Knowledge Without Learning: A Zero-Shot Approach to SAR ATR

Javier E. Garza, George Hellstern
Lockheed Martin Aeronautics Company
Fort Worth, Texas
javier.garza@lmco.com, george.hellstern@lmco.com

Matt Reisman, Kevin LaTourette, Tobé Corazzini, Adam Francisco, Ryan McCormick
Bedrock Research
Highlands Ranch, Colorado
matt@bedrockresearch.ai,
kevin@bedrockresearch.ai,
tobe@bedrockresearch.ai,
adam@bedrockresearch.ai,
ryan.mccormick@bedrockresearch.ai

ABSTRACT

Given the billions of dollars invested by the Department of Defense in artificial intelligence (AI) solutions for the warfighter, it is imperative to develop models for AI systems that allow agents to perform reliably and consistently. When a human switches context from, for example, electro-optical (EO) images to synthetic aperture radar (SAR)-generated images, they can look at the image and make sense of it without prior time invested in reviewing SAR imagery. This is a zero-shot approach to learning. Typical AI systems require additional learning stages for success at context-switching between EO, infrared and SAR images. Without additional representative data, accuracy and confidence levels decrease.

This paper examines how to train an intelligent machine learning (ML) system to evaluate a new situation and make sense of it using zero-shot learning approaches to automatic target recognition (ATR). The use of latent space with real-time transformer-based approaches in computer vision is examined for developing zero-shot algorithms for ATR. This latent space approach encodes image data onto a manifold, using clustering algorithms to identify objects with similar physical features. A mathematical introduction to these concepts is provided, in addition to a description of their application in determining the proximity of images to one another. By leveraging lower-dimensional space to represent the essential features of high-dimensional data, it is possible to map out objects in Euclidean space and determine their similarity to other images that a neural network already knows. Real-Time Detection Transformer and other such models have shown improved mean average precision over You Only Look Once model variants in literature. The combination of these approaches reduces the need for additional data in different modalities, while maintaining model performance.

ABOUT THE AUTHORS

Javier Garza is an LM Associate Fellow and autonomy, AI and ML engineering program manager in Lockheed Martin Aeronautics Company's Advanced Development Programs (informally, Skunk Works®) organization. Garza leads the Advanced Sensing AI portfolio. He has more than 14 years of combined experience with AI, software engineering, open architecture technologies, flight testing and technical leadership. He holds a bachelor's degree in computer science and a master's degree in software engineering from the University of Texas at El Paso and is currently a candidate for a Doctor of Engineering in AI and ML at the George Washington University.

George Hellstern has over 30 years of experience with systems design, including AI solutions for air-to-air combat and sustainment. He is a program manager for autonomy and AI, uncrewed air systems command and control, and human performance. His previous experience includes operational, programmatic and technical work in the Air Mobility Command, the Office of the Secretary of Defense and Skunk Works.

Matt Reisman, Ph.D., is the chief technology officer and co-founder of Bedrock Research. He received his Ph.D. in physics from Washington University in 2017, with a focus on developing noninvasive optical neuroimaging equipment and image-processing techniques. He previously spent seven years leading unclassified research and

development at Lockheed Martin Space Company, which included the development of the first operationally deployed deep learning computer-vision algorithm for the Intelligence Community.

Kevin LaTourette is the CEO and co-founder of Bedrock Research, a leading U.S. aerospace company specializing in the application of multimodal geospatial foundation models for remote sensing. LaTourette previously spent over 15 years as an LM Associate Fellow and chief architect at Lockheed Martin Space Company, where he led the research, development and operational deployment of a change detection and tracking program of record for the Intelligence Community.

Tobé Corazzini, Ph.D., is a principal engineer at Bedrock Research. She has over 20 years of experience in developing models and algorithms for remote sensing in areas spanning climate technology, aerospace and robotics. She previously served as the vice president of software engineering and data science at Aclima, Inc., and she holds a Ph.D. in aerospace engineering from Stanford University.

Adam Francisco is the lead ML engineer at Bedrock research. He has over 10 years of experience with computer science, ML and image processing, with prior roles at Lockheed Martin Space Company and, most recently, Arka.

Ryan McCormick is an ML intern at Bedrock Research. His work has spanned image processing, signal processing and web application development. He previously interned at Dupper Analytics and is a current undergraduate at Stanford University, majoring in computer science with a specialty in AI.

Knowledge Without Learning: A Zero-Shot Approach to SAR ATR

Javier E. Garza, George Hellstern
Lockheed Martin Aeronautics Company
Fort Worth, Texas
javier.garza@lmco.com, george.hellstern@lmco.com

Matt Reisman, Kevin LaTourette, Tobé Corazzini, Adam Francisco, Ryan McCormick
Bedrock Research
Highlands Ranch, Colorado
matt@bedrockresearch.ai,
kevin@bedrockresearch.ai,
tobe@bedrockresearch.ai,
adam@bedrockresearch.ai,
ryan.mccormick@bedrockresearch.ai

INTRODUCTION

The development of artificial intelligence (AI) and machine learning (ML) models for automatic target recognition (ATR) in remote sensing applications has become increasingly important for defense and security purposes. However, the process of training these models is often time-consuming and expensive, requiring large amounts of labeled data. Recent advances in AI and ML have led to the development of new approaches that can reduce the need for extensive labeled data, such as iterative active learning and zero-shot learning.

Iterative active learning is a technique that combines self-supervised pretraining with active learning to automate the data-labeling process. This approach has been shown to be effective in reducing the amount of human labor required for data labeling and can improve the accuracy of ATR models. By using synthetic data to seed the initial model training, iterative active learning can rapidly cover a comprehensive spectrum of diversities across illuminations, perspectives, geographies, target objects and scene complexities. This yields a foundation-model feature space that generalizes across broad collection situations. This approach has been shown to be effective in reducing the amount of human labor required for data-labeling, and in improving the accuracy of ATR models (Reisman et al., 2025). By leveraging the strengths of both self-supervised and active learning, iterative active learning can be used to develop models that can learn from large datasets with minimal human annotation.

Zero-shot learning is a technique that enables AI models to recognize and classify objects without prior training on those specific objects. This approach has been explored in the context of ATR, where it can be used to develop models that can switch between different modalities, such as electro-optical (EO) and synthetic aperture radar (SAR) images, without requiring additional training data. By leveraging latent space and real-time Transformer-based approaches, zero-shot learning can encode image data onto a manifold, using clustering algorithms to identify objects with similar physical features. Creating a zero-shot ML model using foundation models requires a deep understanding of the underlying concepts and techniques. Foundation models are pretrained on large datasets and can be fine-tuned for specific tasks. For example, the Masked Autoencoder (MAE), proposed by He et al. (2022), has been shown to be effective at learning robust and meaningful representations of images. The Cross-Scale MAE (Tang et al., 2023) builds upon this approach by explicitly learning relationships between data at different scales throughout the pretraining process. This allows for the development of models that can learn representations of images at multiple scales, making them more effective for such tasks as image classification and object detection.

