Imperial College London

Department of Earth Science and Engineering

MSc in Environmental Data Science and Machine Learning

Independent Research Project

Final Report

Evaluating the Ability of Self-Supervised Learning to Identify Submesoscale Geophysical Processes using Wave Mode SAR Data

by

Matthew Barrett

Repository: `ese-ada-lovelace-2024/irp-mb1024`

Supervisors:

Björn Rommen

Matthew Piggott

August 2025

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Björn Rommen,Matthew Piggott and Andreas Theodosiou For their guidance, encouragement, and support throughout this project. I am also grateful to the wider community at the European Space Research and Technology Centre (ESTEC) and at Imperial College London for their valuable input and encouragement. I would particularly like to thank Marijan Beg for his logistical support during the MSc programme and Independent Research Project. Finally, I would like to thank my family and friends for their continuous support and encouragement throughout my studies.

AI Acknowledgement Statement

During the course of this Independent Research Project, I made use of the following generative AI tool:

- **ChatGPT-4o** (OpenAI) `https://chat.openai.com` Used to aid my understanding of the methods applied in the research, assist with debugging, and support the formulation of the model architecture structure.

The submitted work is my own and it reflects my own understanding and effort as I declared by signing the Academic Integrity Declaration.

Contents

1 Abstract

The ocean plays a central role in regulating Earth's climate system, making detailed observations of its surface essential. Sentinel's WV offers a rich but underused dataset of oceanic and atmospheric phenomena, due to the lack of automated tools for classification of these processes. This study investigates the potential of self-supervised learning (SSL) as an approach to classify these phenomena without manual annotation. Two contrastive SSL frameworks, SimCLR and DINO, were trained on more than 50,000 unlabeled SAR WV images and were evaluated using a linear probe and clustering. SimCLR with intermediate augmentations achieved the best performance (F-1 - 0.885), approaching the accuracy of an ImageNet-pretrained ResNet18 (F-1 - 0.91), and substantially outperforming the model trained without SSL pretraining. Although distinct clusters could not be formed, images near cluster centres showed consistent geophysical phenomena, indicating that the model learned semantically meaningful features. These findings indicate that SSL can significantly reduce manual labelling requirements in WV SAR classification, though not fully replace annotation due to the inherent ambiguity of scenes containing multiple overlapping processes. Overall, the results position contrastive SSL, as a promising approach for the Earth Observation Division at ESTEC, enabling large-scale and cost-efficient exploitation of Sentinel-1 WV data and providing a methodological foundation to fully benefit from the enhanced finescale monitoring capabilities of the upcoming Harmony mission.

2 Introduction

2.1 Background

The ocean plays a central role in Earth's climate system, acting as a major regulator of global energy and water cycles. The ocean surface, in particular, serves as the interface for the exchange of heat, moisture, and momentum between the atmosphere and the ocean. Comprehensive measurements of this interface are essential to understand ocean atmosphere interactions and for the development of high-resolution climate models (Topouzelis Kitsiou., 2015). Remote sensing is a key tool for observing the Earth's surface and monitoring changes in the climate system, it can be broadly divided into two categories: passive sensing, which relies on external energy sources such as sunlight, and active sensing, which uses its own energy source (Yang et al., 2013).

This research focuses on spaceborne Synthetic Aperture Radar (SAR), a form of active remote sensing capable of acquiring high-resolution sea surface backscatter data regardless of weather or lighting conditions (Yee Kit Chan et al., 2008). SAR can be modulated by a range of physical processes including ocean swell (Collard et al, 2009), upper ocean processes (Jia et al., 2019) and rain cells (Alpers et al, 2016), making it a perfect tool to capture ocean-atmosphere interactions. Of particular interest for this research is Sentinel-1's Wave mode (WV) which captures 20×20 km SAR image vignettes at alternating incidence angles of 23° and 37°, with a ground resolution of approximately 5 meters. These vignettes are acquired every 100 km along the ground track of the satellite (Dai et al. 2022).

WV images exhibit oceanic and atmospheric structures that typically span 1 to 10 km (submesoscale region) which aid in understanding small scale ocean-atmosphere interactions (Nuijens et al., 2024). However, these features are often too small and short-lived to be effectively captured by traditional oceanographic sensors. The high spatial resolution of Sentinel-1's WV enables these phenomena to be observed in unprecedented detail. Figure 1 illustrates examples of these submesoscale processes, highlighting the value of SAR WV imagery for investigating submesoscale dynamics in the open ocean.

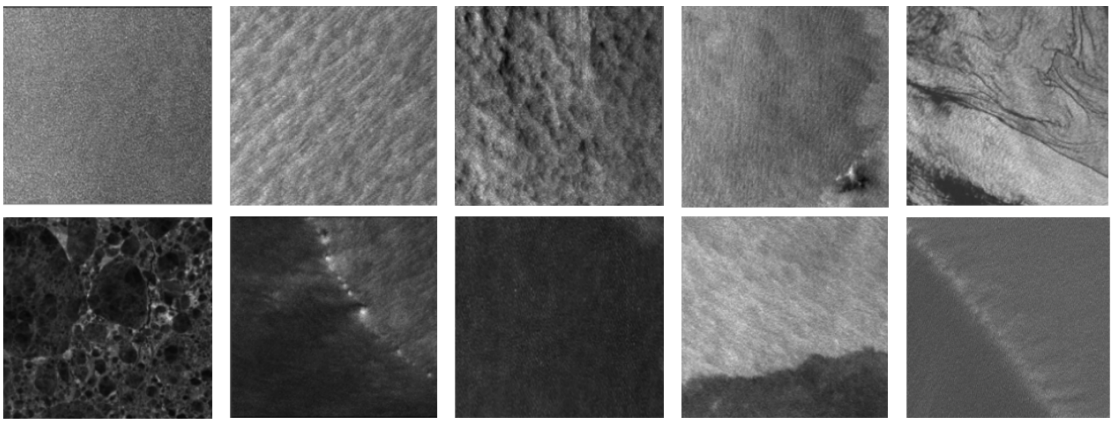

Figure 1: Sentinel-1 WV SAR images illustrating the range of phenomena represented in the dataset. From left to right, top row: pure ocean waves, wind streaks, micro-convective cells, rain cells, and biological slicks. Bottom row: sea ice, icebergs, low-wind area, atmospheric front, and oceanic front.

