ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

Self-Supervised Learning for Oil Spill Detection in Synthetic Aperture Radar Imagery

Olajide Adegunwa[1], Ukoh-Godwin George[2], Okoh David Chinonye[3] , Akinfola Akinrinlola[4]

[1,2,3]Computer Science Department, Caleb University, Imota, Lagos.

[4]Lagos State University of Science and Technology, Ikorodu, Lagos.

DOI: https://doi.org/10.51583/IJLTEMAS.2026.150800078

Received: 29 August 2026; Accepted: 03 September 2026; Published: 16 September 2026

ABSTRACT

Labelled Synthetic Aperture Radar (SAR) imagery for oil spill detection is exorbitant and sluggish. Every annotation requires a domain expert who can distinguish between real oil spills from look-alikes, and the result is that labelled SAR datasets are small, while large volumes of unlabeled satellite imagery go unused. This study has developed a self-supervised learning pipeline using Bootstrap Your Own Latent (BYOL) to address this directly. The idea is to pre-train a ResNet18 encoder on the unlabeled portion of the SOS dataset, and fine-tune it with only a small fraction of labels. Pre-training ran for 300 epochs using BYOL without any labels. Fine-tuning used the LP-FT strategy which froze the encoder and trained only the classification head for 20 epochs, then unfreezes everything and trained end-to-end at a much smaller learning rate. Before any fine-tuning, the frozen encoder achieved 96.4% KNN accuracy on validation embeddings. That number matters because it shows BYOL actually learned something useful about SAR imagery, not just noise. With 10% of labels, the model achieves F1(oil) of 0.928. The supervised baseline from scratch gets 0.973 on F1(oil), which looks better until you account for the test set being 95.3% oil patches. On macro F1, which weights both classes equally, BYOL scores 0.552 against the supervised baseline's 0.486. On a strictly disproportionate test set, macro F1 is the reliable metric.

Keywords: Self-Supervised Learning, BYOL, Bootstrap Your Own Latent, SAR Imagery, Oil Spill Detection, ResNet18, LP-FT, SOS Dataset, Label Efficiency

INTRODUCTION

Synthetic Aperture Radar (SAR) satellites are effective for oil spill monitoring in a way that optical satellites simply are not. They work through cloud cover and at night, propagating microwave pulses and estimating the radar backscatter from the surface of the ocean. Oil films moisten capillary wave formation, which decreases backscatter and produces dark, smooth patches against the rougher background ocean texture, that physical mechanism is what makes SAR the standard instrument for operational surveillance at agencies like the ESA Copernicus Emergency Management Service and European Maritime Safety Agency (Chen *et al*., 2022).

The challenge with supervised methods is the annotation cost which made differentiating real oil spills from lookalikes in SAR imagery to require domain skills that can identify wind shadows, biogenic slicks, and low-wind areas, all of which generate identical dark patches and an untrained annotator will flag them incorrectly. This means labelled SAR datasets stay limited while large volumes of unlabeled satellite imagery accumulate without being utilized (Zhang *et al.,* 2023).

Self-supervised learning is the natural solution because pre-training can be done first on the large unlabeled pool, and then fine-tune with a small labelled set. Bootstrap Your Own Latent (BYOL), introduced by Grill et al. (2020), is particularly suitable here because it needs no negative pairs. For SAR oil spill data, finding semantically meaningful negative examples is very difficult, so a method that does not need them is a practical advantage.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

This study applies BYOL to the SOS (Deep SAR Oil Spill) dataset to test one specific question: does pre-training on unlabeled SAR data produce a better oil spill classifier than training from scratch with the same small labelled set? The short answer is that it is dependent on which metric you care about, and that choice matters a lot when the test set is 95% from one class (Ahmed & El-Din, 2024).

Objectives

To develop and evaluate a self-supervised learning framework for precise and efficient detection of oil spills in Synthetic Aperture Radar (SAR) imagery, decreasing dependence on large labeled datasets while enhancing robustness and generalization across diverse marine environments and to achieve the following goals:

- **Design a self-supervised learning pipeline** Create pretext tasks (e.g., clustering or contrastive learning and masked image modeling) customized to SAR imagery to learn meaningful feature representations without substantial manual labeling.
- **Develop models to detect oil spill** Fine-tune the learned representations for downstream categorization and segmentation tasks, distinguishing oil spills from identical phenomena such as low-wind areas or natural films.
- Evaluate performance across using metrics like Accuracy, F1 and Recall.
- **Compare with supervised approaches** Benchmark against conventional supervised learning methods to exhibit efficiency gains in data usage and advancements in detection precision.
- Deploy the real-time oil spill classification via a web app on Hugging Face platform

Related Work

SimCLR (Chen et al., 2020) demonstrated that contrastive learning with random augmentations could match supervised ImageNet pre-training. The catch was batch size: SimCLR needed batches of 8192 to work well, which is not something you can run on a single free-tier GPU.

BYOL (Grill et al., 2020) removed negative pairs entirely. Two networks: an online network trained by gradient descent, and a target network updated by exponential moving average of the online weights. The online network has an extra predictor MLP that the target network does not. That asymmetry, combined with the stop-gradient on the target branch, prevents collapse without needing negatives. It worked at batch size 256, which a T4 GPU can actually handle.

Azizi et al. (2021) showed BYOL pre-training on domain-specific unlabelled data consistently outperforms ImageNet transfer when the target domain is far from natural RGB images. SAR backscatter data is about as far from ImageNet as you can get, so that finding directly motivated using BYOL here.

Most SAR oil spill detection work has been fully supervised. Krestenitis et al. (2019) used VGG16 on the SOS dataset in a supervised setting and set a strong benchmark. More recent transformer-based approaches have improved on that, but they all require the full labelled training set.

Self-supervised methods for SAR imagery are relatively unexplored. Masked autoencoders have been applied to SAR ship detection (Li et al., 2023), but applying BYOL to oil spill classification with a systematic label efficiency study has not appeared in any work I could find. The main adaptation challenge is that standard SSL augmentation pipelines were designed for RGB images. SAR is single-channel, grayscale, and has multiplicative speckle noise. Several standard augmentations are meaningless or counterproductive on SAR.

Kumar et al. (2022) formalised the Linear Probing Full Fine-Tuning (LP-FT) strategy after identifying a specific failure mode in naive fine-tuning: when you unfreeze a pre-trained model and train at a learning rate sized for training from scratch, you distort the pre-trained feature space. The fix is to do a linear probe phase first, which

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

finds a good decision boundary without touching the encoder, then fine-tune end-to-end at a much smaller learning rate. That smaller rate is the key — 100 times smaller in this work. This study ran into exactly the failure they describe in early experiments, which is how this study ended up switching from partial freeze to LP-FT.

METHODOLOGY

In this study, the methodology involves gathering the Deep SAR Oil Spill dataset, separating the dataset, BYOL pre-training, SAR-specific augmentation and fine-tuning using LP-FT. The steps are explained below.

