DOMAIN-SPECIFIC FOUNDATION MODELS FOR SCIENCE APPLICATIONS: SELF-SUPERVISED LEARNING WITH SAR AND DXA

A DISSERTATION SUBMITTED TO THE GRADUATE DIVISION OF THE UNIVERSITY OF HAWAIʻI AT MĀNOA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

COMPUTER SCIENCE

AUGUST 2025

By

Yannik Glaser

Dissertation Committee:

Peter Sadowski, Chairperson
Huaijin Chen
John Shepherd
Justin E. Stopa
Mahdi Belcaid

Keywords: deep learning, computer vision, self-supervised learning, representation learning, dual-energy x-ray absorptiometry, synthetic aperture radar

ACKNOWLEDGMENTS

Doing a PhD has been an immensely enriching experience that truly would not have been possible without the help and support of countless people.

I would like to thank Dr. Peter Sadowski, my advisor, mentor, and friend. Your guidance, support, and patience have been the defining pieces of this journey. I could not have asked for a better professor to join the university the same semester I did, and I am forever grateful that you gave me the chance to work with you from start to finish. I strive every day to be as patient and generous as you have been to me.

In a similar vein, I would like to thank Dr. John Shepherd and Dr. Justin Stopa, who have also been mentors throughout this journey. Overseeing many of the projects I have worked on from the very beginning, you have been a constant presence throughout and shaped me as a researcher and person. Your expertise and passion for your fields have been an inspiration that I can only hope to emulate.

Finally, I would also like to thank the last two members of my committee, Dr. Mahdi Belcaid and Dr. Huaijin (George) Chen. This dissertation would not have been possible without your input and critical questions. Thank you, Mahdi, for always being a bright presence in the department and for your selfless support and interest. Thank you, George, for spontaneously agreeing to serve as a committee member right after joining UH, and always being at the ready with support and feedback.

I wish I could thank all the collaborators I have had throughout this journey. I can genuinely say I have had nothing but pleasant experiences with every project I have been lucky enough to be a part of. The patience and generous mentorship of everybody I have worked with have shaped my appreciation for academia and admiration for the people dedicating their lives to it.

I would also like to thank everybody who has been part of the Sadowski lab since I joined in 2018, but especially Arianna, Yusuke, Linnea, and Michael. You have been around for most of this process and have always been supportive and a reminder to cherish this process for more than the title at the end. Thank you.

Lastly, I would like to thank my family and friends. But especially, I would like to thank my parents. You have never shown me anything but unwavering support and love; this is as much your achievement as it is mine. Ich werde euch niemals genug danken können.

ABSTRACT

This dissertation explores the use of self-supervised pre-training in non-natural image domains and provides three main contributions to the literature: 1) it adapts current self-supervised learning frameworks to pre-train a model specific to synthetic aperture radar Wave mode imagery; 2) it adapts current self-supervised learning frameworks to pre-train a model specific to dual-energy x-ray absorptiometry; 3) it analyzes embedding characteristics of both models to identify representation quality metrics effective beyond natural image applications.

The immediate goal of this work is to provide embedding models that generalize effectively to a range of downstream tasks in their respective domains. These models serve as highly specific foundation models — they generalize well to in-domain tasks, they are robust to training settings and hyperparameter choices, and they are extremely labeled-data-efficient.

Training models with self-supervised methods that are tuned to the characteristics of the data domain is important because most self-supervised frameworks are highly tuned for optimal performance on natural images. By adapting these frameworks to respect domain-specific characteristics of the data or simply removing natural-image-focused biases, downstream task performance and generalizability can be improved.

The secondary goal is to add to the body of literature exploring representation characteristics, searching for embedding space qualities that indicate a well-performing model without access to labeled data for direct evaluation. This addresses a crucial bottleneck for similar domain-specific pre-training efforts where architecture search, hyperparameter tuning, and comparison between self-supervised methods are all hindered by the need to train each candidate model to completion and evaluate performance on specific downstream tasks. By training novel embedding models for two separate vision domains and extensively analyzing intermediate representations of successful and unsuccessful models, this study seeks to establish a foundation for future research attempting similar pre-training efforts for other computer vision domains beyond natural images.

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

1.1 Problem Statement

Self-supervised learning (SSL) has been a crucial tool for overcoming one of the major bottlenecks in deep learning — the need for labeled data. By training on signals inherently present in the data, SSL has been the driving factor in rapidly improving models in data-rich domains such as natural language processing[1, 2] and computer vision for natural images (NIs) [3, 4, 5, 6, 7], leading to the creation of foundation models [8], models pre-trained on a vast dataset, enabling generalization to a wide range of downstream tasks with minimal or no tuning.

In computer vision, there are two dominant approaches to self-supervised representation learning: masked image modeling (MIM) and joint-embedding self-supervised learning (JESSL). MIM is a reconstructive approach, usually applied in tandem with vision transformers, where a portion of input tokens are masked, and the goal is to reconstruct the image on a pixel [4] or token level [9]. JESSL frameworks rely on a Siamese structure, where the objective is for both Siamese arms to learn similar representations for differently augmented versions of the same original image. These frameworks are more varied, with many mechanisms employed to avoid representational collapse.

Literature on both SSL approaches, however, suffers from a bias toward NI — images that one could take with a camera out in the world — since most of the biggest popular datasets for model development and benchmarking are comprised of NIs [10, 11, 12, 13, 14]. As such, frameworks are often tuned on this modality, and generalization to other vision domains with varying degrees of difference to NIs is suboptimal [15, 16]. Two relevant domains are remote sensing and medical imaging, both of which collect image data with fundamental differences to NIs. In remote sensing, depending on the sensor, data can have varying numbers of channels compared to NIs which usually have 1 channel (grayscale), 3 channels (RGB), or 4 channels at most (RGBA). There are also semantic differences, such as some remote sensing modalities capturing objects of vastly different scales and content, meaning features that the model might learn as useful for NIs would likely not translate. Similarly, while medical imaging can also differ in the number of channels, semantic differences account for much of the domain gap to NIs. Further, depending on the medical modality, pixel intensities might also correspond to set physical qualities of the imaged tissue [17].

Considering JESSL frameworks, it is immediately apparent how the NI bias is problematic for these domains. Augmentations are chosen specifically to encourage invariances desirable in NI applications [3, 18, 16] and the implications of augmentations such as color or brightness distortions change when images do not have regular RGB channels or pixel brightness are unclear. This is also true for geometric augmentations, such as the popular crop and zoom augmentation [18, 19] which could cause representations to disregard critical scale-dependent information for physics information

in remotely sensed data.

MIM is less intuitive in this regard but also has implicit biases that could potentially negatively impact representation quality. Since the objective is image reconstruction, specific remote sensing domains with highly homogeneous images might not be a good fit for this objective as it is not a sufficiently challenging task. Similarly, medical images taken according to strict protocols have relatively low variability across samples, which could cause MIM representations to collapse toward a local minimum, predicting the mean scan while ignoring salient features for downstream tasks.

Ultimately, it is unclear how exactly SSL frameworks affect representation characteristics and thus, transfer performance. It is however, clear that simply applying out-of-the-box frameworks to non-NI domain data is not ideal as it results in suboptimal task performance [16].

Additionally, extensive hyperparameter and framework tuning is especially expensive for SSL, which relies on large vision backbones [14, 20] and benefits heavily from large sample sizes and extended training [14, 21]. Several works have aimed to address this issue by proposing embedding quality metrics that serve as indicators for downstream performance of learned embeddings [22, 23, 24]. However, evaluation of these metrics is also biased toward NI datasets for similar reasons as the SSL frameworks. Additionally, evaluation of how metrics are correlated to downstream task performance is often limited to single-class classification or object detection tasks, missing regression and multi-label classification problems [25, 26], among others. Lastly, most of this work has been conducted using JESSL representations [22, 23, 27], with the literature on MIM representation characteristics severely lagging. As a consequence, while these metrics are a promising tool for reducing the cost of adapting SSL frameworks to new tasks, the current literature is insufficient to tell which metrics are robust and general enough to guide framework selection and adaptation to non-NI domains.

1.2 Motivation

Given the demonstrable success of SSL methods in computer vision, it makes sense to use them to pre-train general models for other, non-NI, vision domains. Domain-specific pre-training promises higher representation quality with better downstream task transfer performance and in-domain generalizability[28, 29, 30]. Remote sensing and medical imaging especially are two domains that could benefit immensely from more focused SSL pre-training frameworks: both have fundamentally different image characteristics from NI data, and while this depends on the exact image-acquisition methods it is generally true for most of the data collected in both domains; both require domain experts for labeling and have a diverse set of potential downstream applications, making it unfeasible to compile sufficiently large labeled datasets for every valuable application; both have a historical backlog of images acquired over many years or even decades, making it possible to compose relatively large, unlabeled datasets for pre-training purposes; both have high-impact downstream applications with real-world implications that would benefit from the availability of domain-specific pre-trained

models to improve performance.

Exemplary data types from both domains are synthetic aperture radar (SAR) imagery from remotely-sensed, satellite-borne systems and dual-energy x-ray absorptiometry (DXA) from medical imaging.