Classification of these processes by WV remains under used due to the lack of automated tools for detecting such features. Traditional classification methods rely on manual annotation, which is both time-consuming and inefficient. Deep learning has been widely applied to automate the classification of remote sensing data, however, most existing models rely heavily on supervised learning (X. Zhu et al, 2017). This presents a major challenge in the context of SAR imagery, where high-quality annotations are scarce and costly to produce.

2.2 Objective

This project explores an alternative approach to manually labelling data, through self-supervised learning (SSL). SSL is a branch of unsupervised learning in which a model is trained on a pretext task, where the overall goal of the model is not to solve the pretext task directly, but to use it as a means to learn meaningful structures of the data, that can be used to perform downstream tasks that aim to match the performance levels of supervised learning (Gui et al., 2024). Although SSL is relatively new in the SAR domain, it offers strong potential for learning from unlabeled data (Tao C et al., 2023)).

SSL can be broadly divided into generative or contrastive learning, or a combination of the two methods (Liu et al., 2020). When used in computer vision contrastive learning typically works by applying semantic-preserving transformations to input images. These augmented images are then passed through an encoder to generate

embeddings, and then optimised so that embeddings from the same image are closer together than those from different images (Jaiswal et al., 2021). This approach tends to perform better in tasks related to classification than generative learning (Ji et al., 2023).

Generative learning alternatively uses a different method, it trains a model to reconstruct the original input (e.g., pixels, tokens) from corrupted or masked versions (Zhang et al., 2022). This approach focuses more on pixel-level reconstruction as opposed to capturing high-level semantic structures. Specifically, for WV SAR data, the limited pixel-to-pixel variation associated with these images means that generative models can score a low reconstruction loss by averaging all the pixels in each block, causing model collapse or the learning of trivial features.

Due to the time constraints of this research, this study focuses only on contrastive SSL for WV SAR classification. To evaluate the performance of SSL for this task, two methods will be used, those being SimCLR (Simple Contrastive Learning of Representations) and DINO (Self-Distillation with No Labels). Both of these methods have demonstrated success in previous applications of SSL (Chen et al., 2020), (Caron et al., 2021) and also differ in architecture and pretext task formulations enabling a broader analysis of contrastive SSL.

SimCLR is a well established, simple contrastive framework in which two augmented views of the same image are encoded with a convolutional neural network which extracts features in a local to global manner and projected into a latent space and then optimised to maximise the agreement between positive pairs whilst separating negative pairs. Due to the negative sampling required in training, it relies heavily on a large batch size. (Chen et al., 2020).

DINO, in contrast, is a self distillation based contrastive learning method which does not require negative sampling. Instead it uses a Vision Transformer which allows the model to analyse the data from a global perspective from the outset. DINO uses a teacher-student architecture where the student network learns to match the teacher's output distribution across multiple local and global augmented views. This design allows DINO to capture more global, high-level semantic structures and remain effective even with smaller batch sizes (Caron et al., 2021).

2.3 Research Aim

The aim of this research is to evaluate how effectively self-supervised methods can extract meaningful features related to geophysical processes in the submesoscale region of the ocean directly from Sentinel-1 WV imagery, using only a minimal dataset and without supervision. If successful, the model developed in this study could be further fine-tuned with larger datasets and extended training times to create a fully functioning classifier capable of identifying geophysical features in WV images. Such a model could support research projects within the Earth Observation Division at the European Space Research and Technology Centre (ESTEC), enabling large-scale, cost-effective analysis of ocean–atmosphere interactions. Ultimately, with the advent of the Harmony mission extending Sentinel-1 observation capability from late 2029 onwards, this would improve environmental monitoring and contribute to a better understanding of the changing climate system.

3 Methods

3.1 Data Preprocessing

The WV data used in this study was obtained from the Alaskan Satellite Facility database using a script provided by ESTEC. The data covers globally distributed ocean imagery acquired between 2024 and 2025 across all months. The data was downloaded in NetCDF format, with the sigma0 backscatter imagery serving as the primary input. Sigma0 is the radar backscatter coefficient, representing the amount of radar signal reflected back to the sensor per unit area, normalised for incidence angle. It is expressed in decibels (dB) and is influenced by surface roughness, dielectric properties, and geometry (ESA, 2013). In the context of Sentinel-1 WV imagery, sigma0 provides information about the state of the ocean surface, including features caused by wind, waves, and submesoscale atmospheric–ocean interactions. A spatial resolution of 100 metres was selected, as it was deemed appropriate for capturing phenomena at the submesoscale region without unnecessarily increasing data volume. The sigma0 bands were extracted from the NetCDF files in the form of single-channel grayscale images, which serve as the input for self-supervised model training. The final data set contains 50,832 of these images.

3.2 Normalisation

As the focus of this research is based on image classification, the normalisation was carried out locally for each image using a Min/Max method, which scales all values between 0 (minimum of the dataset) and 1 (maximum of the dataset). The formula is:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}, \quad x' \in [0, 1] \tag{1}$$

x Original value

$x_{\min}, x_{\max}$ Minimum and maximum values in the data

x' Normalised value

Global normalisation was avoided because SAR images vary widely in absolute backscatter intensity. Applying a single global scale would result in most images

appearing nearly black, masking the atmospheric processes present in the image. The 0th and 100th percentile pixel intensities of each image were clipped to the 1st and 99th percentiles, respectively, to limit the influence of extreme outliers in training that are irrelevant to atmospheric features such as specular reflections from man-made structures like boats.