Satellite imagery is a critical component of remote sensing, and the use of multispectral and temporal imagery has become increasingly important for such applications as land-cover classification and crop-yield prediction. Cong et al. (2022) proposed the SatMAE framework, a self-supervised learning approach for temporal or multispectral satellite imagery based on MAE. This approach has been shown to be effective in learning representations of satellite imagery that are beneficial for downstream tasks, such as land-cover classification and semantic segmentation.

From a multimodal perspective, Sastry et al. (2025) proposed the TaxaBind framework. They introduced the concept of a unified embedding space for ecological applications that covers six modalities: ground-level images, geographic location, satellite images, text, audio and environmental features. This approach has been shown to be effective at learning a joint representation space that can be used for various downstream ecological tasks, such as species-

distribution modeling and habitat classification. By leveraging the strengths of multiple modalities, the approach can be used to develop models that can learn robust and meaningful representations of ecological systems.

The development of zero-shot learning models for remote sensing applications is a rapidly evolving field, driven by advances in deep learning techniques, representation learning, foundation models, iterative active learning, and multispectral and temporal satellite imagery. The application of these advancements to SAR data has the potential to revolutionize the field of remote sensing. This paper begins by providing an overview of SAR imaging, after which it provides context for current approaches to ATR using SAR data. Then, it summarizes the relevant approaches in EO that are key enablers for a foundation model that works with SAR data. A description of the approaches and techniques that were implemented is provided, as well as a description of the dataset used for this research, followed by a model performance evaluation. Finally, the paper provides an overall assessment of the results and contributions of this research.

BACKGROUND

Introduction to Synthetic Aperture Radar

SAR imaging works by using a moving platform, such as an aerial vehicle or satellite, to transmit electromagnetic waves and collect the resultant backscatter signals using an antenna, then combining them to form a visual representation (Moreira et al., 2013). Regardless of lighting or weather conditions, SAR sensors can be used to collect data, which makes their use an attractive alternative to EO collection methods. Although this is an advantage, imagery produced by SAR systems differs from EO imagery and requires special processing to develop a visual representation. Figure 1 shows an example of a SAR image, and Figure 2 shows an example of an EO-based satellite image in the same location.

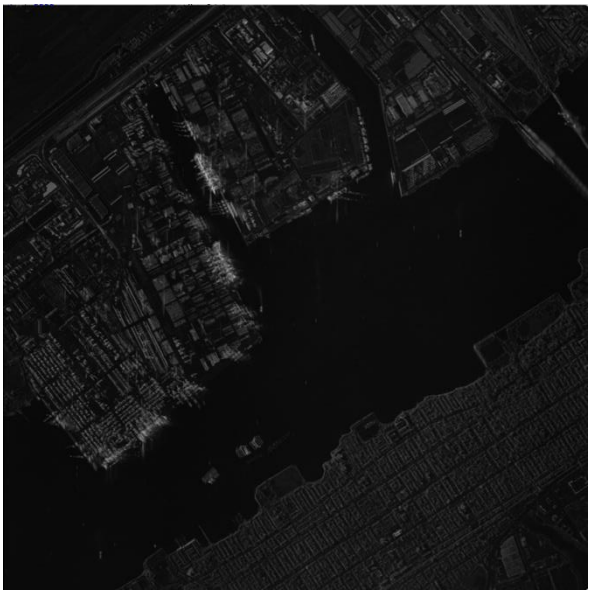

Figure 1. SAR Imagery Example (Data from Umbra SAR Open Data, accessed on 13 June 2025 from https://registry.opendata.aws/umbra-open-data. Licensed under CC BY 4.0.)

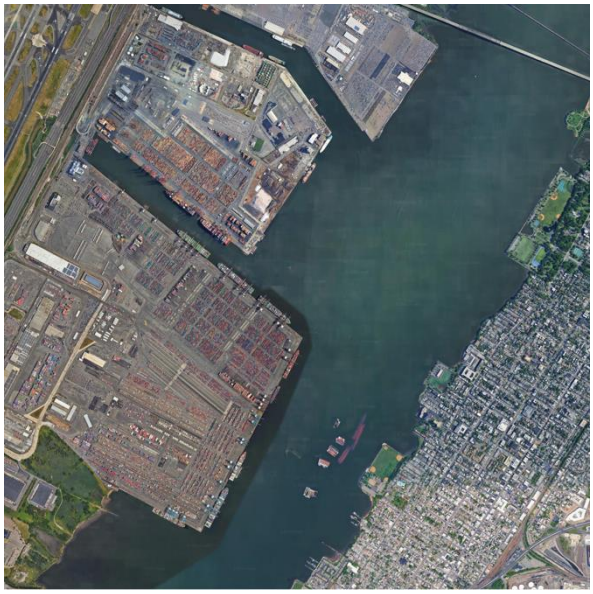

Figure 2. EO Satellite Imagery Example (Image © 2025 Google, Maxar Technologies, CNES/Airbus)

ATR Approaches to SAR

Li et al. (2023) surveyed recent developments in ATR using SAR, having focused on deep learning progress since 2017. They emphasized how deep learning, especially convolutional neural networks (CNN), had surpassed traditional methods by enabling automatic feature extraction and achieving much higher classification accuracy. The authors covered a range of CNN-based approaches designed to tackle common SAR challenges: limited data, class imbalance

and the difficulty of extracting reliable features due to the unique imaging properties of SAR and its sensitivity to azimuth angles. They discussed widely used datasets, such as those in Moving and Stationary Target Acquisition and Recognition (MSTAR) and OpenSARShip, along with standard evaluation metrics, such as accuracy, precision, recall and F1 scores. At the same time, they acknowledged the shortcomings of MSTAR, particularly its small size and consistent collection conditions, which often led to overfitting.

CNNs were clearly the dominant deep learning architecture in SAR ATR, outperforming earlier models, such as restricted Boltzmann machines and deep belief networks. To address data scarcity, the survey highlighted such methods as data augmentation (including such techniques as generative adversarial networks [GAN]) and electromagnetic simulation), transfer learning and few-shot learning. It also covered the use of more complex data types, such as polarimetric SAR and complex-valued SAR data, which offered richer information. Several CNN-based models achieved very high accuracy in MSTAR, often above 99%. For example, Xie et al. (2019) reached 99% with their umbrella network. Huang et al. (2019) reported 99.79% using a group squeeze-excitation CNN. Wang et al. (2020) achieved 99.55% with SSF-Net and 99% with DNet on the 10-class MSTAR dataset. GANs also proved helpful with data augmentation, with some models hitting 99.5% accuracy when trained on GAN-augmented data.

Attention mechanisms and capsule networks showed strong results too. One example is the capsule-based model from Ren et al. (2021), which reached 99.18% on the 10-class MSTAR dataset. The survey additionally reviewed efforts to enhance real-time recognition capabilities and to understand and mitigate adversarial attacks on SAR ATR systems. The authors wrapped up with suggestions for future work. These included building larger and more diverse SAR datasets to reduce overfitting, improving CNN architectures, incorporating domain knowledge, optimizing for real-time use, and making models more interpretable and robust against adversarial attacks. Even though the SAR problem space has been explored in the context of some open datasets, it is important to highlight and consider approaches that have been applied to the EO modality.