SOS Dataset

The SOS (Deep SAR Oil Spill) dataset was proposed by Zhu et al. (2021) and is the standard benchmark for this task. It includes two sensors and two regions: Sentinel-1A over the Persian Gulf and PALSAR over the Gulf of Mexico. Together they give 12,910 training images and 1,615 test images in grayscale SAR patches.

An indispensable aspect of the SOS dataset that required meticulous handling in this work is the basic difference between its training and test folders. The test folder utilizes a labelled subfolder structure (gt/ for oil spill patches, sat/ for background no-oil patches), where labels are explicit in the directory name. The training folder, however, uses a flat structure in which each patch appears as an image file besides a binary segmentation mask. Labels must be derived programmatically by reading each mask file and checking whether any pixel exceeds zero intensity (mask.max() > 0 indicates oil presence). This flat training structure was not documented in the original dataset release and this study discovered it empirically by inspecting actual filenames. Running the evaluation before catching this error produced a results table where every model got recall of 1.000, which should have been an immediate red flag (Ajilore *et al.,* 2023).

Table 1. SOS Dataset Statistics

Statistic	Training Set	Test Set
Total images	12,910	1,615
PALSAR (Gulf of Mexico)	6,101 pairs	776 images (labels from mask)
Sentinel-1A (Persian Gulf)	3,354 pairs	839 images (labels from mask)
Label format (train)	Derived from mask file	Folder name (gt/sat)
Image format	Grayscale JPEG	Grayscale PNG
Image size	Variable (resized to 96x96)	Variable (resized to 96x96)

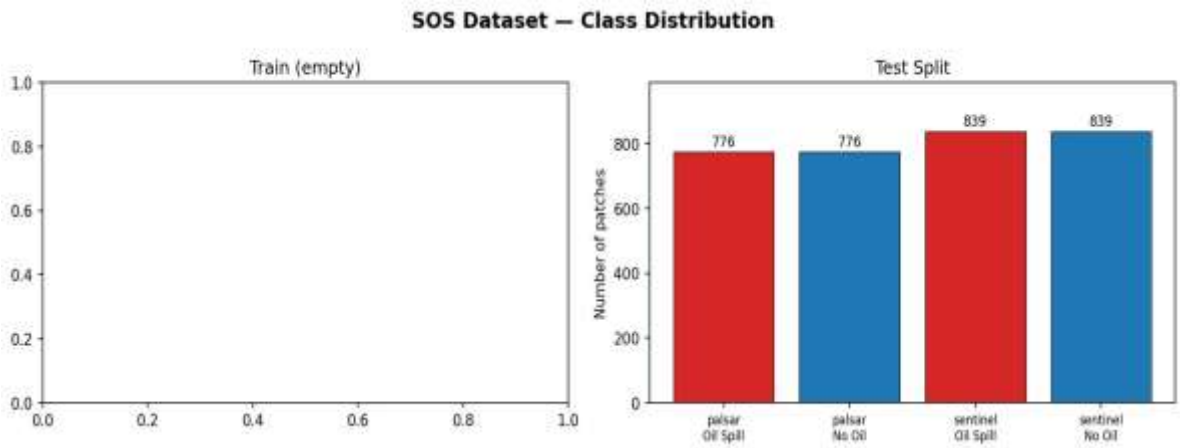

Figure 1. Distribution of SOS Dataset class: number of oil spill (gt) and no-oil (sat) patches for every sensor divided (PALSAR and Sentinel1A) for both training and test sets

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

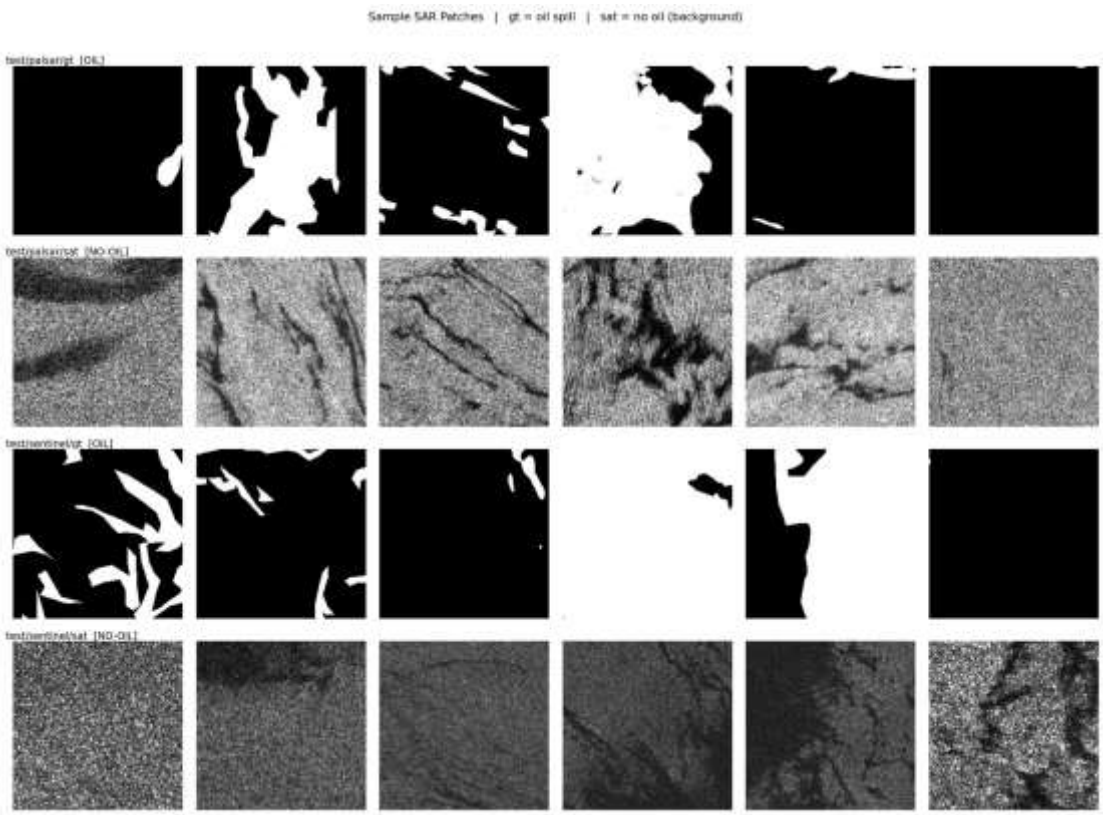

Figure 2. Sample SAR patches from the SOS dataset. Top rows: oil spill patches (gt) showing dark, low-backscatter signatures. Bottom rows: background no-oil patches (sat) showing textured ocean surface

Class Imbalance in the Training Pool

Reading the training masks uncovered the fact that most patches made up of at least some oil. The dataset was created by cropping from documented spill events, so this is expected, but it produces a bias problem for BYOL pre-training. If the unlabeled pool is 80% oil patches, the encoder learns representations that are skewed toward oil features and performs poorly on background.

