAR measurements and machine learning algorithms," *Remote Sens.*, vol. 12, no. 23, 2020, Art. no. 3976.
[19] X. X. Zhu et al., "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[20] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS J. Photogrammetry Remote Sens.*, vol. 152, pp. 166–177, 2019.
[21] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3735–3756, 2020.
[22] R. Ressel, A. Frost, and S. Lehner, "A neural network-based classification for sea ice types on X-band SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 7, pp. 3672–3680, Jul. 2015.
[23] H. Li et al., "Adversarial examples for CNN-based SAR image classification: An experience study," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1333–1347, 2020.
[24] A. S. Li, V. Chirayath, M. Segal-Rozenhaimer, J. L. Torres-Perez, and J. van den Bergh, "NASA NeMO-net's convolutional neural network: Mapping marine habitats with spectrally heterogeneous remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5115–5133, 2020.
[25] H. Yang and W. Yang, "GSCCTL: A general semi-supervised scene classification method for remote sensing images based on clustering and transfer," *Int. J. Remote Sens.*, vol. 43, pp. 5976–6000, 2022.
[26] P. K. Sethy, "Identification of wheat tiller based on AlexNet-feature fusion," *Multimedia Tools Appl.*, vol. 81, no. 6, pp. 8309–8316, 2022.
[27] O. Kechagias-Stamatis and N. Aouf, "Automatic target recognition on synthetic aperture radar imagery: A survey," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 36, no. 3, pp. 56–81, Mar. 2021.
[28] X. Han, Y. Zhong, L. Cao, and L. Zhang, "Pre-trained alexnet architecture with pyramid pooling and supervision for high spatial resolution remote sensing image scene classification," *Remote Sens.*, vol. 9, no. 8, 2017, Art. no. 848.
[29] Y. Zhou and M. Wang, "Remote sensing image classification based on AlexNet network model," in *Proc. Int. Conf. Frontier Comput.*, Springer, 2020, pp. 913–918.
[30] M. Castelluccio, G. Poggi, C. Sansone, and L. Verdoliva, "Land use classification in remote sensing images by convolutional neural networks," 2015, *arXiv:1508.00092*.
[31] M. Ye et al., "A lightweight model of VGG-16 for remote sensing image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 6916–6922, 2021.
[32] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 12116–12128.
[33] L. Jiao et al., "Transformer meets remote sensing video detection and tracking: A comprehensive survey," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 1–45, 2023.
[34] Y. Yuan and L. Lin, "Self-supervised pretraining of transformers for satellite image time series classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 474–487, 2020.
[35] S. Rubab et al., "A novel network level fusion architecture of proposed self-attention and vision transformer models for land use and land cover classification from remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 13135–13148, 2024.
[36] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[37] C. Wang et al., "A labelled ocean SAR imagery dataset of ten geophysical phenomena from sentinel1 wave mode," *Geosci. Data J.*, vol. 6, pp. 105–115, 2019.
[38] C. Wang et al., "Classification of the global sentinel-1 SAR vignettes for ocean surface process studies," *Remote Sens. Environ.*, vol. 234, 2019, Art. no. 111457.
[39] Y. Du, W. Song, Q. He, D. Huang, A. Liotta, and C. Su, "Deep learning with multi-scale feature fusion in remote sensing for automatic oceanic eddy detection," *Inf. Fusion*, vol. 49, pp. 89–99, 2019.
[40] C. Wang et al., "Automated geophysical classification of sentinel-1 wave mode SAR images through deep-learning," in *Proc. 2018 IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 1776–1779.
[41] Z. Gan, R. Henao, D. Carlson, and L. Carin, "Learning deep sigmoid belief networks with data augmentation," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2015, pp. 268–276.
[42] J. Buus-Hinkler et al., "Raw AI4Arctic sea ice challenge dataset," Nov. 2022. [Online]. Available: https://data.dtu.dk/articles/dataset/Raw_AI4Arctic_Sea_Ice_Challenge_Dataset/21284967
[43] B. Wu et al., "Visual transformers: Token-based image representation and processing for computer vision," 2020, *arXiv:2006.03677*.
[44] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. 2009 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[45] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," 2020, *arXiv:2005.00928*.
[46] W. Ramirez, P. Achanccaray, and M. A. Pacheco, "A comparative study of deep learning architectures for classification of natural and human-made sea events in SAR images," *Discover Artif. Intell.*, vol. 2, no. 1, 2022, Art. no. 1.
[47] J. Maurício, I. Domingues, and J. Bernardino, "Comparing vision transformers and convolutional neural networks for image classification: A literature review," *Appl. Sci.*, vol. 13, no. 9, 2023, Art. no. 5521.
[48] S. Ufermann and R. Romeiser, "Numerical study on signatures of atmospheric convective cells in radar images of the ocean," *J. Geophysical Res.: Oceans*, vol. 104, no. C11, pp. 25707–25719, 1999.
[49] V. Turkar, R. Deo, Y. Rao, S. Mohan, and A. Das, "Classification accuracy of multi-frequency and multi-polarization SAR images for various land covers," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 3, pp. 936–941, Jun. 2012.
[50] K. Lu, Y. Xu, and Y. Yang, "Comparison of the potential between transformer and CNN in image classification," in *Proc. 2nd Int. Conf. Mach. Learn. Comput. Appl.*, 2021, pp. 1–6.
[51] H.-P. Wei, Y.-Y. Deng, F. Tang, X.-J. Pan, and W.-M. Dong, "A comparative study of CNN-and transformer-based visual style transfer," *J. Comput. Sci. Technol.*, vol. 37, no. 3, pp. 601–614, 2022.