Class Incremental Object Detection

Class incremental object detection aims to prevent catastrophic forgetting while learning to recognize new types of objects. To address this issue in CNNs, Shmelkov et al. (2017) proposed two loss functions for use during incremental learning with Fast R-CNN. They replaced its backbone with ResNet-50 and evaluated their improvements using the PASCAL VOC and COCO datasets. The authors used a subset of classes from the VOC dataset to train the network, froze it, and then fine-tuned it on the last class. Then, they experimented with adding multiple classes from COCO in groups, one at a time. The authors found that there was a slower decrease in mean average precision (mAP) when adding groups of new classes, versus incrementally adding each new class. They noted that the approach had successfully maintained mAP values on previously known classes.

Iqbal et al. (2023) integrated a clustering method into Faster R-CNN to enable open-world object detection. They used a margin-based loss function that combines the hinge loss and an outlier regularization parameter to differentiate between known and unknown classes. The model is composed of a region proposal network (RPN) that provides object proposals, and a region of interest head that performs localization and classification. The RPN is used to assign labels to unknown classes, and Helmholtz energy functions are subsequently used to represent them in latent space. This method is reliant on very large datasets, requiring extensive compute capabilities.

Representation Learning

Contrastive learning obtains representations of data using positive and negative pairs, pushing similar data points closer in latent space and pulling apart data points that differ. The effective implementation of this method has been explored in several publications, such as Oord et al. (2019). The authors proposed using contrastive predictive coding for unsupervised feature extraction from data. They developed InfoNCE, a contrastive loss function that uses noise-contrastive estimation. By deriving latent representations of input observations using an autoregressive process, their approach encodes them for future use. The loss function uses positive and negative examples to converge and was found to be performant in multiple domains with different classifiers.

He et al. (2020) proposed Momentum Contrast, which matches encoded queries to dictionaries of encoded keys using a contrastive loss function during visual-representation encoder training. This approach iteratively builds a large dictionary with such representations by using a queue. The queue is effectively used as a buffer, where new encoded

representations from the current mini batch are added while older ones are removed and used for encoder progression. This approach performs unsupervised learning, which occurs when features are learned from the data without labels.

Chen et al. (2020) developed SimCLR, which uses data augmentations, an encoder network, a nonlinear projection head and Normalized Temperature-Scaled Cross Entropy Loss to perform image-based, self-supervised contrastive learning. The authors claimed that using two augmentations in order — specifically, applying random cropping, then applying random color distortion — improved generalization. They noted that their implementation exhibited better performance when trained with large batch sizes and for a longer period. The authors compared their approach with supervised approaches using 12 different datasets and ResNet-50, finding that they were able to achieve similar performance.

Radford et al. (2021) developed Contrastive Language Image-Pretraining (CLIP), a foundational visual language model (VLM) enabling zero-shot classification. In their approach, they simultaneously trained two encoders — one for images, and one for text — to successfully predict corresponding text and image pairs. Then, the text encoder could embed associated text and be used for zero-shot prediction. The authors performed pretraining with a newly created dataset containing 400 million image and text pairs. They mentioned that the baseline performance of the approach was only slightly better than that of a supervised ResNet-50 on most datasets. However, it is important to note that it was also able to outperform ResNet-101 with five ImageNet variants. By using this approach, the authors effectively mapped the text and image pairs to a shared embedding space.

Non-Contrastive Representation Learning

Grill et al. (2020) introduced Bootstrap Your Own Latent, which showed that representation learning could be accomplished without using negative pairs on CNNs. The approach was composed of two networks: an online network that learned via backpropagation, and a target network that used the exponential moving average of the online network. Given a view of an image, the online network was used to predict the output of the target network. By learning from multiple views of the same image using a prediction of the other representation, it was possible to enable meaningful feature learning. Similarly, Caron et al. (2021) implemented a self-distillation method named DINO for vision transformers (ViT). In their approach, they used a student and a teacher network, where both networks received differently augmented versions of the same image. The objective of the student network was to predict the output distribution of the teacher network. The authors leveraged centering and sharpening to prevent representation collapse and found that their approach naturally segmented objects.

Kim et al. (2024) used a VLM to create pseudo labels for new classes. A specific prompt that determines whether given labels in an image should be kept or deleted after they have been classified by an object detector was fed into the VLM, generating pseudo labels. Then, they combined the generated pseudo labels with truth labels and retrained the object detection model to incrementally improve its knowledge. An evaluation was performed using PASCAL VOC and COCO. For the COCO dataset, in a scenario with 70 previous classes and 10 new classes, they achieved an improvement of 1.2% mAP over the next-best method. They found that their approach struggled when there was a small amount of pseudo-labeling images available, or when there was a small number of previously known classes in new images.

Foundation Models

As summarized in Table 1, many vision-based foundation models have been released recently. However, no existing approaches have focused on detecting and classifying objects using the SAR modality. This effort aims to develop a SAR-specific foundation model and illustrate its utility for zero-shot maritime object recognition.