The European Space Agency's (ESA) Sentinel-1 mission has been collecting data continuously since 2014 across various acquisition modes, amassing a multi-million image archive [31]. While imagery collected by acquisition modes focused on landmasses has been of interest to diverse studies with multiple applications [32, 33, 34, 35], the wave-mode (WV) images of the ocean surface are, by comparison, relatively understudied. These images are collected over the open ocean, and their primary purpose is to monitor and study the sea state (e.g. wind, waves, and currents) [36]. However, as more recent studies have explored, they also capture a wide range of other information about atmospheric conditions and geophysical phenomena [37]. The almost 10M WV images collected by the S1 mission could therefore be used for various downstream applications ranging from sea ice observation to monitoring atmospheric stability regimes, but actual utilization has been limited. These images have fundamental differences from NIs, including systematic backscatter noise and not being object-centric like many NI datasets, which makes transferring pre-trained models on NI suboptimal and brings into question biases of standard SSL frameworks.

Despite the sensitive nature and often high logistical and monetary cost of acquisition of medical data making it generally harder to create datasets larger than several thousands of images, DXA has been a modality included in several long-running longitudinal National Institute of Health (NIH) studies, accumulating around 100k readily available scans. DXA captures comprehensive body-composition information, but applications of deep learning to the modality have often been focused on fracture risk or fracture detection [38, 39, 40], with fewer works exploring more opportunistic measures that could be derived from DXA, which is a standard measure for body-composition and bone density assessment [41, 42, 43]. Object-centric, information contained in DXA scans is more diffuse than in NI, and relative pixel intensity is explicitly physically meaningful and relevant to downstream applications, something that many SSL frameworks deliberately encourage invariance to through brightness augmentations.

While it is beyond the scope of any single work to make domain-spanning SSL pre-training framework recommendations for all of remote sensing or medical imaging, focusing on specific applications from both domains will help shed light on how similar pre-training efforts can be undertaken effectively in the future while also providing much-needed, domain-specific pre-trained models for downstream applications. Additionally, given the scope of such an effort, opportunistically studying representation characteristics throughout the pre-training process will also benefit other SSL pre-training applications. An empirical study of embedding quality metrics that have been proposed in the SSL literature for architecture search, hyperparameter tuning, and model-selection outside of common evaluation datasets can supplement the literature on desirable embedding attributes

beyond of NI data.

1.3 Research Question

This dissertation aims to primarily address the two following research questions

1. Can domain knowledge be combined with current SSL approaches to learn a SAR 20-km ocean image embedding model with improved transfer performance over approaches trained on NI data?

2. Can domain knowledge be combined with current SSL approaches to learn a DXA image embedding model with improved transfer performance over approaches trained on NI data?

Both research questions contain two main hypotheses that are tested explicitly. **(1)** An embedding model trained in a self-supervised manner on in-domain data can improve over one simply trained on a standard NI dataset (ImageNet-1K [10]). And **(2)** the self-supervised framework used to train the embedding model can be improved by including domain-specific modifications to the original framework.

Opportunistically, this dissertation also evaluates a secondary question given embedding models emerging from exploring the primary research questions:

3. Which embedding quality metrics generalize well outside NI applications and are well-correlated with general downstream task performance.

To evaluate the primary research questions, similar approaches are taken, tailored to each data domain, available data, and compute resources. For self-supervised frameworks, two JESSL frameworks are evaluated, SimCLR [3] and BYOL [44] with a sampling of standard computer vision architectures serving as encoder architectures. A standard MAE [4] with a ViT-S backbone is trained as a representative MIM approach. All models are evaluated on domain-relevant regression and classification tasks. Standard approaches from the pre-training literature are used to estimate the quality of the pre-trained models, ranging from linear separability of the fixed embeddings to end-to-end finetuning.

In tandem with the downstream task performance analysis, embedding quality metrics are monitored throughout model training. The observed metrics during pre-training are evaluated for correlation with task performance on downstream tasks for the different evaluation settings to see which metrics are most predictive of a well-behaved embedding model. Unsupervised metrics designed specifically for this purpose, such as RankMe [22] and α-ReQ [23] will be monitored.

1.4 Contributions

- A generalizable, pre-trained SAR WV foundation model that has been thoroughly tested and evaluated and is publicly available.
- A generalizable, pre-trained DXA foundation model that has been thoroughly tested and evaluated. The model can process multiple anatomical regions and will be made publicly available to researchers.
- Comprehensive empirical results on using embedding space metrics for non-NI imaging data. Unsupervised representation quality metrics are evaluated against unusual downstream tasks on unusual data, providing practitioners with valuable insights into the limits and efficacy of these metrics and their utility for hyperparameter tuning, model selection, and guiding framework alterations.
- A repeatable recipe for training domain-specific self-supervised foundation models in non-NI vision domains.

1.5 Publications:

The following publications are presented in chronological order and related to this dissertation, showcasing the evolution of the work:

- Yannik Glaser, Peter Sadowski, Thomas Wolfgruber, Li-Yung Lui, Steven Cummings, and John Shepherd. Hip fracture risk modeling using dxa and artificial intelligence. In Journal of Bone and Mineral Research, volume 35, pages 200–200. John Wiley and Sons and The American Society for Bone and Mineral Research, 2020.
- Brandon Quach, Yannik Glaser, Justin Edward Stopa, Alexis Aur´elien Mouche, and Peter Sadowski. Deep learning for predicting significant wave height from synthetic aperture radar. IEEE Transactions on Geoscience and Remote Sensing, 59(3):1859–1867, 2021. doi: 10.1109/TGRS.2020.3003839.
- Yannik Glaser, John Shepherd, Lambert Leong, Thomas Wolfgruber, Li-Yung Lui, Peter Sadowski, and Steven R Cummings. Deep learning predicts all-cause mortality from longitudinal total-body dxa imaging. Communications medicine, 2(1):102, 2022
- Yannik Glaser, Ralph C., Foster, Douglas C., Vandemark, Linnea, Wolniewicz, Peter, Sadowski, Justin E., Stopa. Wvnet: A sar wave-mode foundation model. AGU Fall Meeting Abstracts. 2023.

- Yannik Glaser, Justin Edward Stopa, Linnea M. Wolniewicz, Ralph Foster, Doug Vandermark, Alexis Aur´elien Mouche, Bertrand Chapron, and Peter Sadowski. WV-Net: A foundation model for sar ocean satellite imagery. AMS Artificial Intelligence for the Earth Systems. 2025. (Under review)
- Yannik Glaser, Thomas Wolfgruber, Arianna Bunnell, Peter Sadowski, John Shepherd. A DXA body composition foundation model. (In preparation)

1.6 Dissertation Outline

The dissertation will be structured as follows. Chapter 2 will provide information and background on the current state of self-supervised learning, introducing popular frameworks and different fundamental approaches. This chapter will also introduce transfer learning, the concept of a domain gap, and its relevancy to applying self-supervised learning to non-NI imagery. Finally, Chapter 2 will introduce the concept of unsupervised representation quality metrics.

Chapter 3 will be a slightly altered version of a work that is under review for publication at the time of the dissertation defense. This chapter is focused on using self-supervised learning to build a SAR foundation model and downstream applications of said foundation model.

Chapter 4 will repeat the structure of Chapter 3 but will use self-supervised learning for DXA medical imaging data and downstream applications in that modality.

Chapter 5 will analyze embedding quality metrics on the models trained in the previous two chapters. It will evaluate the metrics for robustness to downstream tasks and application domain.

The final chapter describes the conclusions of this dissertation and future work.

CHAPTER 2
BACKGROUND

2.1 Domain data

While in computer vision, most of the research attention and effort is often devoted to a limited number of applications, there are countless underutilized, promising datasets available in scientific vision domains. Two such datasets are SAR S-1 wave mode (WV) images from remote sensing and DXA scans from medical imaging. Both are from domains that are popular for applied computer vision research, but both WV mode and DXA scans have gone underutilized in favor of more popular datasets. However, both hold potential for many downstream applications and present interesting challenges, especially in the context of transfer learning.

2.1.1 Synthetic aperture radar (SAR) S-1 wave mode

Synthetic Aperture Radar (SAR) is an active microwave remote sensing technology. Unlike passive optical sensors that rely on ambient light, SAR systems transmit their own microwave signals and record the reflected energy, or backscatter, from the Earth's surface. This self-illumination allows SAR to acquire high-resolution imagery, disregarding cloud cover, smoke, or darkness. The all-weather, day-and-night imaging capability of SAR makes it a primary instrument for systematic environmental monitoring and disaster management.

The Sentinel-1 (S-1) mission consists of multiple polar-orbiting satellites (S-1A launched in 2014, S-1B in operation between 2016 and 2022, and S-1C, launched in 2024) equipped with 5.405GHz C-band SAR instruments. The satellites collect imagery with four different acquisition modes: Interferometric Wide Swath (IW), Extra Wide Swath (EW), Stripmap (SM), and Wave (WV) all of which serve different observational requirements. The primary mode over land is the Interferometric Wide (IW) swath mode, which offers 250 km swath coverage at moderate resolution, making it suitable for most terrestrial applications.

Over the open ocean, WV is used exclusively. This mode collects global sea surface roughness data regardless of cloud cover and time of day, creating a large dataset of SAR WV images. These images, about 60,000 per month per satellite, are 20×20 km in size with 5 m resolution, are taken every 100 km along the orbit alternating between two incidence angles of 23.8°and 36.8° [45]) and can show various features beyond just sea surface state [37].

2.1.2 Dual-energy X-ray absorptiometry (DXA)

DXA is a non-invasive, low-X-ray-dose imaging modality widely utilized for bone density and quality assessment in osteoporosis [46], but also serves as the criterion measure for quantifying body

composition [41, 47]. This imaging technique relies on the ability to differentiate body tissues based on the differential attenuation of two distinct low-dose X-ray beams, enabling a three-compartment model analysis (bone, fat, lean mass). Specifically, for scans utilized in this work, taken by various Hologic (Hologic Inc., Marlborough, MA) scanners (Hologic QDR Series, Hologic Horizon A), the process involves capturing images using alternating low- and high-energy X-rays, filtered through air, tissue, and bone, resulting in a six-channel, 16-bit image. While some previous work has focused on the ratio (R) of attenuation at low energy to that at high energy as a key parameter, exploiting its relationship with the chemical composition of the material being imaged [17], this study will primarily focus on the high- and low-energy channels. Clinical software can be used to derive body composition variables from the attenuation values in the DXA images.