Junfei Xia received the B.S. degree in oceanography from Nanjing University, Nanjing, China, in 2017, and the M.S. degree in ocean engineering in 2019 from the University of Miami, Coral Gables, FL, USA, where he is currently working toward the Ph.D. degree in oceanography.

His research interests include applications of machine learning and deep learning in physical oceanography, with an emphasis on interpolation and classification using drifter data and synthetic aperture radar (SAR) imagery, deep learning, computer vision, dynamic systems, drifter data analysis, SAR image classification, and submesoscale ocean dynamics.

Roland Romeiser (Senior Member, IEEE) received the Dipl.-Phys. degree from the University of Bremen, Bremen, Germany, in 1990 and the Dr.rer.nat. and Habilitation degrees from the University of Hamburg, Hamburg, Germany, in 1993 and 2007, respectively.

From 1990 to 2008, he was with the Institute of Oceanography, University of Hamburg. From August 1998 to July 1999, he spent a year with the Applied Physics Laboratory, Johns Hopkins University, Laurel, MD, USA, as a Feodor Lynen Fellow of the Alexander von Humboldt Foundation. In April 2008, he joined the Rosenstiel School of Marine and Atmospheric Science, University of Miami, Coral Gables, FL, USA, where he is now a full Professor. He has wide experience in the field of remote sensing of ocean currents, waves, and winds by airborne and spaceborne microwave radars. His recent research interests include advanced synthetic aperture radar processing and algorithm development for current and wave retrievals.

Dr. Romeiser was an Associate Editor for IEEE JOURNAL OF OCEANIC ENGINEERING from 2000 to 2020 and IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING from 2014 to 2017.

Wei Zhang received the B.S. degree in ocean science from Nanjing University, Nanjing, China, in 2015, and the Ph.D. degree in meteorology and physical oceanography from the University of Miami, Coral Gables, FL, USA, in 2020.

He is currently a Research Scientist with the Cooperative Institute for Marine and Atmospheric Studies, University of Miami's. Prior to this, he was a Postdoctoral Research Associate with Princeton University, Princeton, NJ, USA, and NOAA's Geophysical Fluid Dynamics LaboratoryPrinceton. He also held a Visiting Research Position with NOAA's Global Systems Laboratory from 2021 to 2024. He has authored or coauthored more than 20 research papers. His work improves understanding of the limits of atmospheric predictability and the signal-to-noise paradox in climate models, and advances the use of high-resolution coupled models and data-driven methods in climate science. His research interests include climate modeling and prediction, extreme weather, and the application of machine learning in Earth system science.

Dr. Zhang is an Associate Editor for *Weather and Forecasting*, and is on the editorial boards of *Acta Meteorologica Sinica*, and the *Journal of Meteorological Research, Atmosphere, and Frontiers in Environmental Science*.

Tamay Özgökmen received the B.S. degree in mechanical engineering from Bosphorus University, Istanbul, Tü$\breve{u}$rkiye, in 1988, the M.S. degree in mechanical engineering from the University of Miami, Coral Gables, FL, USA, in 1990, and the Ph.D. degree in mechanical engineering from Dartmouth College, Hanover, NH, USA, in 1995.

In 1995, he joined the Rosenstiel School of Marine and Atmospheric Science, University of Miami as the Rosenstiel Postdoctoral Scientist, was awarded tenure in 2006, and promoted to Full Professor in 2010. He served in large collaborative projects, such as National Science Foundation's Climate Process Team (2003–2009), Office of Naval Research's Lateral Mixing (2009–2013), and CALYPSO (2018-present) programs, Department of Defense awards (called Multi University Research Initiatives) on 4-D Coherent Structures (2010–2017) and Machine Learning as well as DARPA's Ocean of Things program to release thousands sensor platforms in the ocean. In the aftermath of the Deepwater Horizon spill in 2010, he assembled the Consortium of Advanced Research of Hydrocarbon Transport in the Ocean to conduct multiplatform expeditions in the Gulf of Mexico, laboratory experiments and leading-edge computations, resulting in some 220 peer-reviewed publications by 400 unique authors and many of Ph.D. degrees awarded by institutions across the United States. He has authored 150 peer-reviewed publications on numerical and observational explorations of ocean turbulence. He is also one of the inventors of the biodegradable drifter, which was made commercially available and is now used by researchers around the world.