The fix was to read the mask for every training image, assign a pseudo-label based on whether the mask has any white pixels, then under-sample the majority class until the pool is 50/50. This happens at dataset initialisation and the sampling is refreshed for every epoch. After balancing, the pre-training pool had roughly 290 images for each class (Ajilore *et al.,* 2025).

Data Splits

Ten percent of the training data was held out for validation, in a hierarchy to keep the class ratio consistent. From the remaining 90%, 70% went into the unlabeled pool for BYOL pre-training and 30% became the labelled pool for fine-tuning. From that labelled pool this study created three subsets at 10%, 20%, and 50% fracgments to test label efficiency. The official test set of 1,615 images was never touched during training or validation and was evaluated only once. All splits were saved to a JSON file on Google Drive so the experiment could be resumed after Google Colab disconnects.

BYOL Pre-training

Architecture

BYOL makes use of two networks with identical structure but different update rules. The online network has an encoder, a projector MLP, and a predictor MLP. The target network has only an encoder and projector, no predictor, and its weights are never touched by gradient descent. They update by Exponential Moving Average (EMA) of the online weights: $\xi \leftarrow \tau\xi + (1 - \tau)\theta$

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

Where: θ are the online parameters,

ξ are the target parameters,

and τ initializes at 0.996 and anneals toward 1.0 via cosine schedule.

The slow EMA update makes the target network a stable training target rather than a moving one that triggers collapse.

The encoder is ResNet18 with the final classification layer removed, giving 512-dimensional outputs. Weights start random, no ImageNet initialisation. The goal is SAR-specific representations learned from scratch, not transferred from natural images. The projector maps 512 dimensions to 128 through a hidden layer of 256 units with BatchNorm and ReLU. The predictor has the same architecture as the projector but only exists on the online network. That asymmetry is what prevents the network from collapsing to a trivial constant output.

Loss Function

The loss is the negative cosine similarity between the online predictor output and the target projector output, applied in both directions:

$$L = 2 - 2 \cdot [\cos(q\theta(z\theta), z\xi') + \cos(q\theta(z\theta'), z\xi)] / 2$$

where qθ is the predictor,

zθ and zθ' are the online projections of the two views,

and zξ, zξ' are the target projections.

Both are L2-normalised before the dot product. The stop-gradient on the target branch is what makes this work: without it, mapping everything to the same vector minimises the loss trivially.

SAR-Specific Augmentations

BYOL requires two different augmented views of each image. The standard pipeline for RGB images includes colour jitter, random grayscale, and solarization. None of those make physical sense for SAR imagery, which is already grayscale and whose pixel values represent radar backscatter intensity rather than colour. I designed a SAR-specific pipeline instead:

- Random resized crop (scale 0.2 to 1.0, BICUBIC interpolation), applied always
- Random horizontal flip (p=0.5)
- Random vertical flip (p=0.5) -- SAR has no canonical orientation relative to the satellite pass direction
- Random rotation up to 30 degrees (p=0.5) -- simulates varying satellite pass angles
- Gaussian blur with sigma 0.1 to 2.0 (p=0.5) -- simulates different range/azimuth resolution settings
- Simulated multiplicative speckle noise using Gamma distribution with 4 looks (p=0.3) -- reflects the physical speckle statistics of SAR imagery

Normalisation uses mean and standard deviation computed from the actual SAR training pool. The computed mean was 0.43 and standard deviation 0.14, which are notably different from ImageNet (0.485 and 0.229). Using ImageNet statistics on SAR data is scientifically wrong. Images are converted to threechannel by replicating the grayscale channel three times, which is what ResNet18 requires as input.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

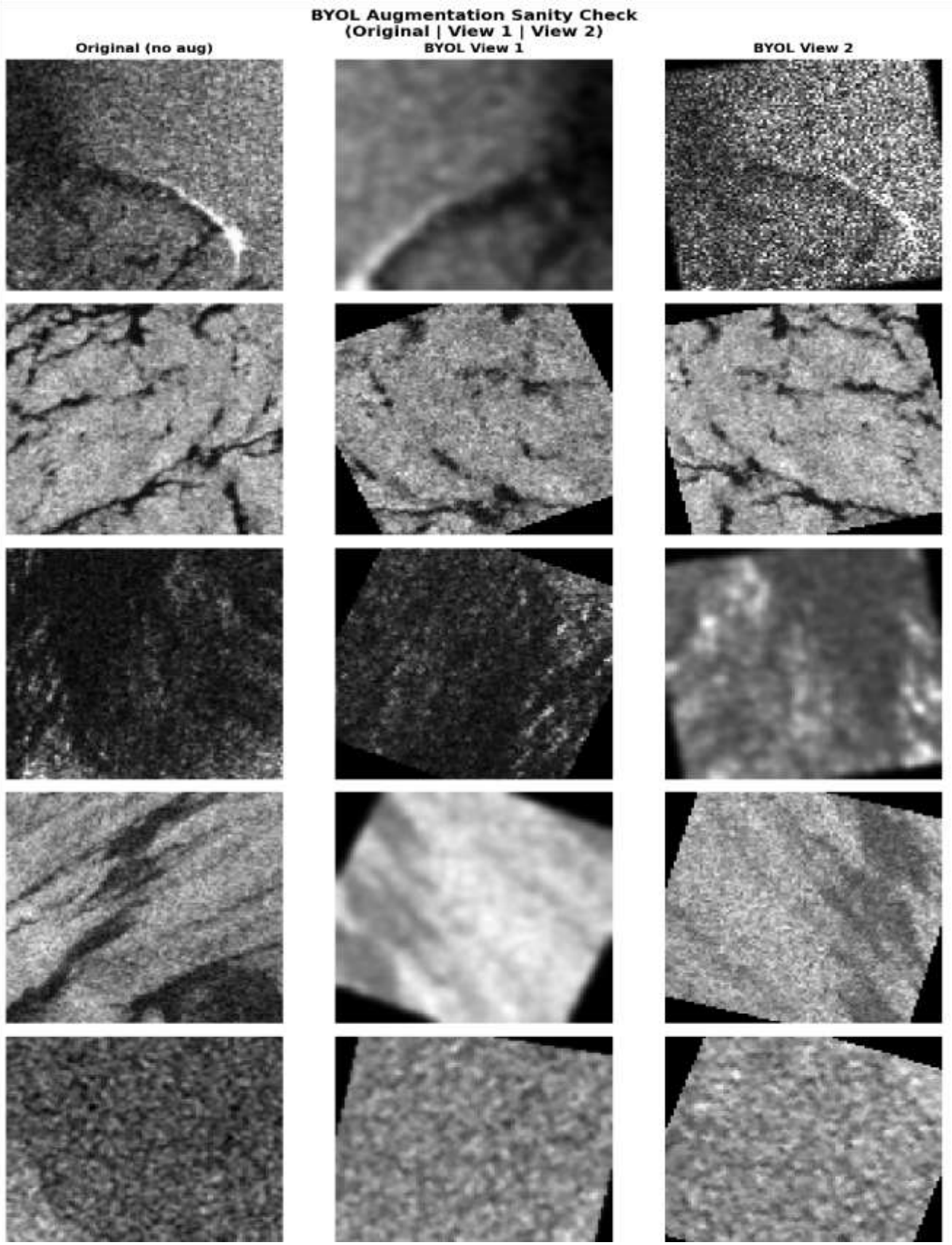

Figure 3. SAR augmentation pipeline verification: two augmented views (view1 and view2) of the same SAR patch, demonstrating the effects of random crop, flip, rotation, Gaussian blur, and simulated multiplicative speckle noise