The primary clinical application of DXA is in the evaluation of bone health, specifically for the diagnosis of osteopenia and osteoporosis. Osteoporosis is a systemic skeletal disease characterized by low bone mass and microarchitectural deterioration of bone tissue, leading to an increased risk of fracture. DXA provides a quantitative measure of bone mineral density, typically at the lumbar spine and proximal femur, which are common sites for osteoporotic fractures [46]. Beyond diagnosis, DXA is crucial for fracture risk assessment. Low BMD is a strong, independent predictor of future fractures [48]. The results from a DXA scan are often integrated into fracture risk assessment tools, such as the Fracture Risk Assessment Tool (FRAX) [1], which combines BMD with other clinical risk factors (e.g., age, sex, prior fracture history) to estimate the 10-year probability of a major osteoporotic fracture. Both proximal femur and spine DXA scans can also be used to directly predict fracture risk [40, 49] and even detect fractures [39].

While initially developed for bone densitometry, the application of DXA in body composition analysis has become equally significant. Unlike simpler two-compartment models (fat mass and fat-free mass) used by methods like hydrostatic weighing, DXA provides a more detailed three-compartment assessment. A total body DXA scan precisely quantifies total and regional fat mass, lean soft tissue mass, and bone mineral content. Additionally, whole-body DXA scans, capturing varied body composition information, have been demonstrated to also contain all-cause mortality risk information [50] and have even been explored to derive markers of biological age based on body composition [51]. Because DXA is a relatively low-radiation, low-cost imaging method that is already commonly applied in multiple contexts, it has great potential for opportunistic screening as well [52].

2.2 Self-supervised representation learning

Self-supervised representation learning emerges in various areas of the deep learning literature. One early approach is generative, using an auto-encoding objective to pre-train individual layers

[1] https://frax.shef.ac.uk/FRAX/

of a deep belief network in a greedy manner [53], replacing restricted Boltzmann machines used for the same purpose [54]. Deep end-to-end auto-encoders (AEs) became a more viable strategy as training recipes improved with works such as denoising AEs [55, 56], variational AEs [57], deep canonically correlated AEs (DCCAE) [58] (which bear resemblance to more modern self-supervised learners, taking advantage of multi-view data), and split-brain AEs [59] which do cross-channel predictions between color channels of input images. This, along with denoising AEs, is related to other approaches with objectives aimed around restoring intentionally removed information. Examples include converting an image to grayscale before restoring original pixel RGB values [60, 61] or context encoders [62] which restore masked sections of an input image, a precursor to the modern MIM objective. The original paper introducing vision transformers (ViTs) also presented preliminary experiments on a self-supervised objective, masking image patches with *mask tokens* and using a decoder to reconstruct the masked patches from context [63].

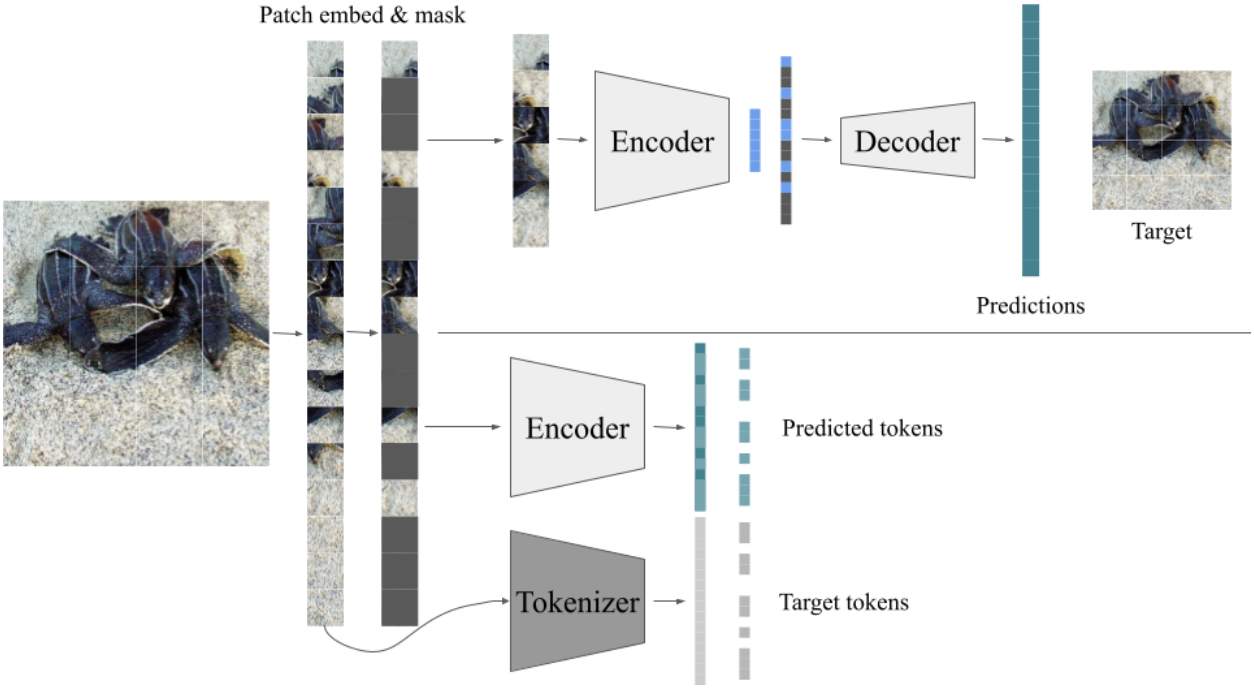


Figure 2.1: Overview and comparison of current MIM frameworks. The top panel shows a masked autoencoder with a pixel-level reconstruction objective, and the bottom panel shows BEiT with a token-level objective.

The masked autoencoder (MAE)[4] fleshed out this concept, achieving competitive performance with other self-supervised representation learning frameworks. Importantly, due to a lightweight decoder for pre-training and forgoing encoding the masked 2.1 patches, this framework is especially computationally efficient and highly scalable. BEiT [64] adapted the reconstruction objective to include a discrete VAE tokenizer [6] to encode image patches and changed the objective to recon-

struct tokens instead of pixel values to be more in line with the successful BERT [1] framework for natural language processing. In practice, the MAE framework, being similar in performance and much more efficient, is usually chosen over BEiT.

Other methods take advantage of closely correlated data, similar to DCCAEs. In the absence of multi-view data entirely, subsequent frames in videos offer similarly inherently correlated images. [65] for instance, take advantage of this by predicting the motion of the camera between subsequent frames. This idea of multi-view invariant representations as a training objective is central to modern JESSL frameworks as well, originating with general Siamese networks initially used for signature verification [66]. The idea of a contrastive loss between Siamese arms is then formalized in [67, 68]. The first work to not use samples explicitly labeled as the same class but instead derive the notion of *positive pairs* from subsequent video frames while contrasting with *negative pairs* from frames of unrelated videos in a triplet loss [69]. The N-pair loss introduced by [70] generalizes the triplet notion to arbitrarily many negatives and is taken to the extreme by [71]. In this work, each sample is effectively treated as its own class, the positive pair being constituted by current and past representations (tracked by a memory bank) of any given sample, while every other image serves as a negative. This work does not rely on any class information or additional information about the data (such as two frames appearing in the same video), working on completely unlabeled image datasets.

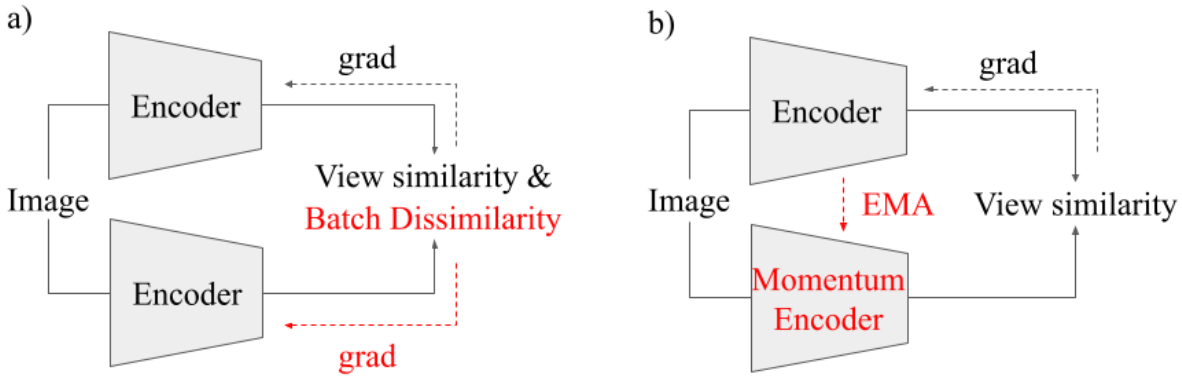


Figure 2.2: Illustration of contrastive and non-contrastive JESSL frameworks, with differences highlighted in red based on [72]. Panel a) shows a simplified contrastive framework similar to SimCLR [3] where gradients are calculated based on view similarity and similarity to all other images in a batch. Panel b) shows a non-contrastive framework, similar to BYOL [44] where one encoder arm is an exponential moving weight average of the other arm, and loss is only calculated based on positives' similarity.