Table 1. Overview of Recent Vision-Based Foundation Models

Model Name	Release Year	Type	Key Architecture Components	Primary Training Data Scale/Type	Core Capabilities	Notable Features/Innovations
GPT-4V	2023	Multimodal VLM	Transformer, CNNs, Attention	Diverse text and image datasets	Image/video understanding, Q&A, object detection, text extraction, reasoning	Advanced visual reasoning, pixel bounding-box extraction, real-time multimodal (GPT-4o)
Google Gemini	2023-2025	Multimodal FM	Transformer, SigLIP Vision Encoder	Trillions of tokens (text, image, audio, video)	Image/video understanding, Q&A, object detection, segmentation, content generation	Multimodal integration, scalable variants (Nano, Pro, Ultra), real-time streaming data-processing
LLaVA Family	2023-2025	Multimodal VLM	CLIP ViT-L/14, Vicuna LLM, MLP Projector	GPT-4 generated instruction data (150K), LAION, CC, SBU, OCR-VQA, TextVQA, ScienceQA	Visual instruction following, VQA, OCR-based reasoning, detailed descriptions, multi-turn conversations	Efficient two-stage instruction tuning, synthetic data generation, continual learning (LLaVA-c), smaller models outperforming larger baselines (LLaVA-MORE)
PaliGemma	2024-2025	Multimodal VLM	SigLIP Vision Model, Gemma LLM (Decoder-Only Transformer)	Curated medical datasets (PaliGemma-CXR), PaLI-3 recipes, WebLI	Image/short-video captioning, object detection, text reading, VQA, medical diagnosis/segmentation/report generation	Lightweight, open-source, multi-task medical imaging, text-only prompting for segmentation
Qwen-VL	2023-2025	Multimodal VLM	Qwen-LM, ViT (Openclip's ViT-bigG), Position-Aware Adapter	1.4B cleaned image-text pairs (LAION, CC, COCO Caption), multi-task (VQA, Grounding, OCR), agent data	Image/video understanding, Q&A, grounding, text-reading, document parsing, interactive visual agent	Multilingual, multi-image input, fine-grained visual understanding, dynamic resolution processing, absolute time-encoding for video
SigLIP	2023-2025	VLM (Vision-Language Pretraining)	Standard ViT (Image/Text Tower), MAP Head	WebLI (10B images, 12B alt-texts, 109 languages)	Zero-shot classification, image-text retrieval, VLM transfer, specialized domain adaptation	Pairwise sigmoid loss (efficiency), strong multilingual encoders, integration with domain adaptation techniques (CycleGAN)
MetaCLIP	2024	VLM (Vision-Language Pretraining)	CLIP-Like Architecture, ViTs	MetaCLIP dataset (2B image-text pairs)	Zero-shot image classification, image similarity, image search, captioning/generation, image combination	Transparent, algorithmic data curation, demonstrates visual SSL can match language-supervised pretraining at scale
DINOv2	2023	Pure Visual FM	ViT	LVD-142M (curated from 1.2B uncurated web images, ImageNet, Google Landmarks)	General-purpose visual features, zero-shot classification, segmentation, depth estimation	Self-supervised learning, no fine-tuning/labels needed for SOTA, efficient training (2x faster, 3x less memory), robust data pipeline
InternImage	2024-2025	Pure Visual FM (CNN-Based)	Deformable Convolution v3/v4 (DCNv3/v4)	ImageNet-22K, Joint 427M	Image classification, object detection, instance/semantic segmentation, image generation	CNN-based foundational backbone, DCNv4 for significant speed/performance gains, efficient annotation strategies
Segment Anything Model (SAM)	2023	Specialized (Segmentation)	ViT-H Image Encoder, Prompt Encoder, Lightweight Transformer Mask Decoder	SA-1B (11M images, 1.1B+ masks) via model-in-the-loop data engine	Promptable segmentation (points, boxes, masks), zero-shot generalization	Data engine for large-scale annotation, fast inference, fine-grained detail limitations addressed by community
SAM 2	2024	Specialized (Segmentation)	Transformer-Based Image/Video Encoder, Prompt Encoder, Memory Mechanism, Mask Decoder	SA-V (51,000 videos, 12.61 masklets/video)	Promptable image/video segmentation, real-time (44 fps), zero-shot generalization, interactive refinement, occlusion handling	Unified image/video architecture, streaming memory, significant speed/accuracy improvements over SAM for images
Grounding DINO	2024-2025	Specialized (Object Detection)	Swin Transformer (image), BERT (text), Feature Enhancer, Cross-Modality Decoder	Abundant vision datasets (COCO, Objects365, GRIT, V3Det, RefCOCO families)	Open-set object detection, phrase grounding, referring expression comprehension	Joint text/image training, few-shot learning by training text embeddings, open-source reproduction (MM-Grounding-DINO)
YOLO-World	2025	Specialized (Object Detection)	YOLOv8-Based, Vision-Language Modeling	Expansive datasets	Real-time Open-Vocabulary Detection (OVD)	Efficient "prompt-then-detect" with offline vocabulary, high speed for latency-sensitive applications
YOLO-NAS	2023	Specialized (Object Detection)	Neural Architecture Search (NAS), Quantization-Aware Blocks	COCO, Objects365, Roboflow 100	Object detection (high accuracy, speed)	Quantization-Aware NAS for minimal accuracy drop in INT8, superior accuracy/latency tradeoff, scalable variants

METHODOLOGY

Dataset Overview

The recent proliferation of commercial SAR sensors has greatly expanded the availability of diverse imagery for wide-ranging applications. Umbra's Open Dataset (Umbra, n.d.) provides dozens of terabytes of raw data spanning all phases of the SAR image-formation pipeline to enable exploring modern AI/ML techniques in this unique image domain. For this study, two of the roughly 50 distinct datasets available from Umbra were leveraged for foundation model pretraining and maritime-specific applications, respectively: ad_hoc and ship_detection_testdata.

The ad_hoc dataset provides comprehensive global imagery with wide diversity, consisting largely of feature-dense regions, such as cities and dense suburban areas. This is critical for effective foundation model development. Overlapping spatial imagery collected at different times also gives critical diversity to enable effective contrastive foundation model pretraining. Ignoring things that are typically irrelevant and should not be among the learned features, such as perspective, shadow and season changes, helps the foundation model to focus on relevant semantic features that do matter during pretraining. This yields a stronger encoder for low- or zero-shot object characterization.

Similar diversity and temporal coverage exist in the ship_detection_testdata dataset, which is used for foundation model fine-tuning and manually extracting relevant ship examples (Figure 3). Additionally, manually annotated ship bounding boxes in the open-source Umbra ship-detection dataset (Zhiyong, 2024) were extracted to assess the separability of different ship types within a pretrained foundation model.

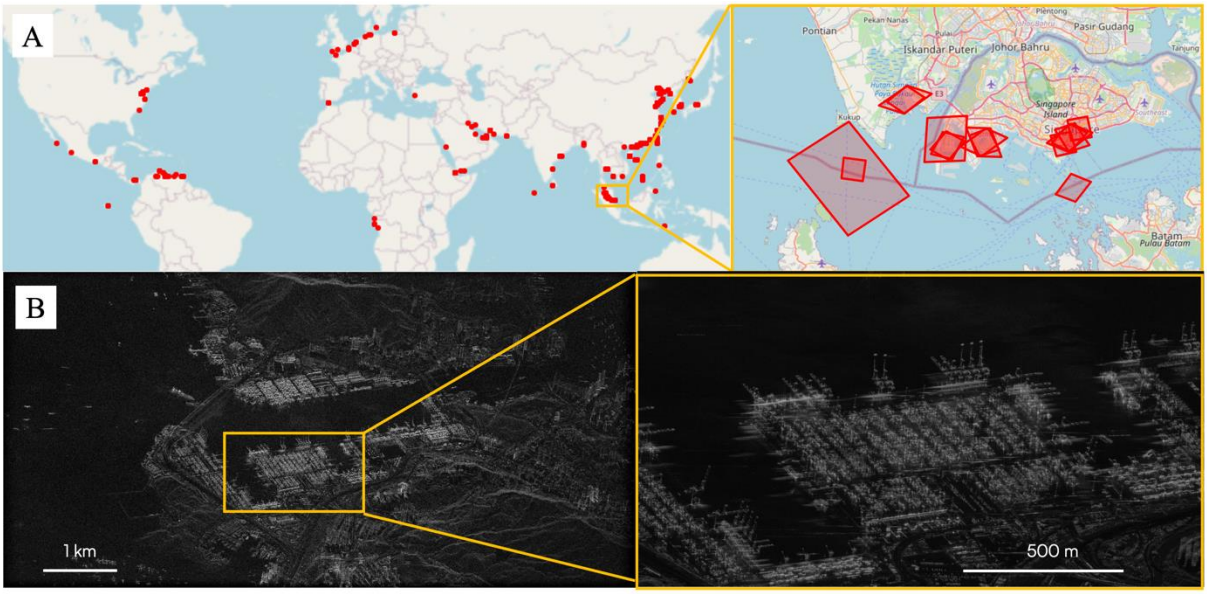

Figure 3. A) The Umbra Open Dataset ship-detection dataset spans many diverse ports all over the globe, with sizeable image footprints and multiple collections over the same area, providing critical content for foundation model pretraining. B) Example Umbra image over the Port of Hong Kong, illustrating the full breadth of its high-resolution images and examples of smaller, more manageable tiles extracted for foundation model pretraining.