3.3 Size

All images were cropped to 200×200 pixels during data augmentation. The majority of the images ranged from 200–220 pixels in both height and width, and any images smaller than 200x200 pixels were padded with the per-image mean intensity to avoid biasing the normalisation process and to minimise the impact on subsequent augmentations. NaN values within each image were filled using the nearest-neighbour method, preserving the local structure of the surrounding pixels and ensuring that these filled values would become insignificant during model training.

3.4 Data augmentation

Data augmentations can be seen as one of the most important factor within SSL. The goal of applying augmentations is to suppress insignificant features within an image while retaining the meaningful ones. In this research, the augmentations used combined both general practices adopted in previous SSL frameworks and domain-specific considerations of Sentinel-1 WV imagery.

Previous SSL frameworks specify random cropping and colour distortion as required augmentations, representing spatial and distributional transformations respectively (Chen et al., 2020). These are necessary because neural networks can otherwise exploit simple spatial and colour distributions in the augmented images to achieve trivial solutions. However, colour distortion was deemed not relevant in this study since SAR WV imagery is grayscale. Instead, photometric augmentations are used to alter the overall brightness and contrast of the image. A potential source of trivial learning within SAR WV arises from the intensity of radar backscatter, due to a large variance in absolute backscatter. This issue was mitigated during preprocessing through local normalisation, ensuring that the network could not rely on raw backscatter intensity as a shortcut for learning.

Augmentation design is challenging, and in practice, a wide range of augmentations

are often applied randomly to images (Liu et al., 2023). Due to this, the approach taken in this study was to apply progressively stronger spatial transformations alongside increasing levels of photometric augmentations, to align with the goal of contrastive learning in suppressing insignificant features and learning invariant representations of the image.

Table 1: Augmentations per model and augmentation strength.

Model	**Set**	**Augmentations**
SimCLR	Light	Rotation; crop; blur
SimCLR	Intermediate	Rotation; crop; horizontal flip; blur; translation
SimCLR	Heavy	Rotation; crop; horizontal flip; blur; sharpness; random erasing
DINO	Light	Rotation; crop; horizontal flip; blur
DINO	Heavy	Rotation; crop; horizontal flip; blur; translation; sharpness

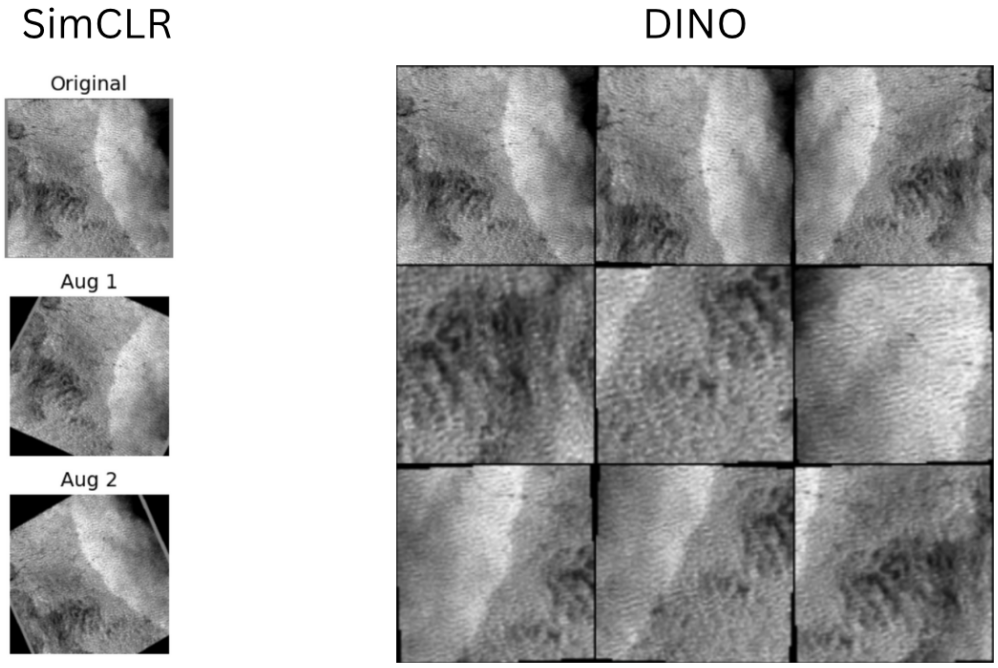

Figure 2: Examples of SimCLR (left) and DINO (right) augmentations applied to a SAR WV image. In SimCLR, the original image is followed by two augmented views. In DINO, the top row shows the original image and two global crops, while the bottom two rows contain six local crops

3.5 SimCLR Architecture

Encoder: The encoder backbone is a convolutional neural network composed of five blocks, each containing two convolutional layers with increasing filter sizes: 32, 64, 128, 256, and 512. Each layer is followed by batch normalisation to standardise the activations to ensure a stable mean and variance in each layer and a Rectified Linear Unit (RELU) activation function to introduce non-linearity into the model. Max pooling is added after each block to reduce the spatial dimensions while retaining the most important information. The final adaptive average pooling layer compresses spatial information into a 512-dimensional feature vector.

Projection head: The projection head reduces the encoder's feature vector to 128 dimensions before applying the contrastive loss. This design choice follows the SimCLR framework (Chen et al., 2020), which introduced a non-linear projection head to improve the effectiveness of contrastive training. The authors observed that while the contrastive loss benefits from operating in the projection space, the encoder representations remain more suitable for downstream tasks. This is hypothesized due to the information loss introduced by the contrastive objective. Since performance across different projection dimensions is relatively consistent, the feature vector is reduced to 128 dimensions in this study to improve training efficiency.