Finally, the framework proposed in [73] is the first modern incarnation of JESSL, formally introducing the notion of creating multiple *views* of the same image through data augmentation. Multiple contrastive frameworks followed, refining this idea. SimCLR [3], a simplified version of which is depicted in Figure 2.2, iterated on the framework from [73] by removing comparisons

between intermediate feature maps in the encoder network, introducing a projection head between the encoder and the final output, and doing a more extensive data augmentation study, popularizing the set of augmentations used by most NI JESSL frameworks since. Momentum contrastive learning (MoCo) [74] introduced a memory bank to artificially increase the effective batch size, alleviating training cost due to contrastive learning's reliance on large batches to prevent representational collapse. [75] take advantage of a memory bank while simplifying the update rule to shifting each embedding toward the mean of its closest neighbors retrieved from the memory bank. Instead of multi-view data, CLIP [7] takes advantage of language-image pairs, training a separate text and image encoder network using a contrastive objective. SimCLR may be the most widely used framework in practice due to its simplicity and adaptability.

A second branch of non-contrastive multi-view invariance methods emerged first in the form of clustering approaches such as [76, 77]. These frameworks employ a two-step training procedure, first embedding images and assigning clusters and then predicting cluster assignments and updating model weights, with [77] using the Sinkhorn-Knopp algorithm to update the cluster-assignments and avoid collapse, an approach also used by SwAV [78] who assign embeddings to predetermined prototype vectors that become the target for Siamese prediction from multiple views. Bootstrap your own latent (BYOL) [44] is the first framework to forego the clustering step in favor of an asymmetric Siamese structure, shown in Figure 2.2. Here, both arms consist of the main vision backbone and an MLP projector; one arm, the *student* also has a smaller MLP predictor network, while the other arm, the *teacher* stops at the projector. For two views generated from an original image through augmentation, the role of the student, given one view, is to predict the continuous output of the teacher's projector for the other view, both ℓ_2-normalized. The student is then updated based on its mean-squared error using gradient descent while the teacher's weights are an exponential moving average of the student's weights, which the authors refer to as self-distillation. This relies solely on positive views, making this family of frameworks non-contrastive. Various follow-up works propose other self-distillation approaches, mostly differing in how they avoid representational collapse. DINO [18] avoids projection and prediction heads by centering and sharpening the teacher outputs, while SimSiam [72] aims to further simplify the framework to the minimal components necessary to avoid collapse, landing on a small predictor head for one Siamese arm and employing a stop-grad to only update the main encoder's weights based on a cosine similarity loss incurred by that arm. This idea of reducing self-supervised learners to essential components is taken to the extreme in [29], primarily for the purposes of theoretical study. Instead of self-distillation, the last class of JESSL frameworks is based around objectives targeting the covariance matrices of embeddings of related data. [79] argue this family of frameworks originates with the canonical correlation view introduced by [80] and can certainly be traced back to the DCCAE framework [58]. Barlow Twins [81] proposes a simple loss function pushing the cross-correlation matrix between the embeddings of two views toward the identity matrix, penalizing large off-diagonal terms while

encouraging diagonal values close to 1. VICReg [26] builds on this loss by keeping the off-diagonal regularization of the cross-correlation matrix while additionally regularizing variance between the dimensions of each embedding to prevent collapse and minimizing distance between embeddings of different views instead of explicitly maximizing the diagonal of the cross-correlation matrix.

Lastly, more recent works have begun combining the MIM and JESSL objectives. iBOT [9] train ViTs using a combination of the BEiT objective, reconstructing masked image patch representations, and a self-distillation objective, minimizing the cross-entropy between the `[CLS]` embeddings of the teacher and student arm. DINOv2 [5] refine this training recipe by systematically including components drawn from several other works, careful hyperparameter tuning, and pre-training on a novel dataset of 142M images. This avenue of research is promising but extremely computationally and data intensive.

While this is an overview of various frameworks for self-supervised representation learning in computer vision, it is by no means comprehensive. For practitioners, that presents the practical problem of choosing which framework to use. Fortunately, several works have attempted to introduce unifying perspectives to understand the different families of frameworks, with results broadly indicating that, under careful hyperparameter tuning, learned representations should be equivalent within some margin. [82, 83, 84] all show, from different angles, performance gaps between frameworks are not due to the inherent potential of the frameworks. Other works indirectly support this assertion, such as [85] linking self-distillation frameworks like BYOL with the Barlow Twins objective, or [86] linking the Barlow Twins objective to an upper bound on contrastive losses.

2.3 Transfer Learning

Transfer learning has long been a central technique in computer vision — essentially reusing weights trained on a particular task to solve related tasks or problems in related domains. This approach drastically reduces the need for extensive labeled data and computational resources, making it possible to develop performant models for specialized domains with data limitations. The standard practice involves fine-tuning models pre-trained on large, general-purpose datasets like ImageNet[10], perhaps the most popular pre-training dataset used in computer vision. The models resulting from pre-training on ImageNet usually display strong transfer performance [87] leading to it having been used to produce then-state-of-the-art results in various domains, such as human pose estimation [88], general pose estimation [89], skin disease diagnosis [90], and satellite images analysis [91].

However, as the domains get more dissimilar to the original NIs used for pre-training, transfer performance can suffer [92]. The concept of dissimilarity is known as the domain gap and has been characterized in relation to its impact on transfer performance by [92] (see Figure 2.3).

This is especially relevant for self-supervised learning, as pre-training efforts have moved from weakly supervised [13] to self-supervised [14]. In fact, of the previously discussed self-supervised frameworks, all of them are trained and evaluated exclusively on NI datasets in the original papers.

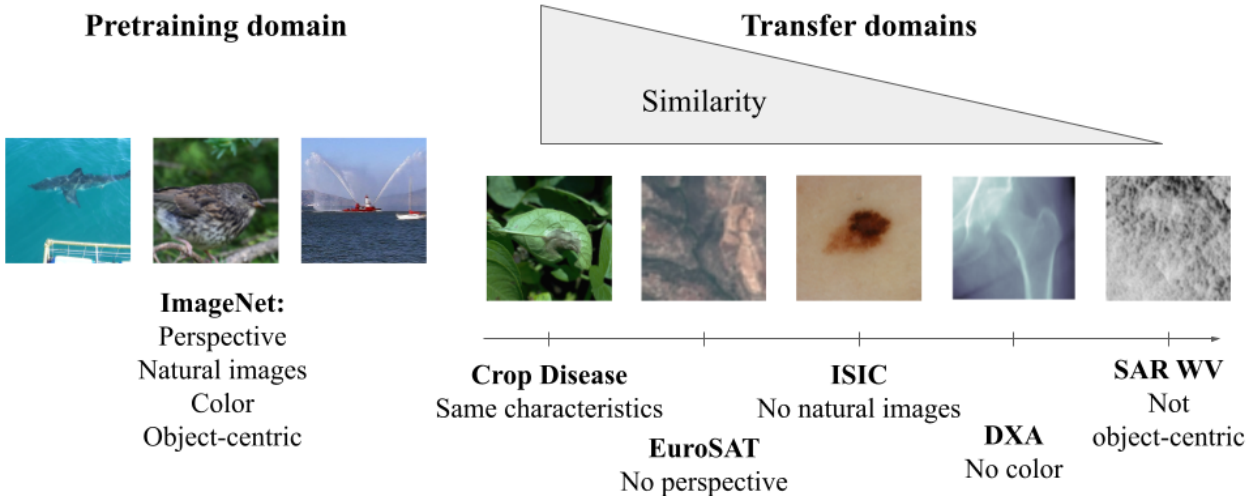


Figure 2.3: Illustration of domain gap concept from [92].

Worryingly, in the suite of common datasets used in benchmarking self-supervised representation learning frameworks, the only dataset representative of non-NI data is the EuroSat dataset [93], appearing in some works doing extensive experiments on framework characteristics [79, 28].

This has obvious implications for the applicability of these frameworks to new domains. Not only do representations learned on in-domain data perform better on transfer tasks [15], but vision backbones and framework hyperparameters are demonstrably overfit to ImageNet [94, 95] and other NI datasets [16] and are ideally set on a case-by-case basis for each new application [29]. Additionally, domain-specific data may come with characteristics that can be exploited to improve representation learning.

For these reasons, many works that are pre-training models in a self-supervised manner for specialized image domains have emerged. For instance in medical imaging, [96] and [97] use an unchanged SimCLR framework but alter pre-training to a multistep process, first initializing with NI-pre-trained weights, then pre-training on as much data in the target domain as possible, taking advantage of self-supervision, and finally finetuning on task-specific labeled data, showing better final performance and robustness across various medical imaging tasks. Several works utilize a CLIP objective to pre-train multimodal models on images and radiology reports (such as [98] or [99]) or image-text pairs scraped from scientific articles like BioMedCLIP [100]. [101] offer a comprehensive review of work in this area. Some work also focuses on DXA, such as [102] using a contrastive loss for cross-modal alignment and automated registration. [103] include a self-supervised step between previously segmented bones of the hand in the training, interestingly, to compensate for the small available sample size in the study. However, it appears that there is no general-purpose DXA pre-trained model available that is shown to transfer well to multiple downstream tasks.

There have been similar efforts for satellite imagery, with MAEs being a popular choice. Sat-

MAE [104] tailors masking strategies to multi-spectral characteristics of the data while Scale-MAE [105] adds scale information to the positional encoding, indicating the area covered by an image, and includes a more elaborate decoding procedure to recover large- and small-scale features. The NENYA framework [106] pre-trains on MODerate-resolution Imaging Spectroradiometer (MODIS) data using the SimCLR framework with an added rotation augmentation and finetune the model to predict sea surface temperature at a global scale. [107] give an overview of all self-supervised efforts utilizing SAR data; notably, WV mode is almost completely absent except for this present effort.