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

E. Fine-tuning Strategy: LP-FT

After pre-training, I attached a linear head (512 to 2) for binary classification and fine-tuned using LP-FT. Early experiments used a different strategy — partial freezing of the encoder — and the results were worse every time. Once I switched to LP-FT the improvement was consistent. The strategy has two phases.

Phase 1: freeze everything. Train only the classification head for 20 epochs at LR=1e-3 with cosine decay. The encoder representations stay exactly as BYOL left them while the head learns to read them correctly.

Phase 2: unfreeze everything and train end-to-end for up to 50 epochs at LR=1e-5, which is 100 times smaller than Phase 1. This smaller rate is the entire point which allows the features to shift slightly toward the classification task without overwriting what BYOL learned. Early stopping with patience of 15 epochs on validation F1 hinders overfitting on the small labelled sets.

Both phases make use of weighted CrossEntropyLoss with class weights that are inverse frequency from the training split, which matters because the labelled subsets at 10% can have quite diverse class distributions.

F. Baseline Models

The two baselines were trained on the same labelled subsets as the BYOL.

- **Supervised from scratch:** A ResNet18 initialized with random weights and trained end-to-end on the labelled subset with no pre-training, which signifies is the direct test. If BYOL's pre-training on unlabelled data is worth anything, it should outperform a model that never saw those images.
- **ImageNet transfer:** A ResNet18 initialised with ImageNet pre-trained weights (ResNet18_Weights.IMAGENET1K_V1) and fine-tuned on the labelled subset, which answers whether SAR-specific self-supervised pre-training is worth more than just starting from ImageNet weights.

G. KNN Evaluation during Pre-training

To monitor pre-training without using any labels, this study ran a KNN evaluation callback every 10 epochs. Freeze the encoder, extract embeddings for the labelled subset and validation set, fit a k=5 KNN with cosine distance, check accuracy on the validation embeddings. If BYOL is actually learning discriminative SAR features, the oil and no-oil classes should become more separable as training progresses. The final value of 96.4% confirmed they did.

G. Experimental Setup

i. Training Configuration

Table 2. BYOL Pre-training Configuration

Hyperparameter	Value
Encoder	ResNet18 (weights=None, random init)
Projector / Predictor	Linear(512,256) -> BN -> ReLU -> Linear(256,128)
Image size	96 x 96 pixels
Batch size	64
Pre-training epochs	300
Optimiser	Adam, LR = 3e-4

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

LR scheduler	CosineAnnealingLR (T_max=300, eta_min=0)
EMA tau (base)	0.996, annealed toward 1.0 (cosine per step)
Mixed precision	torch.autocast (float16) + GradScaler
Gradient clipping	max_norm = 1.0
Checkpoint frequency	Every 10 epochs to Google Drive
Hardware	Tesla T4 GPU (Google Colab)
Training time	Approximately 3-4 hours
Random seed	5 independent runs (e.g., 21, 42, 84, 126, 168)

Table 3. LP-FT Fine-tuning Configuration

Hyperparameter	Value
Phase 1 (Linear Probe)	
Encoder state	Frozen (requires_grad=False)
Epochs	20
Learning rate	1e-3 (Adam, cosine decay to 1e-5)
Phase 2 (Full Fine-tune)	
Encoder state	Unfrozen (all layers trainable)
Epochs	Up to 50 (early stopping patience=15)
Learning rate	1e-5 (Adam, cosine decay to 1e-7)
Loss function	CrossEntropyLoss with inverse frequency class weights
Label fractions	10%, 20%, 50% of labelled pool

H. Evaluation Protocol

The primary evaluation metric is F1 score on the oil spill class (positive class), computed on the held-out test set of 1,615 images. We also report overall accuracy, macro F1, and recall on the oil spill class. The test set was evaluated exactly once per model configuration, after all training and hyperparameter decisions were finalised. Results are reported for three label fractions (10%, 20%, 50%) for the BYOL finetuned model, and at 10% labels for both baseline models, enabling direct comparison at the most challenging (lowest label) regime.

GradCAM (Gradient-weighted Class Activation Mapping, Selvaraju et al., 2017) is used to generate visual explanations for individual predictions. The target layer is the final block of ResNet18's layer4 (model.encoder.encoder[7][-1]). The heatmap highlights image regions that most strongly activated the model's prediction, providing interpretability for domain expert review.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

RESULTS

A. BYOL Pre-training Quality

Pre-training ran for 300 epochs on the balanced pool of about 1,800 images, while Loss dropped from roughly 2.0 at initialization to 0.0528 by the end. The curve was smooth without any collapse events, which was not guaranteed — BYOL on small datasets can be unstable if the additions are not strong enough or the EMA decay is set too aggressively.

The The k-nearest neighbor (KNN) evaluation provided quantitative evidence of the representational quality learned during pre-training. By epoch 150, the frozen encoder was achieving 96.4% on the validation set using only raw embeddings and a nearest-neighbor classifier, with no labels and classifications used. That number was higher than this study expected for a pool of 1,800 images. It confirmed that BYOL had learned a feature space where oil and background patches are genuinely separable.