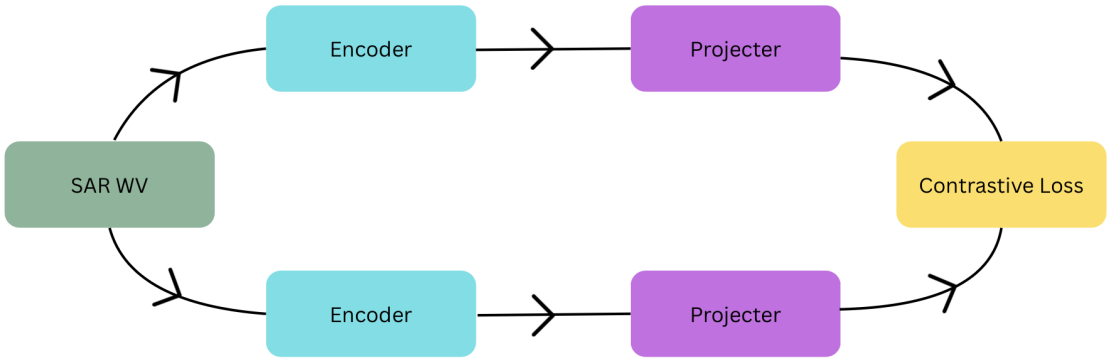

Figure 3: Overview of the SimCLR framework. Two augmented views of each image are encoded, projected into a latent space, and optimised with a contrastive loss to maximise agreement between positive pairs while separating negatives.

Loss Function: For each image in the batch, two augmented views are generated and passed through the encoder–projection head network, producing embeddings i and j. These embeddings are then concatenated, L2-normalised for model stability, and used to compute a cosine similarity matrix between all possible pairs. The temperature parameter (set to 0.5) moderates the penalty on hard negatives as to not punish

negatives which are similar to the positive and are likely in the same downstream class. Over-penalising such cases could unnecessarily separate semantically similar samples, which is undesirable for classification tasks (Wang et al., 2020). A softmax cross-entropy loss is used, where each embedding must correctly identify its corresponding positive within the batch. By minimising this loss, the network learns an embedding space in which semantically similar samples are close together and dissimilar samples are far apart.

$$\mathcal{L}_{\text{SimCLR}} = \frac{1}{2N} \sum_{i=1}^{2N} -\log \frac{\exp(s(z_i, z_{j(i)})/\tau)}{\sum_{k=1 k \neq i}^{2N} \exp(s(z_i, z_k)/\tau)}. \tag{2}$$

where:

N Number of original images in the batch (yielding $2N$ augmented views).

z_i Embedding vector of the i-th augmented view after the encoder–projection head.

$j(i)$ Index of the positive pair corresponding to i (the other augmentation of the same image).

$s(z_i, z_k)$ Similarity function between embeddings, typically cosine similarity.

τ Temperature parameter that rescales similarities before the softmax, controlling the sharpness of the distribution.

3.6 DINO Architecture

Patch Embedding: Each 200×200 image is divided into non-overlapping 20×20 patches, resulting in 100 patches per image. A convolutional layer is applied to each patch in the image, which projects each patch into a 512-dimensional embedding vector, and learnable positional embeddings are added to retain spatial context across patches. This token sequence forms the input to the transformer encoder.

Transformer Encoder: A special classification token (CLS) is prepended to the sequence of patch embeddings, providing a learnable vector that serves as a global summary of the image. After encoding, this token is used as the image-level representation. The sequence, including the CLS token and patch embeddings, is then passed through six Transformer blocks, each following a pre-normalisation structure. In each block, the patch tokens are first normalised and then passed through a multi-head

self-attention layer with eight heads, allowing each patch to attend to all others and capture global contextual relationships. A residual connection ensures stable gradient flow and helps preserve information across layers. Following attention, the tokens are again normalised and processed by a multi-layer perceptron (MLP) with a hidden dimension equal to four times the embedding size ($512 \rightarrow 2048$), using Gaussian Error Linear Unit (GELU) activation for non-linearity. Another residual connection adds the MLP output back to the original input of this stage. Stacking six such blocks enables the model to build progressively richer, long-range dependencies between image regions.

Projection Head: After encoding, only the CLS token is retained as the image-level representation. It is passed through a projection head made of two Linear–GELU layers ($512 \rightarrow 512 \rightarrow 512$). Gelu activation is used here to follow best use for vision transformers to make optimisation more stable and smoothly suppress negative samples. This is followed by a weight-normalised linear layer that maps to 2,048 prototype vectors ($512 \rightarrow 2048$). These prototypes act as cluster centres in the feature space, allowing the model to predict a distribution over the prototype as opposed to directly matching embeddings.

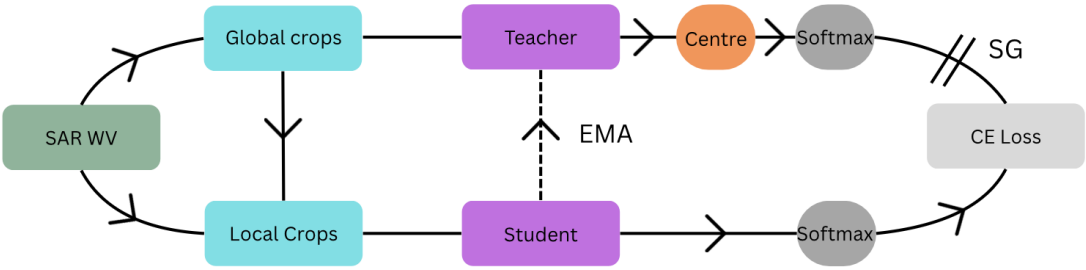

Figure 4: Overview of the DINO framework. Multiple augmented views of each image are generated, with the student network processing both global and local crops, while the teacher network processes only global crops. The student is trained to match the teacher's probability distribution across views using a cross-view cross-entropy loss, with teacher outputs stabilized through centering and a momentum update.

Loss: For each augmented view, the teacher's pre-softmax logits are centered and temperature-scaled to 0.04 to produce confident targets, then converted into a fixed target distribution via softmax, gradients are stopped through the teacher branch. The student logits are scaled by a higher temperature (0.1) to ensure stability and prevent over-penalisation of mismatches. They are then passed through log-softmax.