While self-supervised methods already have a track record of success beyond NI data, efforts to adapt these methods are generally disjointed. Many strategies for adapting self-supervised frameworks have emerged in the literature, including exploiting known data invariances to find views [108, 109], changing augmentation policies to fit the domain [110], feature engineering [105], creating losses based on known properties of the data [111, 112], or changing the underlying architecture of the model [104]. There is no agreed-upon strategy or systematic approach for discerning which adaptation will be beneficial for what task, and exploring the options involves expensive pre-training and model tuning tailored specifically to the data domain using domain and machine learning expertise. Reducing the cost of this process is critical to enabling more work to be done in this area with greater efficiency.

2.4 Representation quality metrics

Given the need for efficient means of evaluating learned embedding quality, several works have emerged proposing metrics for this purpose. Alignment and Uniformity [113] rely on measuring the distribution of embedding across the embedding space (uniformity) and closeness of images from the same class (alignment) to characterize what constitutes a good model. Optimizing for both of these metrics in tandem is shown to lead to good performance in the paper; however, evaluating alignment relies on the availability of labeled data or augmentations for generating views that are well-aligned with the downstream tasks, neither of which are feasible for general-purpose, unsupervised evaluation. Another metric measuring distribution qualities of a representation space is IsoScore [114], proposing a robust measure of isotropy for high-dimensional embedding spaces. The authors point out that this is the only metric that demonstrably measures how uniformly variance is distributed across an embedding space. Results in the paper also indicate that isotropy might not be strictly desirable for good transfer performance, which seems to contradict the previously discussed results suggesting uniformity is desirable. IsoScore is however, only evaluated on NLP applications. Instead of focusing on the distribution of the embedding space, RankMe is an effective rank measure proposed to estimate how much information is captured by an embedding model [22]. The authors show that their effective rank measure is well-correlated with downstream task performance on various tasks. A last perspective of embedding quality comes from analyzing

eigenvalues of the representation matrices. [115] are the first to point out power law behavior in the decay of the eigenvalues, linking the rate of decay as the link between collapsed or whitened representations, common issues in SSL. This is expanded on in subsequent work [116] by analyzing linear regression transfer performance. The authors find that an eigenspectrum decay parameterized by exponent $\alpha \sim 1$ leads to better, more robust transfer performance. α-ReQ [23] is a score designed based on this observation, shown to be well-correlated with downstream performance. A comprehensive review of these measure with some experiments, limited to NIs, is presented in [24] showing mixed performance for all metrics while introducing several new potential performance indicators. Little work has been done thus far to test these scores on non-NI data, regression or multi-label classification problems, or even representations derived from MIM objectives.

CHAPTER 3
WV-NET: A FOUNDATION MODEL FOR SAR OCEAN SATELLITE IMAGERY

This chapter is largely identical to a journal article under review for publication at the time of the dissertation defense.

3.1 Introduction

Machine learning is becoming increasingly important for analyzing remote sensing data. The number of Earth observation satellites in orbit has grown from 150 in 2008 [117] to over 1150 in 2022 [118]. Missions like the European Space Agency's (ESA) Sentinel-1 (S-1) mission generate large amounts of high-resolution images with global coverage. ESA has taken an open-data policy making high-resolution synthetic aperture radar (SAR) imagery readily available for applications ranging from environmental monitoring to climate modeling [119]. Fully leveraging the torrent of S-1 SAR imagery requires automated analysis tools with many potential applications for machine learning [e.g. 120, 121]. However, the machine learning approach generally requires large datasets of training images that have been annotated by experts.

Transfer learning is one common solution to this challenge. A deep neural network model is first pre-trained on a large dataset from a related domain and then fine-tuned on the target task, requiring significantly less labeled data than would be necessary when training from randomly initialized network parameters. The pre-trained model is called a *foundation model* because it can be reused for multiple downstream tasks. Foundation models pre-trained to classify NIs (primarily the ImageNet dataset [10]) are routinely fine-tuned for remote sensing [122, 123, 124]. However, transferring a model from a NI classification task to a remote sensing task can be problematic because the image characteristics are very different. [92], among other works, formalize this difference based on image characteristics like perspective, object-centric-ness, presence of color, and whether the images are of natural objects, concluding that EuroSat imagery has limited similarity to ImageNet, making transfer learning more difficult. This is known as the domain gap, and the deep learning literature has repeatedly shown that a wide domain gap between pre-training and target data domains can hinder transfer learning performance [125, 126, 127]. [128] even go so far as to argue that satellite imagery should be treated as an entirely separate modality from NIs due to its distinct qualities such as unique spatial and temporal scales, the ability to capture multispectral imagery beyond standard RBG channels, and the extremely sparse annotation-to-data-volume ratio often associated with satellite data.

Self-supervised learning (SSL) provides an approach to pre-training a foundation model on unannotated, domain-specific data. Instead of predicting annotations in a supervised manner, SSL

algorithms define some other *pretext task* for pre-training. This approach has long been utilized in natural language processing [129, 130] and has been one of the driving factors for the success of large language models [1, 131, 2]. Recently, contrastive learning has re-emerged as a successful self-supervised form of pre-training, especially for computer vision [3, 18, 9]. Contrastive algorithms have produced impressive results on NI datasets, resulting in general-purpose models that perform on par or better than models being trained in a supervised manner from scratch on the target dataset [18, 3]. Thus, SSL presents opportunities for analyzing remote sensing data. Recent studies have shown that pre-training on remote sensing data instead of NIs yields superior performance on downstream tasks [132], with most proposed methods being self-supervised [133, 104, 105, 134]. To date, these efforts have focused on remote sensing imagery of landmasses or coastal regions.

The objective of this work is to build the first foundation model for open-ocean sea surface images. Our foundation model is pre-trained on imagery from S-1 WaVe (WV) mode, which was designed to capture ocean waves at 5 m resolution in 20 $\times$ 20 km footprints [135, 136, 137]. These images capture various ocean phenomena [138] and have global coverage, with millions of images archived over the last decade. Thus, the data has been used to study ocean fronts [139], air-sea interactions including organized turbulence in the marine boundary layer [140, 141, 142], and other physical, atmospheric, and biological processes [37]. By building a foundation model specific to SAR WV images, we can more fully utilize the large existing database of WV imagery, accelerating research related to the ocean and atmosphere.

Two hypotheses are tested. First, we test whether contrastive SSL can train a SAR WV-mode foundation model that outperforms standard foundation models pre-trained on NIs. Second, we test whether performance of the model can be improved by using domain knowledge to design data augmentations. These hypotheses are tested experimentally using a dataset of almost 10 million S-1 WV images, along with three smaller subsets of annotated images that exemplify target supervised tasks for transfer learning. This dissertation will serve to introduce the model and demonstrate its robustness and superior transfer performance on a set of diverse downstream applications. While in-depth analysis of the model's application to these downstream tasks is beyond the scope of this dissertation, follow-up work is being prepared that more closely explores individual applications. The optimized foundation model is made publicly available under the name WV-Net, and we expect it to be useful for various downstream applications such as studying air-sea interactions, improving constraints on numerical weather predictions, and monitoring sea ice.

3.2 Methods

3.2.1 Datasets

Sentinel-1 launched two satellites, S-1 A and B, in April 2014 and 2016 respectively [31]. A third, S-1 C, was launched in December of 2024. S-1B went out of commission in December 2021. The

S-1 satellites are identical polar-orbiting, sun-synchronous satellites [119]. S-1 operates in the C-band SAR with a center frequency of 5.405 GHz or a wavelength of 5.5 cm. S-1 has a 12-day repeat cycle, flies at an altitude of 690 km, has an inclination of 98.2°, and a repeat period of 98.7 minutes. When both S-1A and S-1B were in operation they were 180° out of phase equating to a 6-day repeat cycle.

Each satellite produces approximately 60,000 images per month. S-1A and went into routine acquisition mode in October 2015 and July 2016, respectively, so in total, there are approximately 165 months and 9.9M S-1A/B images. The WV images are 20 × 20 km scenes and alternate between incidence angles of 23.8° (WV1) and 36.8° (WV2). The along-track separation is 100 km with 5 m pixel spacing. S-1 uses both vertical-vertical (VV) and horizontal-horizontal (HH) polarization, but only one polarization can be obtained for one image. The majority of the WV archive is in VV.

The S-1 *GeoTIFFs* are saved in range-azimuth (cross and along track) coordinates. The images in this study have the North direction facing upwards; therefore, the descending passes are flipped in the range and azimuth directions to make their relative geophysical representation the same. The raw 20-km WV images have 4000 to 5000 pixels in the range and azimuth directions. We implement a similar strategy to [37] by reducing the raw data size while highlighting the geophysical phenomena that influence the sea surface roughness. The scales resolved by this processing are larger than the typical azimuth cutoff of 100 m Stopa et al. [143] and extend to 5 km in three steps: 1) incidence normalization, 2) downscaling, and 3) intensity normalization.

1. *Incidence Normalization:* The radar backscatter (σ_0) depends on the local surface wind, incidence angle (ϕ), relative wind-platform angle (θ), and polarization. The model CMOD5N of Hersbach [144] is used to ameliorate these effects by assuming a constant wind speed of 10 ms^{-1} and a relative wind-platform angle (ϕ) of 45° to estimate the sea surface roughness (SSR) as

$$SSR = \frac{\sigma_0}{CMOD5N\,(10\ ms^{-1}, \theta, \phi = 45°, VV)}. \tag{3.1}$$

2. *Downscaling:* A moving boxcar window of 10 × 10 pixels or 50 m is applied to the SSR data. Every 10th pixel is then selected to reduce the data by a factor of 100, resulting in an image size of 400 to 500 pixels in both range and azimuth.