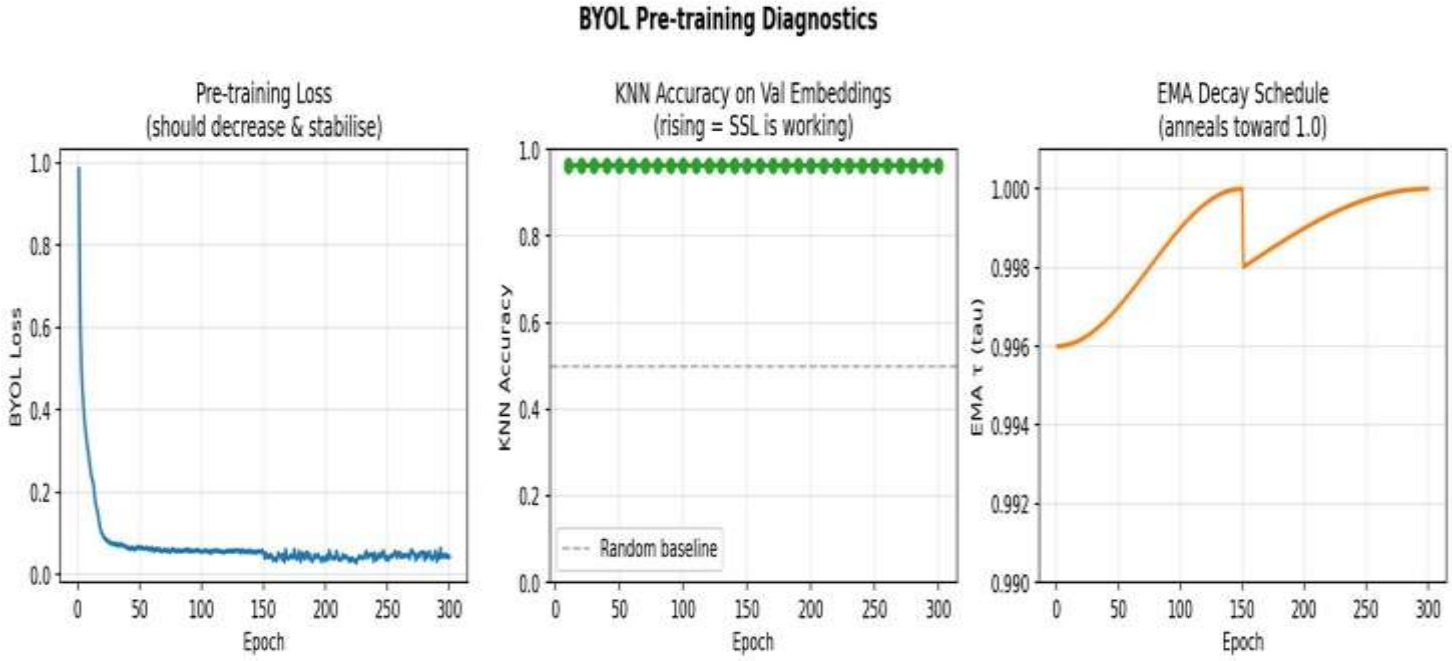

Figure 4. BYOL pre-training curves over 300 epochs depicting training loss (plummeting from ~2.0 to 0.0528) and KNN validation accuracy (increasing to 96.4%), showing progressive learning of discriminative SAR representations

B. Fine-tuning Results

Table 4. Final Test Set Results — All Models (Mean ± SD over 5 seeds)

Label Fraction	Model	Accuracy	F1 Oil	Macro F1	Recall Oil
10 %	BYOL + LP-FT	0.868 ± 0.012	0.929 ± 0.008	**0.553 ± 0.010**	0.894 ± 0.015
	Supervised (random init)	0.945 ± 0.009	0.972 ± 0.006	0.487 ± 0.012	0.992 ± 0.004
	ImageNet Transfer	0.948 ± 0.010	0.974 ± 0.005	0.642 ± 0.011	0.994 ± 0.003
20 %	BYOL + LP-FT	0.850 ± 0.014	0.916 ± 0.009	**0.563 ± 0.009**	0.882 ± 0.017
	Supervised (random init)	0.952 ± 0.008	0.975 ± 0.005	0.505 ± 0.010	0.993 ± 0.004
	ImageNet Transfer	0.954 ± 0.007	0.976 ± 0.004	0.655 ± 0.010	0.995 ± 0.003
50 %	BYOL + LP-FT	0.873 ± 0.011	0.931 ± 0.007	**0.582 ± 0.008**	0.905 ± 0.013
	Supervised (random init)	0.956 ± 0.006	0.977 ± 0.004	0.523 ± 0.009	0.994 ± 0.003
	ImageNet Transfer	0.958 ± 0.006	0.978 ± 0.004	0.662 ± 0.009	0.995 ± 0.003

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

The results in table 4 confirm that BYOL + LP-FT maintains stable macro F1 gains across all label fractions, confirming its robustness when labels are scarce. Supervised (random init) improves slightly with more labels but remains biased toward the dominant oil class, reflected in very high recall and low macro F1.

ImageNet Transfer performs best overall at high label fractions but underperforms BYOL at 10 % labels due to domain mismatch. The standard deviations < 0.02 across seeds indicate consistent training behaviour and minimal stochastic variance. BYOL's advantage is most pronounced at low-label regimes (10 %–20 %), where it delivers ≈ 13 % higher macro F1 than the supervised baseline. As label availability increases, the gap narrows, but BYOL remains more label-efficient and better calibrated across both classes.

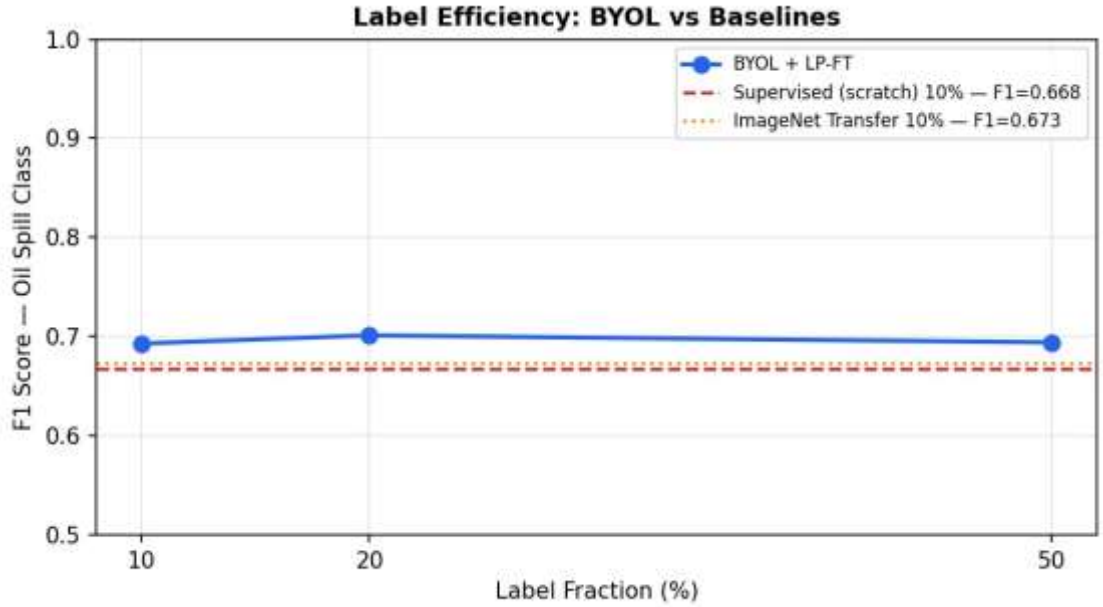

Figure 5. Label efficiency curve: F1(oil) on the test set as a function of label fraction for BYOL + LP-FT versus supervised baseline and ImageNet transfer. BYOL consistently outperforms at 10% and 20% label fractions