The loss is defined as the cross-entropy between the teacher distribution of each view (computed on only global crops) and the student distribution of all views (both global and local crops), skipping same-view pairs. This cross-view cross-entropy is averaged across all pairs and samples in the batch. After each forward pass, a running center (EMA) is updated by taking the batch mean of the teacher logits (pre-softmax) and blending it with the previous center using momentum, which stabilises the targets over time.

The DINO loss averages cross-entropy over student–teacher view pairs:

$$\mathcal{L}_{\text{DINO}} = -\frac{1}{N}\sum_{n=1}^{N}\frac{1}{|V|\,|U|}\sum_{v\in V}\sum_{u\in U\,u\neq v}\sum_{k=1}^{K} q_{n,u}^{(t)}(k)\,\log p_{n,v}^{(s)}(k). \tag{3}$$

$$\mathrm{p}_{n,v}^{(s)}(k) = \text{softmax}\Big(g_s(f_s(x_{n,v}))T_s\Big)_k,$$
$$q_{n,u}^{(t)}(k) = \text{softmax}\Big(g_t(f_t(x_{n,u})) - cT_t\Big)_k,$$

where:

N Number of training samples in a batch.

$x_{n,v}$ The v-th augmented view of sample n.

f_s, f_t Student and teacher encoders.

g_s, g_t Student and teacher projection heads.

T_s, T_t Temperature parameters for the student and teacher softmax distributions.

c Centering vector applied to teacher logits to stabilize training and avoid collapse.

$p_{n,v}^{(s)}(k)$ Probability assigned to class k by the student for view v of sample n.

$q_{n,u}^{(t)}(k)$ Probability assigned to class k by the teacher for view u of sample n.

V Student views (local and global crops).

U Teacher views (global crops).

K Dimension of the output space.

4 Results

4.1 Transfer Learning

Method: The first method used to evaluate the performance of the trained SSL models on the WV dataset (50 epochs) each was transfer learning via a linear probe. In transfer learning, a pretrained model is applied to a new dataset for a related but distinct task (Pan Yang, 2010). Here the quality of the learned features was evaluated by applying the SSL models to a manually annotated dataset of 1,000 Sentinel-1 SAR WV images (preprocessed the same way as the SSL training imagery) separated into 10 classes based on its predominant geophysical processes. Examples of these 10 classes are shown in Figure 1. Frozen embeddings from the SSL models were extracted and trained using a multinomial logistic regression using an 80/20 training/test split, for 5,000 iterations. The dataset was kept small for the test to ensure the performance focused on the learned features from the SSL models.

Results: Table 3 shows the overall F1 scores for all models plus a non-pretrained baseline and a pretrained resnet18 model. For the SSL models trained in this research, SimCLR with intermediate augmentation scores the highest: accuracy 0.885, macro-F1 0.86, weighted-F1 0.88, outperforming SimCLR with heavy and light augmentations and both DINO models. Relative to the model without SSL pretraining (accuracy 0.33, macro-F1 0.28), it outperforms this by +0.555 accuracy and +0.58 macro-F1, showing that self-supervised pretraining learns features with substantial predictive value. For reference, an ImageNet-pretrained ResNet18 encoder (trained on 1.28M images across 1,000 classes) achieved an accuracy of 0.91, a +0.025 improvement over the SimCLR intermediate model.

Table 2: Linear-probe results on the SAR dataset

Model	**Acc**	**Macro F1**	**Weighted F1**
SimCLR Heavy	0.855	0.820	0.850
SimCLR Intermediate	**0.885**	**0.860**	**0.880**
SimCLR Light	0.810	0.790	0.810
Baseline	0.330	0.280	0.310
DINO Heavy	0.640	0.600	0.640
DINO Light	0.740	0.720	0.740
Image-net Resnet18 Encoder	0.910	0.874	0.913

Best model analysis: The performance of SimCLR with intermediate augmentations

is broadly strong across all classes, four classes have F1 $\geq$ 0.90 and nine classes are F1 $\geq$ 0.80. The strongest classes are Pure Ocean Waves (P 0.963, R 1.000, F10.981; n=26), Ice berg (P 1.000, R 0.909, F1 0.952; n=11), Atmospheric front (P 0.917, R 0.917, F1 0.917; n=24), and Rain cells (P 0.909, R 0.909, F1 0.909; n=22). The weakest scoring classes are Low Wind Area (P 0.667, R 0.923, F1 0.774; n=13), followed by Oceanic Fronts (P 0.800, R 0.800, F1 0.800; n=5) and Biological Slicks (P 0.870, R 0.769, F1 0.816; n=26). The size of the datasets for class varies between 5–26 (median = 24), with the smallest test size for Oceanic Fronts of just 5 samples.

Label efficiency: Using the same 80/20 train/test split, we repeatedly subsampled the training labels to $\{1, 5, 10, 25, 50, 100\}$% and trained the same multinomial logistic-regression probe on the frozen SimCLR embeddings. Macro-F1 increases steeply on small increases of labels and then plateaus: $\sim$1% $\rightarrow$ 0.372 $\pm$ 0.031, 5% $\rightarrow$ 0.594 $\pm$ 0.031, 10% $\rightarrow$ 0.675 $\pm$ 0.027, 25% $\rightarrow$ 0.741 $\pm$ 0.027, 50% $\rightarrow$ 0.828 $\pm$ 0.015, and 100% $\rightarrow$ 0.878 $\pm$ 0.000. In relative terms, these correspond to approximately 42%, 68%, 77%, 84%, and 94% of the full-data macro-F1, respectively. The standard deviation also shrinks as more labels are provided (0.031 at 1–5%, 0.027 at 10–25%, 0.015 at 50%, and 0.000 at 100%), indicating increasing stability across random subsamples. Overall, the pretrained representation is label-efficient as most of the eventual performance is achieved with 10–25% of the labels, after which returns diminish and scores approach a ceiling by 50–100%, indicating diminishing returns from additional annotation.