3. *Intensity Normalization:* The image intensity is enhanced by normalizing each image against data within the 1st ($P01$) and 99th ($P99$) percentile

$$SSR_n = 255\left(\frac{SSR - P01}{P99 - P01}\right). \tag{3.2}$$

This alters the SSR to lie within the interval [0,255], where values of $SSR \leq P01 = 0$ and $SSR \geq= 255$. This normal imagery is then an unsigned 8-bit integer, and the matrix is

saved as a portable network graphics (*PNG*) file. Thus the dataset is effectively composed of grayscale images.

The images contain features of interest at multiple spatial scales (Figure 3.1), sometimes in the same image. Below we describe three subsets of the data that have been annotated for supervised learning tasks; these tasks are used to evaluate WV-Net as a foundation model.

GOALI classification dataset: The GOALI dataset consists of 10,000 WV images that were manually annotated by human experts for multilabel classification, meaning any image can have multiple labels simultaneously (Stopa et al. 2025, manuscript in preparation). The labels indicate geophysical phenomena observable in the image. To this we add 6,400 images from Wang et al. [37] that have been re-annotated in a way consistent with GOALI, for a total of annotated 16,400 images. The GOALI images are multilabeled with the following phenomena: wind streaks (WS), micro-scale convective cells (MC), negligible atmospheric variability (NV), rain cells (RC), cold pools (CP), sub-mesoscale air-mass boundaries (AB), low wind areas (LW), atmospheric gravity waves (AW), biological slicks (BS), ocean fronts (OF), internal oceanic waves (IW), icebergs (IB), ships (SH), ship wakes (SW), and other unidentified phenomena (UD). Sample images are shown in Figure 3.1. This dataset is currently unpublished work but will be made publicly available in the future.

Wave height regression dataset: Quach et al. [137] created a dataset of hundreds of thousands of WV images with significant wave height (H_s) by colocating S-1 satellites with altimeter satellites. Here we use a subset of 200,000 images and randomly split the data into sets of 50,000, 50,000, and 100,000 for training, validation, and final evaluation, respectively.

Air temperature regression dataset: Stopa et al. [142] showed that the sea surface roughness observed in SAR is related to atmospheric stratification and therefore air-sea temperature differences. Using ERA5 reanalysis data [145] as ground truth annotations for the sea surface temperature (SST) and air temperature (T_{v10}), we attempt to predict the difference from SAR images. The air temperature is corrected using the COARE algorithm [146] to account for moisture content and derived as a virtual air temperature at a height of 10 m above the sea surface (TV10). The annotated dataset consists of 76,000 images, which are split into 50,000 training, 11,000 validation, and 15,000 testing images.

3.2.2 Model details

WV-Net is primarily trained using the SimCLR contrastive SSL framework [3] with a standard, fully-convolutional ResNet50 as the backend architecture [147]. The SimCLR SSL pretext task is

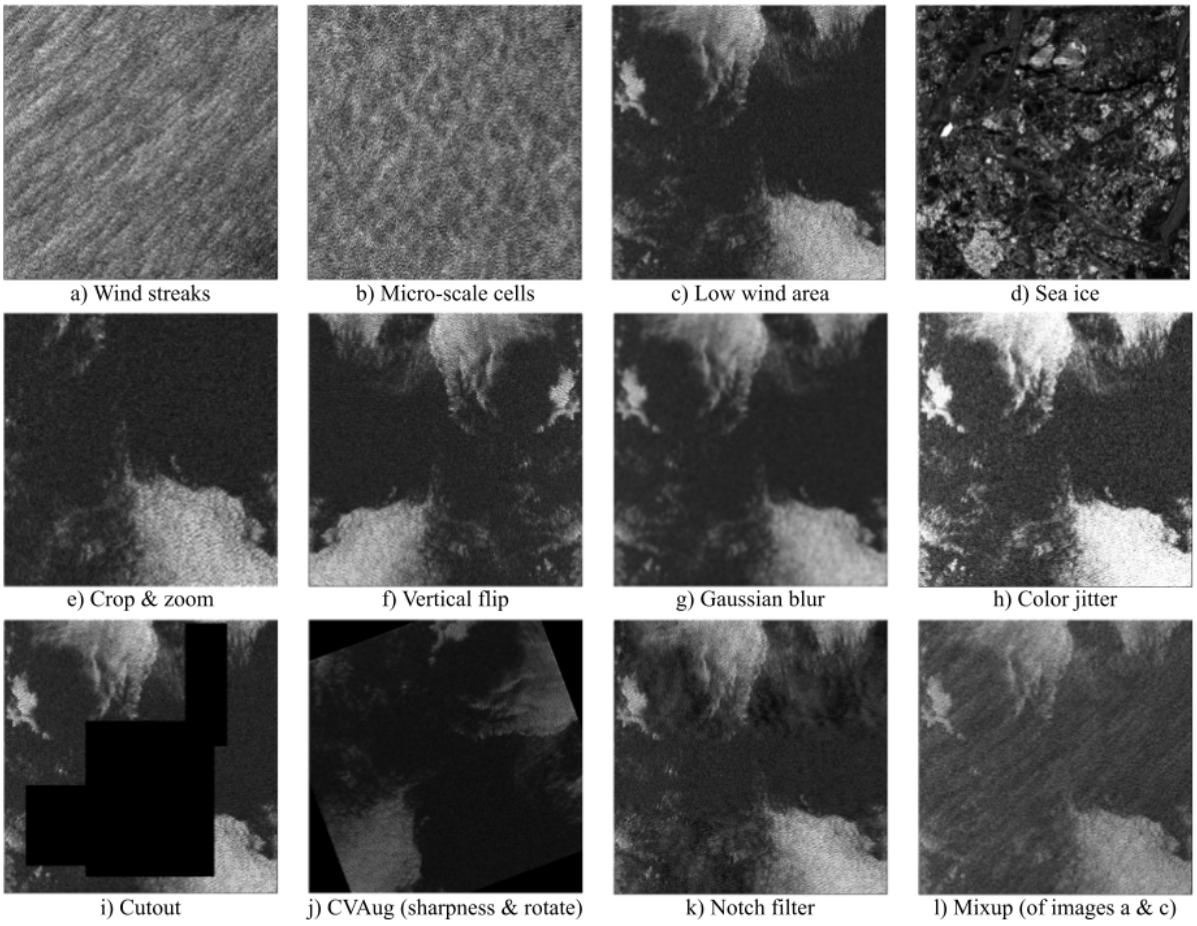


Figure 3.1: (a–d): Sample images of different geophysical phenomena observable in the global S-1 WV archive, titled by their dominant classes. Multiple classes can be present in the same image. (e–h): Augmented versions of the low wind image illustrating the default SimCLR augmentation policies. (i–l): Augmented low wind images illustrating the augmentation policies evaluated in this work. In the actual SimCLR framework, usually multiple augmentations are applied in sequence to the same image.

to learn similar representations for two augmented *views* of an image while discouraging similarity with the representations of any other image in the training data (Figure 3.2).

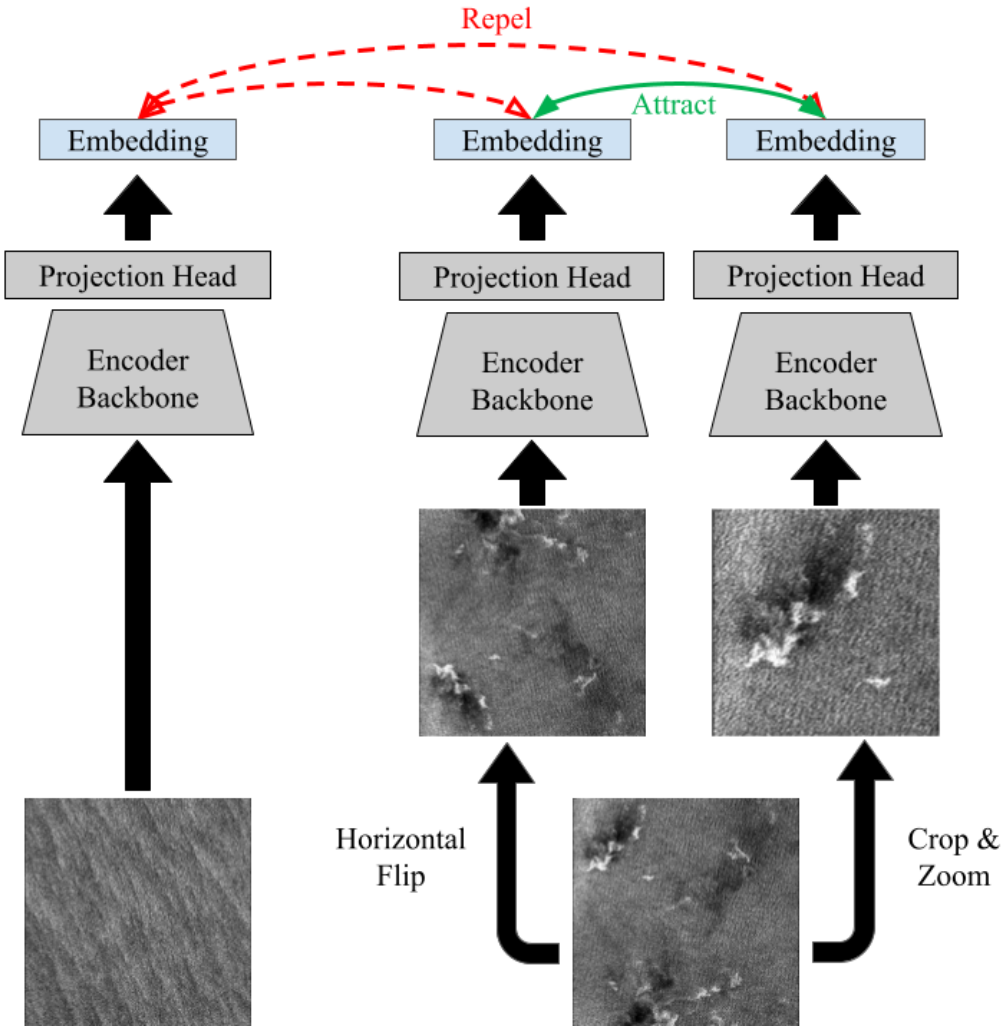


Figure 3.2: In the SimCLR algorithm, images are randomly augmented to create several views of the same image. An encoder network — consisting of a backbone and a smaller projection head — learns to produce an embedding that is similar to embedded views from the same original image and dissimilar to embedded views from all other images. Only the encoder backbone is used for transfer learning.