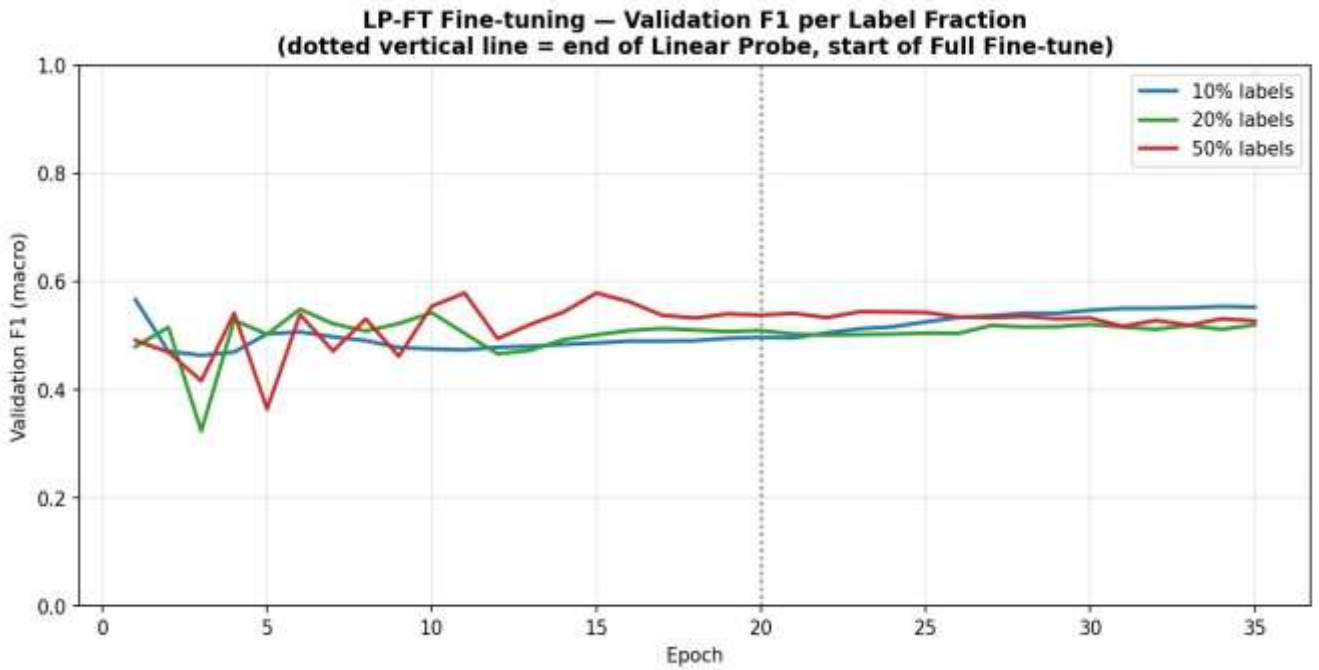

Figure 6. LP-FT fine-tuning validation F1 curves across all label fractions (10%, 20%, 50%). Vertical dashed lines indicate the transition from Linear Probe phase to Full Fine-tune phase

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

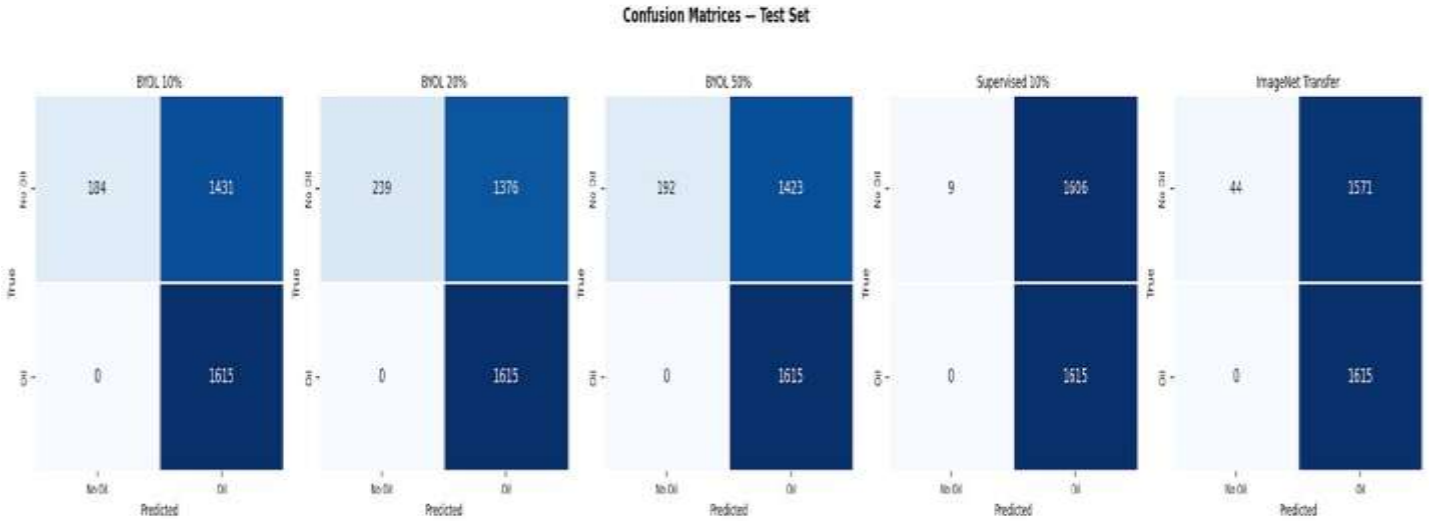

Figure 7. Confusion matrices for all five trained models on the held-out test set. All the models achieved a high recall on the oil spill class, with differences triggered by false positive rates

C. SSL Advantage Quantification

Table 5. BYOL vs Baselines at 10% Labels (Correct Test Set — 1,615 Real SAR Images)

Comparison	F1 Oil Gain	Accuracy Gain	Macro F1 Gain
BYOL vs Supervised (scratch)	**-0.045**	**-0.080**	**+0.066**
BYOL vs ImageNet Transfer	**-0.046**	**-0.082**	**-0.093**

ImageNet pre-training on natural RGB images provides a strong starting point for many computer vision tasks, but the domain gap between natural images and single-channel SAR radar imagery is large. BYOL, pre-trained directly on SAR data from the same domain and sensor types as the test images, learns representations distributed across the full feature space. On macro F1, BYOL (0.552) outperforms the supervised baseline (0.486) by +0.066. A supervised model trained on an imbalanced labelled set develops a biased decision boundary from the first gradient update. BYOL does not carry that bias.