4.2 Clustering

Method: After identifying the best performing model via transfer learning, we analysed its learned representation space through unsupervised clustering. The goal was to test whether semantically similar WV SAR images are pulled together in the embedding space learned by SimCLR. For each image, we extracted the encoder feature vector and applied Principal Component Analysis (PCA) to reduce dimensionality to 50 components, preserving a cumulative explained variance of 0.9311. K-means clustering was then applied in the PCA-transformed space (k=11; init=*k-means++*; n_{init}=20; max_iter=300; seed=42), where k was selected via the elbow method on the inertia curve, selecting the largest value for k beyond which additional clusters yielded only marginal decreases in inertia.

Findings: Clusters showed heavy overlap, the global silhouette score was 0.005, indicating no meaningful separation into distinct clusters. Cluster sizes varied (min

1,797; max 8,269; median 3,471; mean 4,235), but the near-zero silhouette indicates poorly defined boundaries rather than simple imbalance. To visualise these clusters, they were projected into a Two-dimensional space using the t-Distributed Stochastic Neighbour Embedding (t-SNE) projections. In Figure 4, visually it can be seen that there are mixed groups of clusters without clear gaps. Indicating an inability within the model to correctly cluster images into distinct groups based on underlying geophysical features.

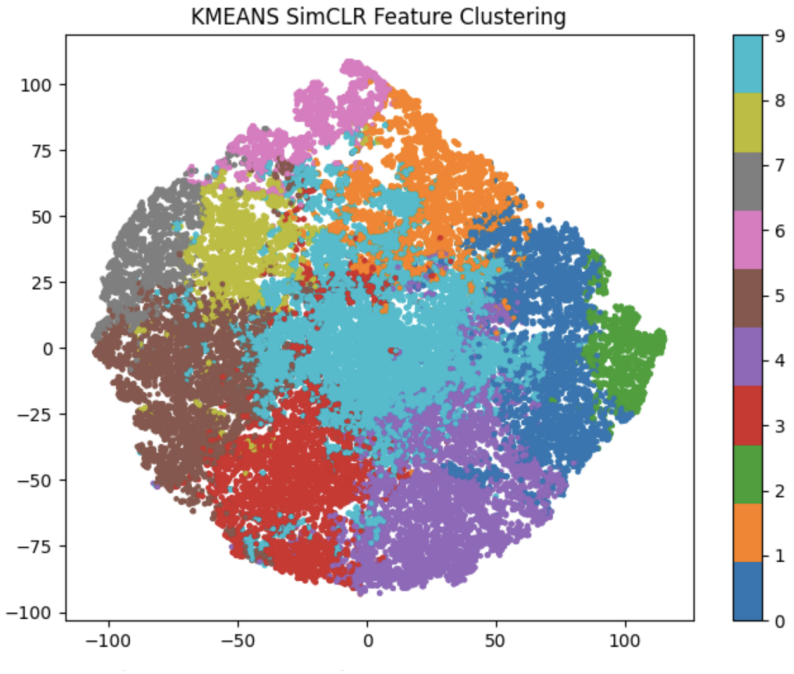

Figure 5: t-SNE of 50D PCA features colored by k-means cluster. Clusters overlap substantially, consistent with a silhouette of -0.005.

Although generally the model could not produce well defined clusters, after analysis of images within each k-means cluster by the shortest euclidean distance to the cluster centroid. It can be seen that the cores of these clusters show clear local semantic similarity, images nearest to each centroid typically share the same dominant geophysical phenomena (e.g., similar wave fields or similar wind patterns). As the distance from the centroid increases, semantic similarity diminishes and mixed process scenes become more common, indicating that the clusters have compact, semantically consistent cores but fuzzy boundaries.

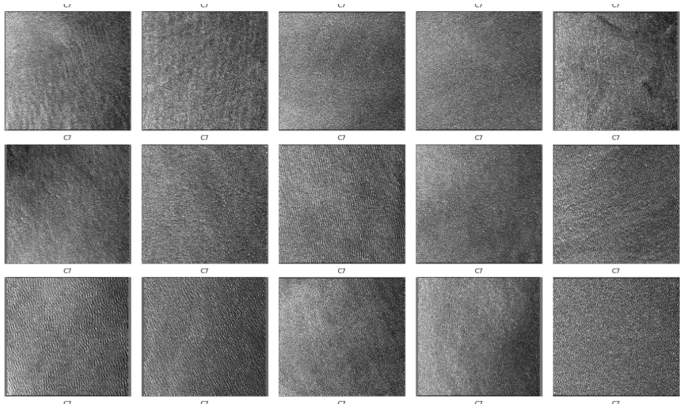

Figure 6: Top-15 nearest images to a cluster centroid in PCA space, illustrating Pure Ocean Waves.

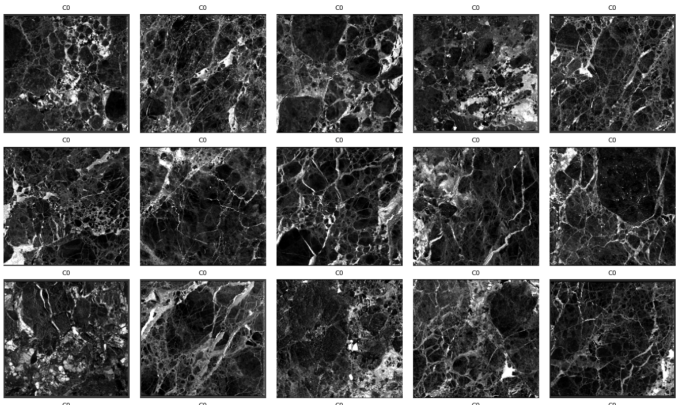

Figure 7: Top-15 Nearest Images to a Cluster Centroid in PCA Space, Illustrating Sea Ice