Training Approach and Experimental Setup

The overall experimental process (Figure 4) begins with model pretraining, which provides a feature-rich encoder that can be used for downstream task-specific exploitation, such as zero-shot recognition of maritime objects. For foundation model pretraining, two key innovations were leveraged to adapt remote-sensing deep learning algorithms to the zero-shot regime: self-supervised pretraining (e.g., MAE [He et al., 2022]) and cross-modal latent space alignment (Theodoridis et al., 2020). These and other, similar innovations are what have made ChatGPT and similar technology ubiquitous.

Foundation models need substantial data volumes to learn all the relevant features in the sensors or modalities of interest. However, this approach can be made significantly more efficient by ensuring that meaningful scene content is present in the imagery. Before the model was pretrained, rich camera geometry modeling and legacy capabilities in image registration were incorporated. This allowed for performing the orthorectification and alignment of all imagery with open-source map data (e.g., Overture Maps). In turn, this enabled the automated extraction of relevant scene content.

After sufficiently diverse data had been curated and tiled, pretraining was carried out via MAE. From 734 total images spanning the Umbra Open Dataset datasets, 132,508 tiles were used, and a SAR-specific foundation model was trained

for downstream tasks. Remote sensing-specific innovations, such as Cross-Scale MAE (Tang et al., 2023), ensure sensor, perspective and resolution agnosticism.

Next, relevant holdout data, including open-source ship-detection maritime chips, was procured and fed through the pretrained encoder to extract the high-dimensional features of each chip. These datasets are typically heavy on civilian and commercial ships, so imagery was hand-curated for the separate military ports present in the open dataset. This allowed for manually extracting military vessels to serve as the rare targets to be zero-shot categorized.

The preprocessing of these image chips must be carried out carefully. This is because SAR deep learning models can be highly sensitive to different algorithms that might drastically alter the realism of the noise and speckle patterns fundamental to the modality. For this effort, three different approaches to image chipping and preprocessing were explored to see how they impact the downstream clustering performance (Figure 5):

1) **Image Resizing:** This is the direct resizing of extracted image chips to a predetermined, consistent size (256 by 256 pixels), using bilinear interpolation and potentially altering the SAR fundamental image characteristics.
2) **Spatial Zero Padding:** This retains the true spatial size of an image by simply zero-padding it up to 256 by 256 pixels. Very few ships at the Umbra resolution are larger than this; those that are would require downsampling. This preserves the true noise characteristics of the SAR images.
3) **Fourier Zero Padding:** This does not preserve the spatial size, done deliberately to ensure that feature assessment occurs on object features, versus grouping things only by their size similarity. However, it allows for properly sampling speckle noise during interpolation instead of performing the unrealistic smoothing that occurs during standard resizing.

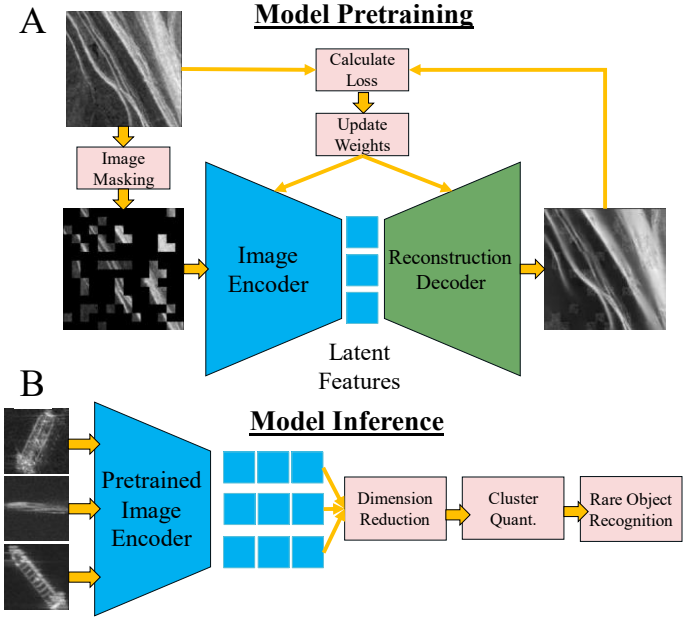

Figure 4. Full experimental algorithm flow, spanning upfront pretraining via MAE (A), and the downstream exploitation of that pretrained encoder for object-specific clustering and rare object recognition (B).

Embeddings were then generated for each unique object by feeding them through the encoder and extracting the latter layer structure. As these features are represented as abstract high-dimensional vectors, dimensionality reduction was performed to visualize and quantify the similarity between different objects' embedding vectors. Several dimensionality reduction techniques exist, but UMAP (McInnes et al., 2018) was preferred due to its rapid speed and scalability to wide ranges of different data volumes.

To quantify cluster quality after reducing the dimensionality, several metrics were considered: Davies-Bouldin Index, Calinski-Harabasz Index, silhouette score, scattering-density between clusters and density-based clustering validation. An unsupervised feedback loop using a combination of these metrics to automatically assign the optimal number of groups helped to demonstrate the ideal combination of algorithm and parameter choice for a given foundation model. Finally, with the optimal clustering techniques and foundation models identified, it was possible to see unique or anomalous groupings in the 2D representation of the latent feature space, enabling the identification of distinctly isolated novel objects. These groupings could then be used as the seed for downstream iterative active learning for true ~~fully~~ zero-shot rare object recognition (Reisman et al., 2025).