A SimCLR training step begins by randomly sampling a mini-batch of N training images. Each image $\mathbf{x}_k$ is transformed twice by random sequences of augmentation policies (sampled from a pool of transformations) to produce two views of the original image, $\tilde{\mathbf{x}}_{2k-1}$ and $\tilde{\mathbf{x}}_{2k}$, resulting in 2N total images. Each view is encoded by a backend network (here a ResNet50) and then a smaller *projector* neural network, resulting in the embedding vectors $\mathbf{z}_{2k-1}$ and $\mathbf{z}_{2k}$. The loss for any *positive pair* of

embeddings $\mathbf{z}_i$ and $\mathbf{z}_j$ originating from the same image is:

$$l_{i,j} = -\log \frac{\exp(sim(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{I}_{[k \neq i]} \exp(sim(\mathbf{z}_i, \mathbf{z}_k)/\tau)} \tag{3.3}$$

where τ is a temperature scalar, $\mathbb{I}$ is the indicator function mapping the positive sample pair to 0 to prevent it from adding to the denominator of the loss and $sim(\cdot, \cdot)$ is the cosine similarity:

$$sim(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u}^T \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|} \tag{3.4}$$

The final loss is the sum of losses $l_{i,j}$ for all positive pairs in a batch where the goal is to maximize the positive pairs' similarity while minimizing the similarity to the rest of the batch. Unless otherwise specified, all hyperparameters are adopted from the original SimCLR work [3] with linear learning rate scaling. Also in accordance with the original SimCLR formulation, the projection head is a two-layer multilayer perceptron (MLP) with a 2,048-unit hidden layer and 128-unit linear output size.

3.2.3 Augmentations

SAR WV mode images are very different from NIs, so experiments were conducted to optimize the choice of augmentations used to train WV-Net. In addition to the augmentations proposed in the original SimCLR [3] paper, we explore various augmentations proposed in the contrastive learning literature, transformations from traditional computer vision, and a transform from signal processing that was inspired by the SAR imaging process.

Cutout is a geometric transform where one or multiple rectangles within the image are zeroed out or replaced with Gaussian noise [148] (see Figure 3.1i). We expect that including the cutout augmentation could, on one hand, replicate some driving factors behind the success of masked-image-modeling in geospatial data [4, 104], and also force the model to pay attention to all areas of the images despite the majority of the dataset being comprised of homogeneous textures across the 20-km frames. In this study, cutout can be applied up to three times, each application having a probability of $p = 0.5$, each with a random scale between 2% and 30% of the image and a random aspect ratio between 0.3 and 3.33. The areas may overlap and are zeroed out.

CVAug (or computer vision augmentations) is a composite of classical computer vision transforms that, intuitively, may complement learning on remote sensing data. Random color inversion is included to encourage a level of invariance to the pixel-intensity information and highlight textures. Random rotation is included because the model should be rotationally invariant. A sharpness transform is included to obscure or highlight texture features, incentivizing more balanced repre-

sentations that do not over-rely on global textures. Each augmentation is applied with probability $p = 0.5$, the sharpness being increased by a factor of 0.5 and the rotation between $\pm 170°$ (see Figure 3.1j).

Notch filtering or stopband filters are common in signal processing, typically applied to reduce noise. Khan et al. [149] use the term *spectral dropout* to describe dropping weak Fourier coefficients from a layer's input distribution. Here, random dominant Fourier features from the raw input vector are dropped instead. Since ocean waves with scales of 50 to 800 m visually dominate most of the images, this augmentation should force the network to consider less dominant features. The notch filter is applied with probability $p = 0.5$ and zeroes out up to 15 of the first 30 Fourier features obtained by doing a 2D Fourier transform over the image. However, the most dominant frequency, the first Fourier component, is excluded (see Figure 3.1k).

Mixup was first proposed as a data augmentation for supervised learning [150]. It creates new training examples by taking a weighted combination of random feature vectors and their labels. Mixup has found popularity in SSL by only combining feature vectors [151, 152, 153]. Further work framed mixup in terms of other noise-injection methods such as adding Gaussian noise to images, showing that it improves over random noise masks because the corrupted example is closer to the data manifold [154]. Mixup is applied with probability $p = 0.5$ and a random mixup strength, m, between 0.1 and 0.4. Explicitly, the augmented image, C, created by mixup is written $C = (1 - m)A + mB$ where A is the original image and B is another, randomly sampled, image from the same batch (see Figure 3.1l).

No-zoom crop modifies the crop-and-zoom augmentation that is universal among multi-view contrastive learning frameworks to create a *no-zoom crop* policy that focuses on random cropping with only minimal scaling. Since WV images are captured from a satellite in constant orbit and have a consistent 20 km footprint, phenomena captured do not vary in scale as much as features in NIs might. Thus, by reducing the zoom component, scale invariance is not as heavily incentivized in the model, allowing features to be more specific.

This means the augmentations broadly fall into the following categories:

- **SimCLR augmentations:** These include random cropping and zooming, random flipping, random color jitter, and random Gaussian blur (see Figure 3.1e-h for examples) and are from the original SimCLR formulation.
- **Literature-inspired augmentations:** *Mixup* [150] and *Cutout* [148] which have been shown to work well in contrastive learning frameworks [155, 3, 154].
- **Computer vision augmentations:** Traditional image processing transformations that seem well-suited for this application. We combined random rotation, random color inversion, and

random sharpness transformations into a single augmentation policy called *CVAug*. We also introduce *no-zoom crop*, which reduces the zooming component of crop-and-zoom.

- **Domain-inspired augmentation:** WV images are often dominated by ocean surface waves, so representing the image in the frequency domain and dropping random frequency components emphasizes or de-emphasizes particular features that could be relevant to sea-surface state. We call this *random notch filtering*.

All augmentations are added to the overall transform pool from which to sample during training. That means that each augmentation policy, be it from the original SimCLR policies or one of the added policies described here, gets applied with some probability to each image. An image may be transformed by any combination of policies, including all or none, and the sampling is repeated for every image in every batch. Examples for each augmentation policy can be seen in Figure 3.1.

The set of augmentations was optimized using a local search strategy. One augmentation policy at a time is introduced to the baseline SimCLR policies, and the performance is evaluated. All models are trained for 100 epochs total where one epoch consists of training on a random 30% sample — roughly 3.8M unique samples — of the full unlabeled dataset. This sample is redrawn every 20 epochs, allowing for reduced computational cost while still exposing the model to the majority of the full unlabeled data at some point during training. All models are trained using 4 V100-32 GB GPUs, 16 CPU cores, and 200 GB of RAM with a global batch size of 512, taking about 6 days to complete 100 epochs. The resulting models are then evaluated on the classification task and the H_s regression task.

3.2.4 Evaluation protocols

To evaluate the quality of the WV-Net embeddings, we conduct experiments in which the embeddings are used for a multilabel classification task, two regression tasks, and an unsupervised image retrieval task. The experiment protocols are summarize below.

Multilabel classification: For all classification tasks a subset of classes from the GOALI dataset (WS, MC, NV, RC, CP, AB, LW, BS, OF, IW, and SI) will be considered as they comprise the majority of the phenomena of interest for downstream applications and together comprise the vast majority of the dataset. All other labels are grouped into a catch-all "Other" class. The classification data is stratified and randomly split into 60%, 20%, and 20% of the original 10,000 images for training, validation and hyperparameter tuning, and final model testing, respectively. All 6,400 images from Wang et al. [37] are held out for testing.

The kNN classifier is trained according to the protocol from Wu et al. [71] with 15 neighbors, chosen based on a hyperparameter sweep. Cosine similarity (Equation 3.4) is used as the distance metric for the kNN model.

The protocol from Chen et al. [3] is directly adopted for linear probing with no modifications. The proposed MLP architecture from Bordes et al. [16] (2 hidden layers with 2048 ReLU [156] units) is adopted, and the model is trained for 200 epochs using the Adam optimizer [157] with a constant learning rate of 0.001.

The fine-tuning procedure from Chen et al. [3] is followed with the batch size reduced to 256 and the learning rate scaled to 0.05 accordingly.