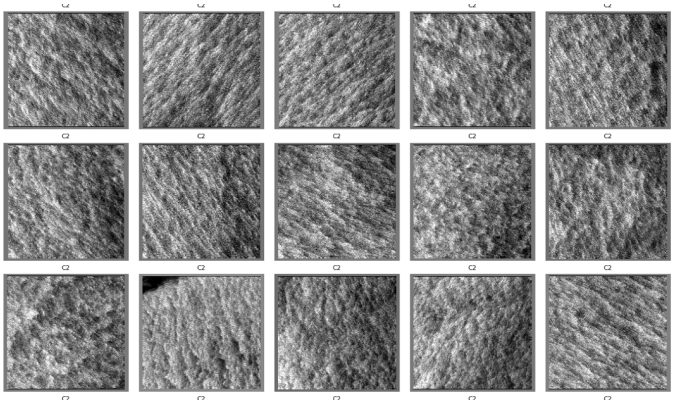

Figure 8: Top-15 Nearest Images to a Cluster Centroid in PCA Space, Illustrating Wind Streaks.

5 Discussion

There was clear difficulty for the model to produce distinct clusters, this outcome is consistent with the nature of the imagery and labels. Many SAR WV scenes contain multiple geophysical processes simultaneously (e.g., waves plus a front or rain cells), so the underlying variation is continuous rather than assigned into mutually exclusive categories. How this is handled during manual annotation varies. For example, for the dataset used in the logistical regression analyses of this research, each image is assigned a single predominant label despite the presence of mixed phenomena (Wang et al., 2019), a strategy that introduces clear bias. More recent multi-label approaches recognise this issue by allowing multiple simultaneous class annotations per image (Wang et al., 2025), reflecting expert disagreement and labeling ambiguity. As a result, any individual cluster will struggle to be unique. Nonetheless, images nearest to the cluster being very similar and generally from the same class indicate that the model is learning meaningful geophysical features, this is further backed up by the models ability to perform well in the logistic regression analysis with minimal labelling.

These findings position the model as a tool that can substantially reduce, rather than replace manual annotation in SAR WV analysis. This is backed up by the label-efficiency trend showing steep early gains and diminishing returns, a small fraction of labeled data is enough for a simple linear probe on frozen features to capture most of the eventual classification performance, with additional labels yielding progressively smaller improvements. Due to many scenes containing multiple processes, assigning a single geophysical label is uncertain and requires a more concrete set of rules to classify each image and therefore fully automatic labeling isn't appropriate for all cases. Therefore, the model serves as a starting point, reducing annotation as opposed to replacing it. Moreover, the model in this study was trained on just over 50,000 images, whereas millions of Sentinel-1 WV vignettes are available. Training on a substantially larger dataset would be expected to further improve model performance

In terms of individual geophysical features in which the model found most difficult to classify, predominantly Oceanic Fronts (OF) and Low Wind Areas (LWA), this difficulty can be largely attributed to a smaller sample size of data available for these processes, reflected in the transfer learning task. These classes had the lowest representation within the labelled dataset (8.5%), and are less frequently captured in WV imagery compared to other geophysical processes (Wang et al., 2019). Also, OF and LWA tend to co-occur with other geophysical features within a WV image,

making it more difficult for single label classification. Finally, the distinctiveness of low-wind regions may have been reduced by the local normalisation step applied during preprocessing, since this procedure removes differences in absolute backscatter intensity that are otherwise a defining characteristic of LWA.

For SimCLR, the simplest augmentation pipeline was the lowest performing. The augmentation did not mask the features sufficiently, allowing the model to identify positives within the batch without learning high-level features. This produced a low training loss but weaker transfer performance. Conversely, overly strong augmentations degraded feature learning by removing too much structure relevant to geophysical phenomena. Overall, all SimCLR variants were able to learn useful features, but transfer performance reflected this balance between too easy and too strong augmentations. As augmentations make the contrastive task prediction harder, representation quality can improve up to a point. The key is to balance the masking of insignificant features with the preservation of geophysically meaningful points. For DINO, lighter performed better than the heavier augmentations in our architecture, suggesting that even lighter augmentations may be worth exploring. Importantly, training loss is not a reliable metric for representation quality in SSL, low loss can reflect inadequate learning. As a result, early stopping based on a few epochs or small batches is less informative, which complicates hyperparametre tuning and training decisions in SSL. Reliable assessment requires downstream probes and visual comparisons of augmentations strategies prior to training.

It is difficult to identify a single reason why SimCLR outperformed DINO, given the limited analysis time and the many design choices made for this task (architecture, augmentation policy, batch size, temperatures, and optimiser settings). A possible explanation is that SimCLR's simpler objective and fewer views preserved more task-relevant structure in WV SAR images, especially with a batch size large enough to provide useful negative diversity. In contrast, DINO, which relies on multi-crop training, was likely affected by the design of the local crops. In our implementation, these crops were either too small or too aggressively augmented, masking or discarding key geophysical points, this is consistent with DINO's weaker transfer on classes dominated by localised structure (e.g., rain cells, oceanic fronts) as opposed to globally dominated phenomena (e.g., micro-convective cells, Pure Ocean Waves). More generally, the results shows that augmentation design has a significant effect on the performance of these models. Both SimCLR and DINO outperformed the baseline logistic regression, and the best-performing SimCLR model achieved performance close to that of the ImageNet-pretrained ResNet18, despite being trained on less

than 5% of the data and using a considerably smaller encoder.