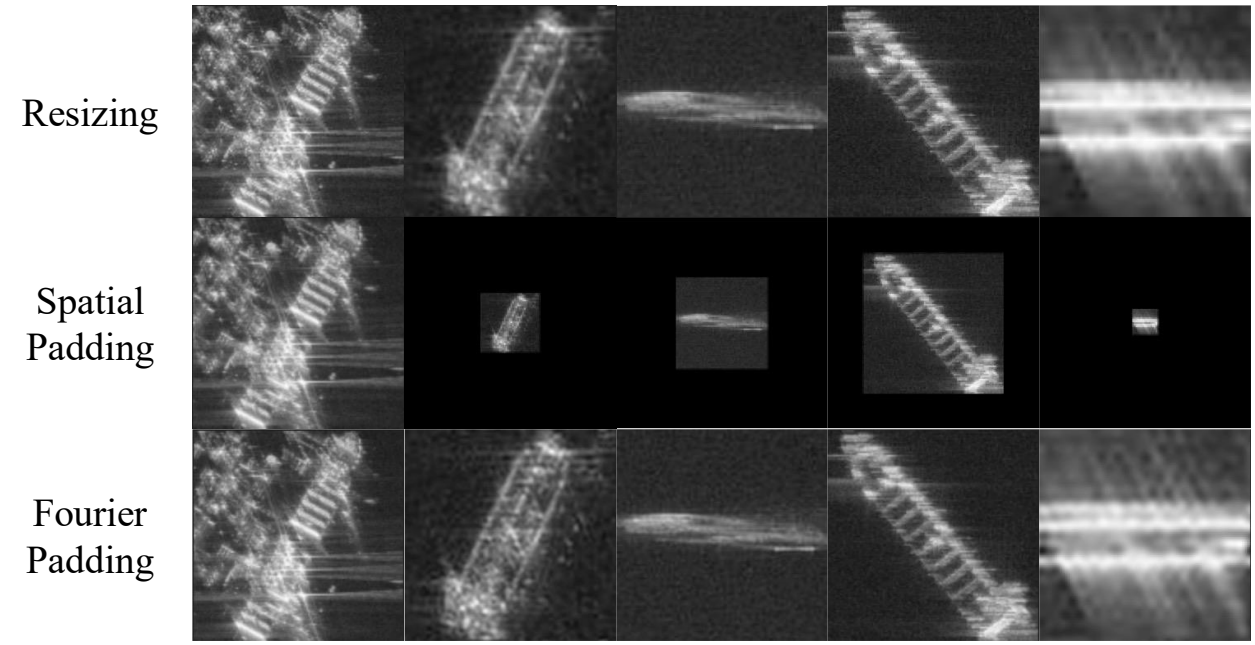

Figure 5. Five examples of different ships preprocessed by each of the three chipping and resizing algorithms, illustrating the differences in preserved spatial size and underlying noise characteristics.

To assess the value of SAR-specific foundation models for zero-shot object categorization, three different model pretraining approaches were used:

1) **Open Source Pretrained Model:** Here, DINOv2 (Oquab et al., 2023) was used, which contains widely diverse visible-image-domain examples but is not specific to overhead imaging or remote sensing. This model reveals whether critical features of SAR images are separable in a well-trained but unrelated foundation model that has never seen a SAR image before.
2) **Custom EO VIS Foundation Model:** Starting with a pretrained open-source model, a remote-sensing-specific foundation model encoder was fine-tuned on open-source imagery from Christie et al., 2018. This model has also never seen a SAR image, but it became highly sensitive to critical overhead image features after training on global satellite imagery, which might suffice for SAR satellite image feature separability.
3) **Custom SAR Foundation Model:** The pretrained EO VIS foundation model was further fine-tuned for the SAR domain by leveraging the full ad hoc Umbra open dataset. This determined whether the enhanced sensitivity to SAR-specific features more cleanly separated objects of interest for zero-shot recognition.

RESULTS

Overall cluster-quality metrics were calculated for each combination of foundation model, chipping/resizing algorithm and clustering metric. Clusters were automatically generated using three different approaches: HDBScan, fixed K-means clustering (20 clusters) and automatic K-means clustering via silhouette scores. Figure 6A shows the optimal clustering from each of these three methods for each of the five cluster metrics. Note that S_DBW and density-based clustering validation are optimized when their metric is smallest, and the other three are optimized when their metric is largest.

A

Max Across All Metrics									
	Open Foundation Model (DINOv2)			Custom EO VIS Foundation Model			Custom SAR Foundation Model		
Metric	Resizing	Spatial Padding	Fourier Padding	Resizing	Spatial Padding	Fourier Padding	Resizing	Spatial Padding	Fourier Padding
Silhouette	0.48	**0.565**	0.445	**0.537**	0.509	0.501	0.574	**0.621**	0.593
Calinski	425.11	**828**	575.5	847.7	**2478.6**	993.1	5591.4	3797.1	**6119.8**
Davies	0.855	0.959	**0.997**	0.765	0.673	**0.827**	0.65	**0.69**	0.591
S_DBW	1.047	**1.024**	1.048	1.052	1.054	**1.048**	1.041	**1.028**	1.041
DBCV	-0.148	-0.338	**-0.434**	-0.135	-0.076	**-0.582**	**-0.452**	-0.028	0.005

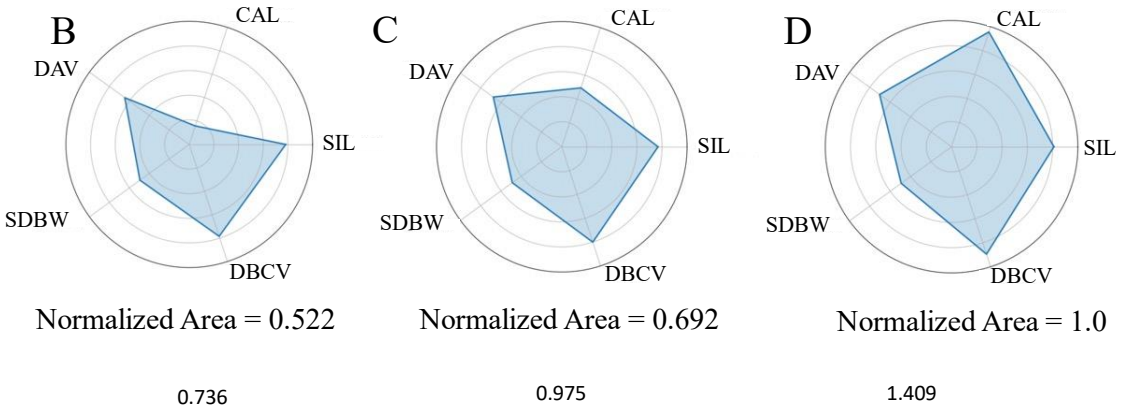

Figure 6. A) Summary of optimal cluster scores from all five metrics, for all combinations of foundation model and preprocessing approach. B-D) Spider plot representations of all five optimal metrics for the Open Foundation Model, the Custom EO VIS model and the Custom SAR model, respectively. Measuring the normalized area of these polygons provides a singular metric to capture the overall performance of each model.

Within each foundation model, the optimal clustering performance (indicated in boldface in Figure 6) is seen to be following either spatial padding or Fourier padding preprocessing in almost all cases. This confirms the expectation that the preservation of spatial features and SAR-specific image characteristics is critical to the maximal exploitation of pretrained encoders for object recognition. The overall optimal clustering performance from any combination of foundation model and preprocessing (indicated in underlined boldface in Figure 6) is seen to occur somewhat diversely across the three foundation models, depending on the metric. Based on the situations that each of these metrics is typically considered for, it appears that the SAR-specific model is optimal for distance-based clustering metrics. By contrast, the general image foundation models, which were tuned on far more examples than the SAR model, are better at density clustering.