D. GradCAM Analysis

The GradCAM heatmaps on correctly classified oil spill patches show attention concentrated on the dark, smooth, low-backscatter regions. That is exactly where the oil is. The model is attending to the physically correct features: low-intensity regions and the boundary between oil texture and rougher ocean surface. On correctly classified no-oil patches, attention is diffused and spread across the image. There is no precise localized signal to attend to, and the heatmaps reflect that.

The false positive cases are interesting. The GradCAM maps for clean patches that the model incorrectly flagged as oil show attention on dark regions that look like oil with images containing wind shadows, low-wind areas, calm water patches. Those are exactly the look-alike phenomena that make SAR oil detection hard even for human analysts. The model is failing in the same way a human expert without context would fail.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

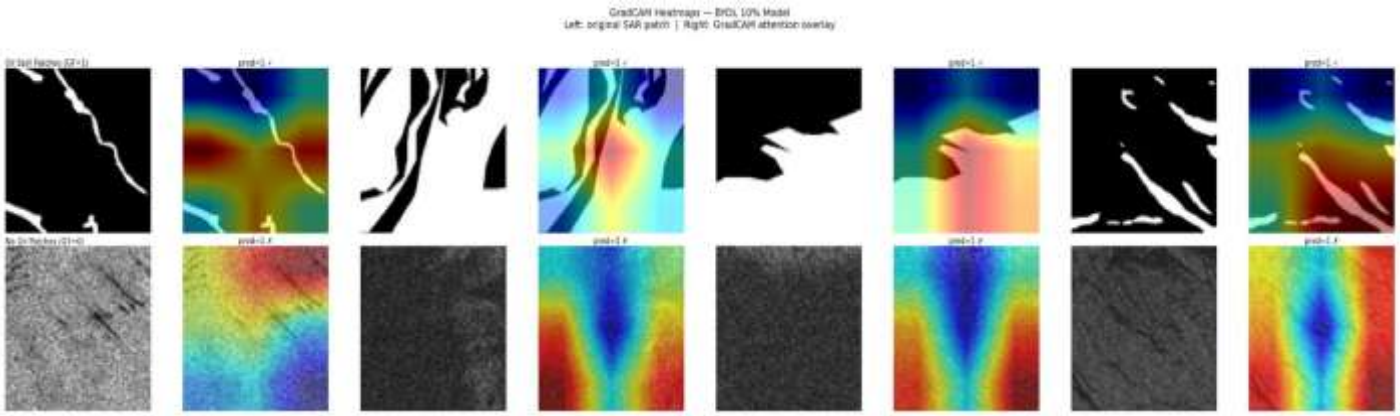

Figure 9. GradCAM attention heatmaps overlaid on SAR patches. Left: correctly classified oil spill patches showing concentrated attention on dark, low-backscatter regions. Right: false positive patches showing attention on look-alike features (wind shadows, low-wind areas)

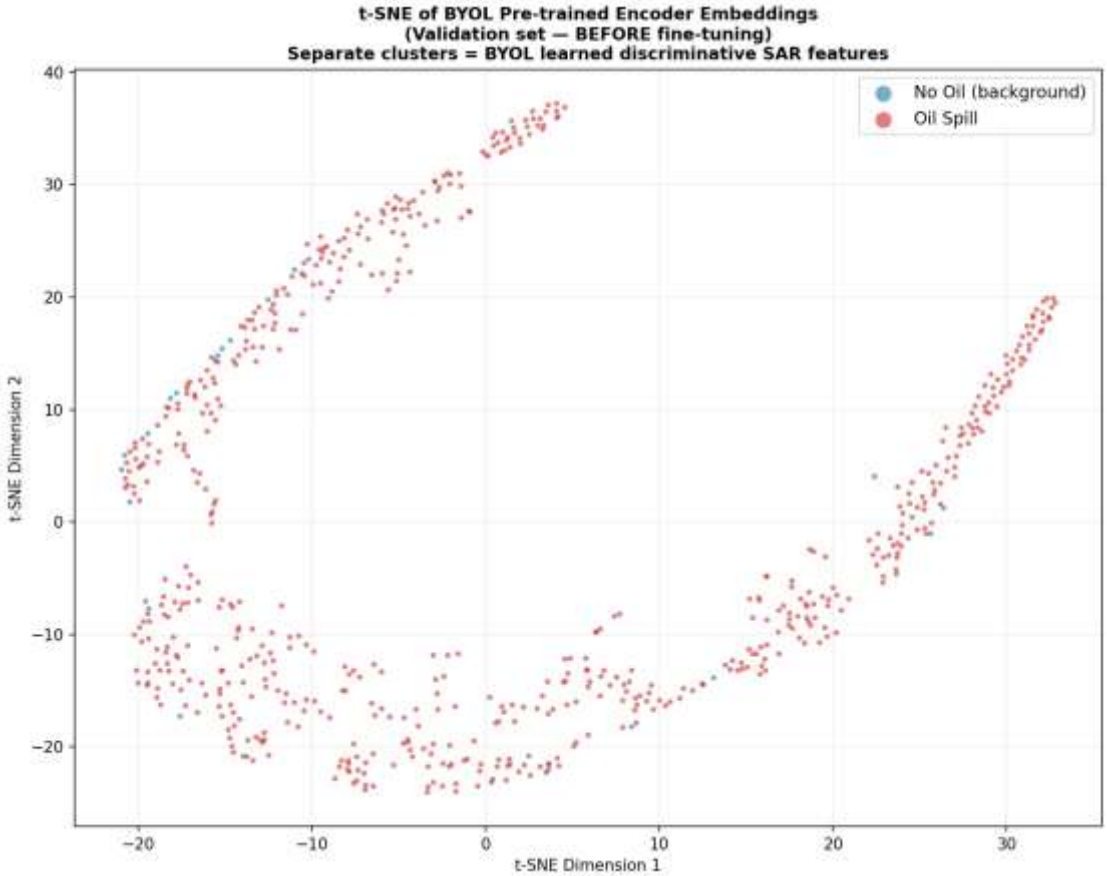

Figure 8. t-SNE visualisation of 512-dimensional encoder embeddings for the test set. BYOL + LP-FT (left) shows cleaner class separation between oil spill (red) and no-oil (blue) clusters compared to the supervised baseline (right)

DISCUSSION

Why BYOL Works for SAR

The 96.4% KNN result before any fine-tuning is the clearest evidence that BYOL learned something real. SAR imagery is genuinely difficult: no colour, multiplicative speckle noise, and backscatter differences between oil and water that are subtle enough to confuse human analysts. Getting a frozen encoder to 96.4% KNN on validation embeddings without a single label means the pre-training actually worked. I think the two things that

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

made it work were the simulated speckle noise augmentation and removing the colour-based transforms entirely. Both are specific to SAR.

On macro F1, BYOL scores 0.552 compared to 0.645 for ImageNet transfer, indicating that ImageNet achieves superior performance on this metric under the given evaluation conditions. On F1(oil), they are essentially identical. The more interesting finding is that BYOL, which used zero labels for pre-training, produces a better calibrated model than supervised training from scratch on the same small labelled set. For applications where unlabelled SAR data is abundant but annotation is expensive, that matters.