6 Limitations

This study has several limitations that should be considered when interpreting the results. More augmentation variants could be studied for these methods, and additional SSL approaches could be incorporated in future work, to give a full analysis of the best SSL model to employ for SAR WV mode classification. On the data side, class imbalance and rarity (e.g., oceanic fronts) increase variance and reduce the F1 score. Preprocessing choices may also have affected results, the 100 m resolution can remove small-scale details and the incidence-angle effects from SAR imaging geometry were not explicitly modelled, incorporating a standardised correction on the angle could enhance the data being trained. Also although local normalistation enhances geophysical feature learning the absolute backscatter loss due it represents a loss of valuable physical information, this loss may reduce the usefulness of the learned representations for downstream tasks. Therefore, developing preprocessing strategies that minimise the loss of absolute backscatter information while still minimising variability, should be a priority for future work. Despite these constraints, the results indicate that SSL is a viable approach to incorporate in WV SAR classification and merits scaling in future work.

7 Conclusion

This research shows that contrastive SSL can serve as a useful tool for advancing our understanding of submesoscale geophysical processes in the open ocean. The best performing model learned semantically meaningful representations of these phenomena, as shown by strong performance in transfer learning tasks and meaningful clustering of image features. However, ambiguity of many ocean–atmosphere processes means that a fully automated classification system is not yet feasible. Instead, SSL provides a powerful way to reduce the amount of manual annotation required compared to purely supervised approaches. With the large amount of continuous data produced from sentinel-1 and the enhanced monitoring of finescale processes to be introduced by ESA's upcoming Harmony mission, SSL-based approaches have strong potential to enable large-scale, cost-effective analysis. Ultimately, such methods could support more detailed environmental modelling and improve our ability to monitor and understand the changing climate system.

8 Bibliography

Alpers, W., Zhang, B., Mouche, A., Zeng, K., & Chan, P. W. (2016). Rain footprints on C-band synthetic aperture radar images of the ocean—revisited. *Remote Sensing of Environment, 187,* 169–185.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 9650–9660.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. Proceedings of the 37th International Conference on Machine Learning, 119, 1597–1607.

Collard, F., Ardhuin, F., & Chapron, B. (2009). Monitoring and analysis of ocean swell fields from space: New methods for routine observations. Journal of Geophysical Research: Oceans, 114(C7).

Dai, Z., Lin, W., & Li, X. (2022). Sentinel-1 wave mode data for ocean monitoring: Characteristics and applications. Remote Sensing, 14(3), 589.

Gui, J., Sun, Z., Wen, Y., Tao, D., & Ye, J. (2024). A survey on self-supervised learning: Algorithms, applications, and trends. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(1), 122–145.

Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D., & Makedon, F. (2021). A survey on contrastive self-supervised learning. Journal of Intelligent Information Systems, 57(3), 423–454.

Jia, Y., Xu, Q., Wang, H., & Zheng, G. (2019). Observations of upper ocean processes with Sentinel-1 SAR wave mode data. Remote Sensing of Environment, 232, 111295.

Ji, X., Henriques, J. F., & Vedaldi, A. (2023). Invariant information clustering for unsupervised image classification and segmentation. International Journal of Computer Vision, 131(1), 85–102.

Liu, X., Zhang, F., Hou, Z., Wang, Z., Mian, L., & Zhang, J. (2023). Self-supervised learning: Generative or contrastive. IEEE Transactions on Knowledge and Data Engineering, 35(1), 857–876.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering, 22*(10), 1345–1359.

Nuijens, L., Stevens, B., Medeiros, B., & Zinner, T. (2024). Submesoscale processes in the atmosphere–ocean system: Observations and implications. Bulletin of the American Meteorological Society, 105(2), 241–260.

Tao, C., Qi, J., Guo, M., Zhu, Q., Li, H. (2023). Self-supervised remote sensing feature learning: Learning paradigms, challenges, and future works. IEEE Transactions

on Geoscience and Remote Sensing. Advance online publication.
Topouzelis, K., & Kitsiou, D. (2015). Detection and classification of mesoscale atmospheric phenomena above sea in SAR imagery. Remote Sensing of Environment, 160, 263–272.
Wang, C., Mouche, A., Tandeo, P., Stopa, J., Longépé, N., Erhard, G., Foster, R. C., Vandemark, D., Chapron, B. (2019). A labelled ocean SAR imagery dataset of ten geophysical phenomena from Sentinel-1 wave mode. Geoscience Data Journal, 6(2), 105–115.
Wang, C., Stopa, J. E., Vandemark, D., Foster, R., Ayet, A., Mouche, A., Chapron, B., Sadowski, P. (2025). A multi-tagged SAR ocean image dataset identifying atmospheric boundary layer structure in winter tradewind conditions. *Geoscience Data Journal, 12*(1), 1–14.
Yang, J., Gong, P., Fu, R., Zhang, M., Chen, J., Liang, S., & Dickinson, R. (2013). The role of satellite remote sensing in climate change studies. Nature Climate Change, 3(10), 875–883.
Yee Kit Chan, K. Voon Wong, & S. Chen. (2008). Synthetic aperture radar (SAR) for ocean remote sensing. *Progress in Electromagnetics Research, 82,* 49–68.
Zhang, H., Cao, Y., Xie, Z., Li, Z. (2022). A survey on masked autoencoder for self-supervised learning in vision and beyond. International Joint Conference on Artificial Intelligence (IJCAI), 4806–4813.
Zhu, X. X., Tuia, D., Mou, L., Xia, G. S., Zhang, L., Xu, F., & Fraundorfer, F. (2017). Deep learning in remote sensing: A comprehensive review and list of resources. IEEE Geoscience and Remote Sensing Magazine, 5(4), 8–36.
European Space Agency (ESA). (2013). Sentinel-1 user handbook. ESA Standard Document.
Wang, T., Liu, J., & Chen, W. (2020). Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In Proceedings of the 37th International Conference on Machine Learning (PMLR, Vol. 119) (pp. 9929–9939).
Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 22(10), 1345–1359.
Chen, X., Zhang, S., Xie, Y., & Liu, Y. (2025). Multi-label classification of remote sensing imagery: A survey and benchmark. IEEE Transactions on Geoscience and Remote Sensing, 63(1), 1–19.