Upon visual inspection, the SAR-specific model has much more consistent clustering than the open source or custom EO VIS models do. The latter two produce highly different 2D clustering representations with very small modifications to the dimensionality reduction and clustering parameters, whereas the SAR-specific model always shows a more consistent structure. To qualitatively investigate these cluster representations, the open-source FiftyOne tool from Voxel51 (*FiftyOne*, n.d.) was utilized. Figure 7 shows example subclusters within this 2D feature space for the overall optimal combination of SAR-specific foundation model and preprocessing algorithm.

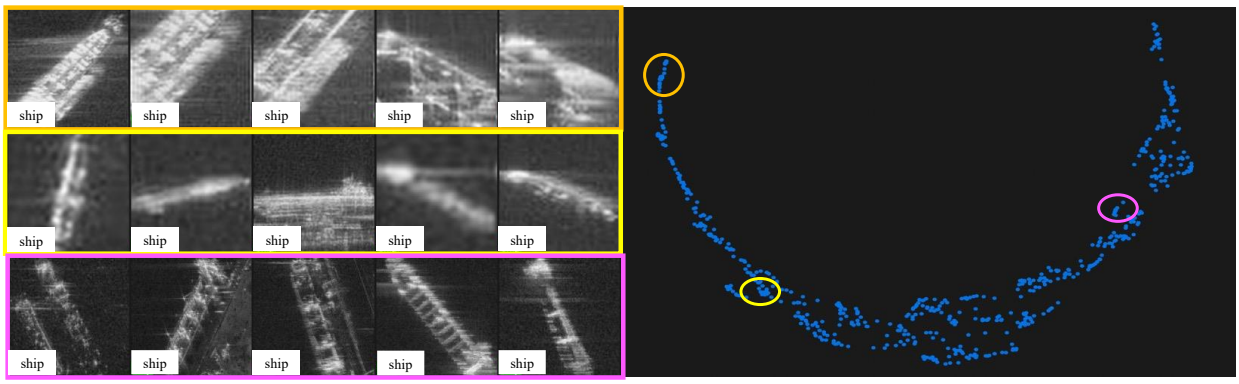

Figure 7. Three subregions of the 2D feature space are highlighted, showing that nearby groupings of ships share similar characteristics and features, despite the foundation model having been trained only on generic, broad SAR imagery.

Finally, hand-curated, rare objects (e.g., military ships chipped from the Port of Hong Kong) were introduced to see how they were interpreted by the encoder. They were then compared to the roughly 1,000 other ships pulled from the open-source ship-detection dataset (Figure 8). Of the 70 military-ship chips introduced, which had been preprocessed using the same techniques as those used for the open dataset, the bulk of them stood independent and isolated from the generic ships (indicated by yellow and orange circles in figures 7 and 8). A few ships appear similar to some of the generic ships (indicated by pink circles in figures 7 and 8), possibly indicating one of two things: either the presence of a small number of military ships in the open set or, more likely, a need for further fine-tuning on more port-specific imagery to diversify the learned ship features.

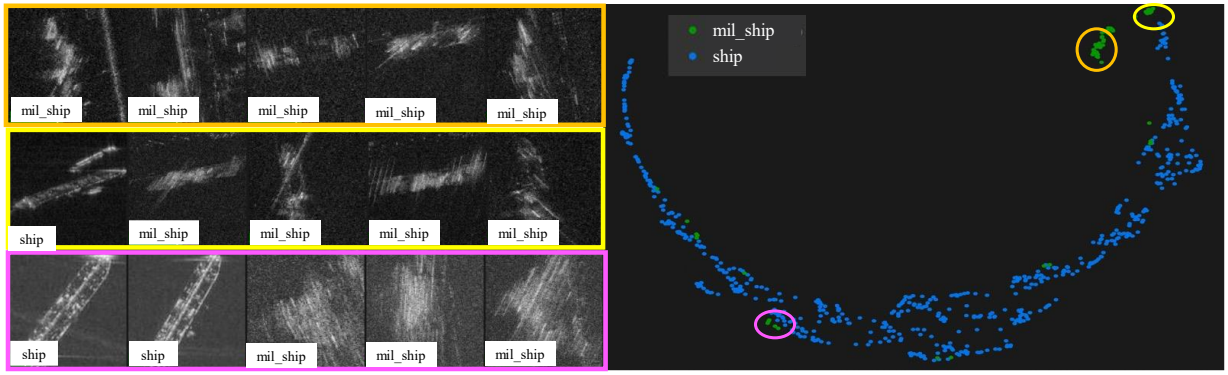

Figure 8. The majority of the military ships introduced to the SAR-specific foundation model are isolated from the other ships, representing their uniqueness and the ability of this foundation model to group related objects in a zero-shot fashion.

CONCLUSION AND FUTURE WORK

This research successfully demonstrated a novel approach to ATR for SAR imagery. It addressed the critical challenge of developing AI systems that perform reliably and consistently even when encountering new contexts without extensive prior training. The work focused on a zero-shot learning paradigm that eliminates the need for the additional learning stages and representative data typically required for context-switching between different modalities, such as EO and SAR images; in such instances, accuracy and confidence levels usually decrease without those added things.

A key contribution of this study is the development and illustration of a SAR-specific foundation model. This is a significant advancement, given that no existing approaches have previously focused on detecting and classifying objects using the SAR modality in this manner. By leveraging latent space with real-time, Transformer-based approaches, image data was encoded onto a manifold, and clustering algorithms were used to identify objects with similar physical features. This approach effectively maps out objects in Euclidean space to determine their similarity

to other images that a CNN already knows. The experimental results underscore the importance of preprocessing techniques in maximizing the utility of pretrained encoders for object recognition. Specifically, it was found that optimal clustering performance consistently occurred when using either spatial padding or Fourier padding preprocessing. This confirmed that the preservation of spatial features and SAR-specific image characteristics is critical. In contrast, simple image resizing often altered fundamental SAR image characteristics, impacting performance.

Overall, this work demonstrated how different foundation models that have been solely trained with unlabeled, unstructured imagery can learn critical features for subtle rare-object discrimination. Most significantly, this study shows the ability of the SAR-specific model to isolate rare, high-value targets. That ability is attributable to its optimally identified clustering metrics and preprocessing techniques. Despite never having seen them before, nor having ever seen a label of any image, object or context, the model was able to perform zero-shot target recognition. This capability is crucial for Department of Defense applications, as it enables operators to make sense of new situations without the time-consuming and expensive process of acquiring large amounts of labeled data for every new target.

To build upon these promising results, several avenues for future research have been identified. The identified clusters can serve as starting labels for the automated iterative active learning process discussed in Reisman et al., 2025. This not only will enable automatically labeling any additionally collected imagery, but also will highlight edge cases in which the algorithm-derived labels are most likely to be inaccurate. This approach will continually refine the model's knowledge and further reduce human labeling effort.

Another future direction is the investigation of a maritime SAR-specific foundation model. This would be more sensitive to the relevant features and contexts seen in and around ports, instead of training on a massive, global SAR dataset. It could be explicitly deployed in regions known to contain ports, and the more generic, global model would be reserved for novel locations and situations. Further development in this area could yield even greater sensitivity to relevant features and contexts found in and around ports.