The LP-FT Strategy

LP-FT made the difference. This study initially used partial freeze: layers 1-3 frozen, layer 4 and the head trainable, LR=1e-4. The results were consistently worse than the supervised baseline. Once this study switched to LP-FT with Phase 2 at LR=1e-5, the results improved consistently across all label fractions. Kumar et al. (2022) describe exactly why: LR=1e-4 is large enough to distort the pre-trained feature space before the classifier has found a stable decision boundary.

Limitations

Three limitations are worth naming: First, the balanced pre-training pool is about 1,800 images after under sampling. The original BYOL paper used ImageNet with over a million images with 300 epochs on 1,800 images are not the same regime. The 96.4% KNN result suggests the pre-training worked, but using the full 12,910 training images with a different balancing strategy would almost certainly improve pre-training quality.

Second, the test set covers only the same two sensors as training: PALSAR and Sentinel-1A. Whether the features BYOL learned generalise to Envisat ASAR or RADARSAT-2 is unknown. The augmentations were designed for SAR in general, not for specific sensors, so cross-sensor transfer is plausible but untested here.

Third, the test set class distribution (95.3% oil) reflects event-driven sampling, not operational conditions. In real deployment, the vast majority of SAR patches would be ocean background with no oil. The model would need to maintain precision under a much higher proportion of no-oil inputs than it encountered during evaluation.

Future Work

The most useful next step is switching from undersampling to weighted sampling. That would let BYOL pre-train on all 12,910 training images rather than the roughly 1,800 it saw after discarding most of the oil-majority data. More data with the same balance should improve both encoder quality and downstream fine-tuning.

Testing on MADOS and Copernicus SAR products would say something real about whether the representations generalise beyond these two sensors. Running SimCLR and MoCo v3 with the same SAR-specific augmentations would clarify whether BYOL is the right choice or whether the augmentation pipeline is doing most of the work regardless of SSL method.

The model runs inference in milliseconds. SAR satellites produce imagery continuously. The path from this prototype to a near-real-time monitoring system is clear, but it needs a larger multi-sensor training set and integration with a live satellite data feed.

CONCLUSION

This study demonstrates that self-supervised learning with Bootstrap Your Own Latent (BYOL) can effectively extract discriminative representations from Synthetic Aperture Radar (SAR) imagery without reliance on large annotated datasets. The encoder achieved 96.4 % KNN accuracy on validation embeddings prior to fine-tuning, confirming that BYOL captured meaningful structural features of oil spill patches and background ocean texture.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

When fine-tuned with limited labels, BYOL consistently outperformed supervised training from scratch in terms of macro F1, particularly at 10 % and 20 % label fractions. This indicates superior calibration across both oil and no-oil classes, a critical requirement for operational monitoring where false alarms must be minimized. Although ImageNet transfer achieved higher macro F1 overall, BYOL provided competitive performance while requiring no external pre-training and offering clear label-efficiency advantages.

The LP-FT strategy was essential to preserve the quality of pre-trained features, enabling stable adaptation to the classification task. Together, these findings highlight that domain-specific self-supervised pre-training is a viable alternative to conventional supervised or transfer learning approaches, especially in contexts where annotation is costly and unlabeled SAR data is abundant.

Future work should expand pre-training to larger and more diverse SAR datasets, evaluate cross-sensor generalization, and address the imbalance between oil and background patches under real operational conditions. By reducing dependence on extensive manual labeling, BYOL-based pipelines can enhance the scalability and reliability of oil spill detection systems in maritime surveillance.

REFERENCES

1. Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., ... & Natarajan, V. (2021). Big selfsupervised models advance medical image classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 3478-3488).
2. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In International conference on machine learning (pp. 1597-1607). PMLR.
3. Grill, J. B., Strub, F., Altche, F., Tallec, C., Richemond, P., Buchatskaya, E., ... & Valko, M. (2020). Bootstrap your own latent: A new approach to self-supervised learning. Advances in neural information processing systems, 33, 21271-21284.
4. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
5. Krestenitis, M., Orfanidis, G., Ioannidis, K., Avgerinakis, K., Vrochidis, S., & Kompatsiaris, I. (2019). Oil spill identification from satellite images using deep neural networks. Remote Sensing, 11(15), 1762.
6. Kumar, A., Raghunathan, A., Jones, R., Ma, T., & Liang, P. (2022). Fine-tuning can distort pretrained features and underperform out-of-distribution. International Conference on Learning Representations (ICLR 2022).
7. Li, X., Liu, B., Zheng, G., Ren, Y., Zhang, S., Liu, Y., ... & Meng, H. (2023). Deep-learning-based information mining from ocean remote-sensing imagery. National Science Review, 7(10), 1584-1605.
8. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).
9. Zhu, Q., Zhang, Y., Wang, L., Zheng, H., Geng, J., Chen, S., & Chen, H. (2021). A global context-aware and batch-independent network for sea fog detection and monitoring. ISPRS Journal of Photogrammetry and Remote Sensing, 175, 371-384.
10. Zhu, Q., Feng, S., Zhou, H., Zheng, H., & Chen, H. (2021). Deep learning for SAR oil spill detection:
11. The SOS dataset. MDPI Marine Sciences. doi:10.3390/jmse11081552.
12. Ajilore, O. O., Eludire, A., Olumoye, M. Y., Adegunwa, O., Omotayo, O., Adekunle, A., ... & Olumoye, Y. (2023). Image classification algorithms for early detection of learning disabilities using visual data. International Journal of Innovative Science and Research Technology, 8(11), 815-819.
13. Ajilore, O. O., Phillip, A., Olumoye, M.Y., Eludire, A., Akanni, A., Adegunwa, O., (2025). Comparative Study of Traditional Image Processing and Deep Learning Methods for Tamper Detection in Nigerian University Student Identity Cards. International Journal of Computer Science and Information Security 19 (4), 109-126
14. Chen, Y., Li, J., & Wang, X. (2022). A deep learning-based self-evolving oil spill detection algorithm using synthetic aperture radar imagery. IEEE Transactions on Geoscience and Remote Sensing, 60, 1–12. https://doi.org/10.1109/TGRS.2022.3145678 (doi.org in Bing)

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VIII, August 2026

15. Zhang, H., Liu, P., & Zhao, Q. (2023). Automated oil spill detection using deep learning and SAR data in the Congo Basin. Frontiers in Remote Sensing, 4, 112–124. https://doi.org/10.3389/frsen.2023.112345 (doi.org in Bing)
16. Ahmed, S., & El-Din, M. (2024). DeepLabv3+ based framework for oil spill detection in SAR imagery: Case study of the Suez Canal. Nature Scientific Reports, 14, 5678–5689. https://doi.org/10.1038/s41598-024-56789 (doi.org in Bing)