Future efforts will also focus on optimizing these models for real-time operational deployment. They will explore computational efficiencies to ensure high-speed processing for latency-sensitive applications, and will assess hardware integration requirements for battlefield applications. Although the Umbra Open Dataset provides diversity, further testing with even more varied SAR datasets — notably, those with different sensor parameters, environmental conditions and clutter types — will enhance the model's generalization capabilities.

This SAR-specific foundation model can be applied well beyond maritime targets, to other critical object classes or types of targets relevant to defense and security, such as land vehicles or infrastructure. Doing so will demonstrate its broader utility. Also, as AI models become more integrated into critical defense applications, future work should investigate potential algorithmic biases within the SAR-specific foundation model. From there, methods to improve model interpretability could be explored to ensure trust and explainability in decision-making processes.

REFERENCES

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). *Emerging Properties in Self-Supervised Vision Transformers* (No. arXiv:2104.14294). arXiv. https://doi.org/10.48550/arXiv.2104.14294

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*, 1597–1607. https://doi.org/10.48550/arXiv.2002.05709

Christie, G., Fendley, N., Wilson, J., & Mukherjee, R. (2018). *Functional Map of the World* (No. arXiv:1711.07846). arXiv. https://doi.org/10.48550/arXiv.1711.07846

Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D. B., & Ermon, S. (2022). SatMAE: pre-training transformers for temporal and multi-spectral satellite imagery. *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 197–211.

FiftyOne: Multimodal Data Platform for Computer Vision. (n.d.). Retrieved June 19, 2025, from https://voxel51.com/fiftyone

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., & Valko, M. (2020). *Bootstrap your own latent: A new approach to self-supervised Learning* (No. arXiv:2006.07733). arXiv. https://doi.org/10.48550/arXiv.2006.07733

He, K., Chen, X., Xie, S., Li, Y., Dollar, P., & Girshick, R. (2022). Masked Autoencoders Are Scalable Vision Learners. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15979–15988. https://doi.org/10.1109/CVPR52688.2022.01553

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). *Momentum Contrast for Unsupervised Visual Representation Learning* (No. arXiv:1911.05722). arXiv. https://doi.org/10.48550/arXiv.1911.05722

Huang, G., Liu, X., Hui, J., Wang, Z., & Zhang, Z. (2019). A novel group squeeze excitation sparsely connected convolutional networks for SAR target classification. *International Journal of Remote Sensing*, *40*(11), 4346–4360. https://doi.org/10.1080/01431161.2018.1562586

Iqbal, M. A., Yoon, Y. C., Khan, M. U. S., & Kim, S. K. (2023). Improved Open World Object Detection Using Class-Wise Feature Space Learning. *IEEE ACCESS*, *11*, 131221–131236. https://doi.org/10.1109/ACCESS.2023.3335602

Kim, J., Ku, Y., Kim, J., Cha, J., & Baek, S. (2024). *VLM-PL: Advanced Pseudo Labeling approach for Class Incremental Object Detection via Vision-Language Model*. 4170–4181. https://doi.org/10.1109/CVPRW63382.2024.00420

Li, J., Yu, Z., Yu, L., Cheng, P., Chen, J., & Chi, C. (2023). A Comprehensive Survey on SAR ATR in Deep-Learning Era. *Remote Sensing*, *15*(5), Article 5. https://doi.org/10.3390/rs15051454

McInnes, L., Healy, J., & Melville, J. (2018, February 9). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. arXiv.Org. https://arxiv.org/abs/1802.03426v3

Moreira, A., Prats-Iraola, P., Younis, M., Krieger, G., Hajnsek, I., & Papathanassiou, K. P. (2013). A tutorial on synthetic aperture radar. *IEEE Geoscience and Remote Sensing Magazine*, *1*(1), 6–43. https://doi.org/10.1109/MGRS.2013.2248301

Oord, A. van den, Li, Y., & Vinyals, O. (2019). *Representation Learning with Contrastive Predictive Coding* (No. arXiv:1807.03748). arXiv. https://doi.org/10.48550/arXiv.1807.03748

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., … Bojanowski, P. (2023, April 14). *DINOv2: Learning Robust Visual Features without Supervision*. arXiv.Org. https://arxiv.org/abs/2304.07193v2

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language

Supervision. *Proceedings of the 38th International Conference on Machine Learning*, 8748–8763. https://doi.org/10.48550/arXiv.2103.00020

Reisman, M. D., LaTourette, K., LeDuc, D., Shagnea, P., & Lindenbaum, A. (2025). Synthetic data-seeded active learning for automated ATR data labeling. *Automatic Target Recognition XXXV, 13463*, 179–189. https://doi.org/10.1117/12.3053822

Ren, H., Yu, X., Zou, L., Zhou, Y., Wang, X., & Bruzzone, L. (2021). Extended convolutional capsule network with application on SAR automatic target recognition. *Signal Processing*, *183*, 108021. https://doi.org/10.1016/j.sigpro.2021.108021

Sastry, S., Khanal, S., Dhakal, A., Ahmad, A., & Jacobs, N. (2025). *TaxaBind: A Unified Embedding Space for Ecological Applications*. 1765–1774. https://doi.org/10.1109/WACV61041.2025.00179

Shmelkov, K., Schmid, C., & Alahari, K. (2017). Incremental Learning of Object Detectors without Catastrophic Forgetting. *2017 IEEE International Conference on Computer Vision (ICCV)*, 3420–3429. https://doi.org/10.1109/ICCV.2017.368

Tang, M., Cozma, A., Georgiou, K., & Qi, H. (2023). Cross-scale MAE: a tale of multi-scale exploitation in remote sensing. *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 20054–20066.

Theodoridis, T., Chatzis, T., Solachidis, V., Dimitropoulos, K., & Daras, P. (2020). *Cross-Modal Variational Alignment of Latent Spaces*. 960–961. https://openaccess.thecvf.com/content_CVPRW_2020/html/w56/Theodoridis_Cross-Modal_Variational_Alignment_of_Latent_Spaces_CVPRW_2020_paper.html

Umbra. (n.d.). *Umbra Synthetic Aperture Radar (SAR) Open Data - Registry of Open Data on AWS* [Dataset]. Retrieved June 19, 2025, from https://registry.opendata.aws/umbra-open-data/

Wang, W., Zhang, C., Tian, J., Ou, J., & Li, J. (2020). A SAR Image Target Recognition Approach via Novel SSF-Net Models. *Computational Intelligence and Neuroscience*, *2020*(1), 8859172. https://doi.org/10.1155/2020/8859172

Xie, Y., Dai, W., Hu, Z., Liu, Y., Li, C., & Pu, X. (2019). A Novel Convolutional Neural Network Architecture for SAR Target Recognition. *Journal of Sensors*, *2019*(1), 1246548. https://doi.org/10.1155/2019/1246548

Zhiyong. (2024). *umbra-ship* [Open Source Dataset]. Roboflow Universe. https://universe.roboflow.com/zhiyong/umbra-ship