The fine-tuned classification models are trained to minimize the sum of binary cross-entropy losses over all individual classes to allow for the multilabel property:

$$\mathcal{L}_{cls} = -\sum_{i=1}^{N}(\sum_{c=1}^{C} y_{i,c}\log(\hat{y}_{i,c}) + (1-y_{i,c})\log(1-\hat{y}_{i,c}))) \tag{3.5}$$

The micro-averaged AUROC is computed by summing the predictions for each class and then calculating an AUROC curve for the aggregated predictions. The F1-score is also reported.

Regression: Linear probing and MLP protocols are left unchanged from the classification protocols except for adjusting the output function. Hyperparameters were minimally adjusted for the end-to-end fine-tuning scenario since the hyperparameters used for the classification task showed instability. The final parameters are 10^{-6} weight decay, a backbone learning rate of 0.007, an output-layer learning rate of 0.025, a dropout rate of 0.5, and the weighted loss described below. All other hyperparameters are left unchanged. The fine-tuned regression models use a softplus output unit and are trained using a weighted combination of the mean absolute (or L1) error (MAE) and the mean squared error (MSE) weighted in favor of the MSE:

$$\mathcal{L}_{reg} = \sum_{i=1}^{N} 0.1 * |y_i - \hat{y}_i| + (y_i - \hat{y}_i)^2 \tag{3.6}$$

Models are evaluated using the mean absolute error (MAE) and root mean squared error (RMSE).

Image retrieval: The embeddings are evaluated for one-shot image retrieval performance, following the kNN-retrieval approach from Caron et al. [18]. Experiments are conducted on the rarest classes from the combined classification dataset, occurring in no more than 1,000 images (<0.05% of the total dataset), consisting of seven total classes. Models are evaluated in terms of mean average precision (mAP) averaged over all classes.

3.3 Results

Experiments were conducted to test two hypotheses: (1) a self-supervised model trained on WV data will outperform a model trained on ImageNet, and (2) the self-supervised model can be

Framework	Backend Architecture	Classification (AUROC)	Wave height (RMSE)
SimCLR	ResNet50	**0.935**	**0.564**
	ConvNeXt	0.911	0.991
	ViT	0.881	1.153
BYOL	ResNet50	0.931	0.922
	ConvNeXt	0.889	1.133
	ViT	0.885	1.155
SwAV	ResNet50	0.925	0.780
	ConvNeXt	0.915	1.126
	ViT	0.927	1.168
MAE	ViT	0.920	1.160

Table 3.1: Validation set performance of different contrastive framework and backend architecture combinations. The best model per task is highlighted in bold.

improved by selecting pretext tasks based on domain-specific properties of the satellite images. We first perform experiments to select a self-supervised framework and vision backbone architecture. With those selected, we optimize the set of augmentations used in the SimCLR training algorithm. WV-Net was then trained using the optimized augmentations for an extended period. This model is compared to an ImageNet model, a model pre-trained in a supervised manner on Sentinel-1 land cover data [158], and a WV model trained with the default SimCLR augmentations, testing hypotheses (1) and (2), respectively.

3.3.1 Framework and backend choice

Since there are multiple possible choices for contrastive self-supervised frameworks, we chose to evaluate one representative member of each of the framework families proposed by Balestriero et al. [79]. SimCLR [3] for the deep metric learning family, bootstrap your own latent (BYOL) [44] for the self-distillation family, and swapping assignments between multiple views of the same image (SwAV) [78]. Similarly, there are multiple potential choices for families of backend architecture that have shown promise in a broad range of computer vision tasks. We chose a ResNet50 [147] to represent a standard convolutional architecture, ConvNeXt-T [159] to represent a more modern version of a convolutional architecture, and a ViT-S/16 [63] to represent vision transformers. The model sizes were chosen to have roughly the same number of trainable parameters and constrained to fit the available compute budget. All hyperparameters were set in accordance with the original framework papers. Table 3.1 details the performance results for fine-tuned models on the classification and wave height tasks, showing clear dominance by the SimCLR + ResNet combination.

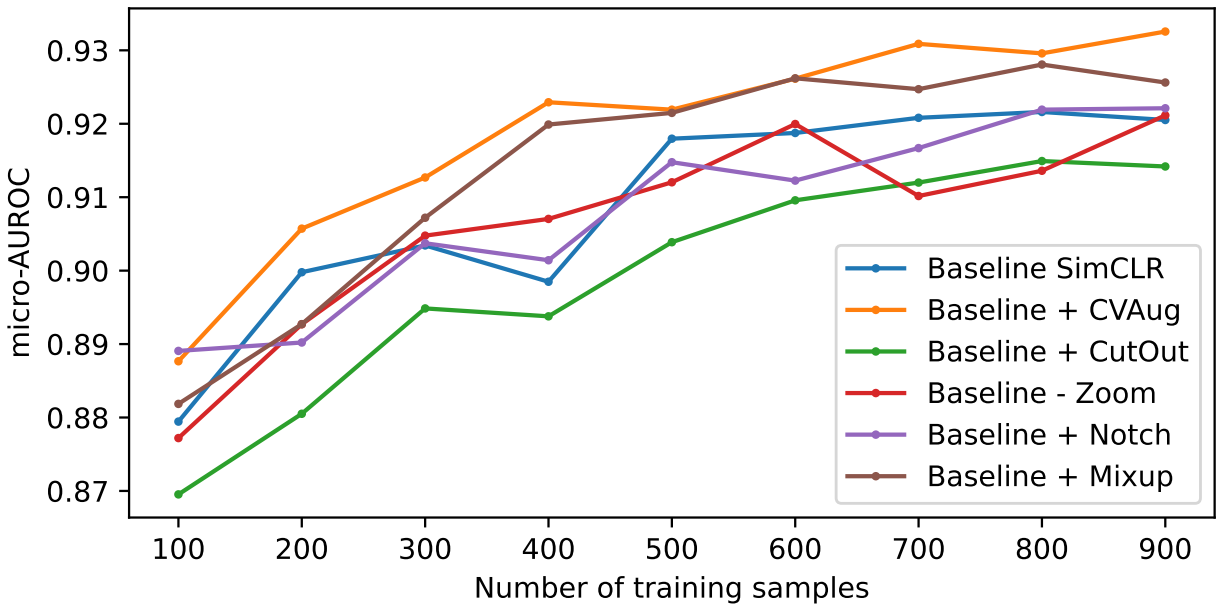


Figure 3.3: Performance of various embeddings (Micro-AUROC, higher is better) vs. number of labeled training samples in the multilabel classification task. This experiment used the MLP transfer learning protocol

3.3.2 Optimization of WV-mode-specific data augmentations

Beyond absolute downstream performance, it is also important to understand how the models perform in data-constrained environments. An advantage of transfer learning is that it drastically reduces the need for labeled examples, lowering the barrier of entry for solving science problems with machine learning. Figure 3.3 shows the MLP transfer performance of the differently pre-trained models for low numbers of training samples for image classification. Mixup and CVAug robustly perform better than most other models with micro-AUROC statistics greater than 0.92 and are only trained with 900 images. The performance differences are larger than those observed on the full training dataset in Table 3.2. Therefore, for rare classes or in situations when the training datasets are small, CVAug and mixup notably improve the model performance.

The linear probing results are illustrative of the overall trends observed in this part of the study and show that Mixup and CVAug consistently improve performance for both the classification and regression tasks. However, the domain-inspired notch filter policy, Cutout [148], and reduced zooming in the crop-and-zoom augmentation did not improve performance — thus these augmentations were not included in the final model.

Evaluation method	Model	Classification		Wave height	
		AUROC↑	F1-Score↑	MAE↓	RMSE↓
kNN	ImageNet	0.925	**0.675**	-	-
	Baseline SimCLR	0.928	0.674	-	-
	Baseline + Cutout	0.920	0.660	-	-
	Baseline + CVAug	**0.935**	**0.690**	-	-
	Baseline + Mixup	**0.930**	**0.676**	-	-
	Baseline + Notch	0.926	0.668	-	-
	Baseline - Zoom	0.927	0.661	-	-
Linear	ImageNet	0.952	0.730	0.447	0.601
	Baseline SimCLR	0.953	0.735	0.428	0.580
	Baseline + Cutout	0.943	0.705	**0.380**	**0.516**
	Baseline + CVAug	**0.954**	**0.742**	**0.413**	**0.555**
	Baseline + Mixup	**0.953**	**0.738**	**0.392**	**0.531**
	Baseline + Notch	0.949	0.720	**0.427**	**0.578**
	Baseline - Zoom	**0.955**	0.722	0.485	0.647
MLP	ImageNet	0.931	**0.715**	0.479	0.656
	Baseline SimCLR	0.933	0.709	0.406	0.561
	Baseline + Cutout	0.919	0.681	**0.360**	**0.494**
	Baseline + CVAug	**0.941**	**0.728**	**0.385**	**0.533**
	Baseline + Mixup	**0.938**	**0.725**	**0.373**	**0.514**
	Baseline + Notch	0.924	0.694	0.410	0.564
	Baseline - Zoom	0.926	0.687	0.488	0.645
Fine-tuned	ImageNet	0.931	0.759	2.696	3.001
	Baseline SimCLR	0.932	0.765	0.424	0.604
	Baseline + Cutout	**0.933**	0.749	0.448	0.629
	Baseline + CVAug	**0.935**	**0.768**	0.433	0.614
	Baseline + Mixup	**0.935**	**0.770**	**0.393**	**0.556**
	Baseline + Notch	**0.934**	0.748	**0.419**	**0.598**
	Baseline - Zoom	0.931	**0.765**	0.444	0.624

Table 3.2: Comparison of different augmentation methods used during training with SimCLR. The original (baseline) SimCLR method trained on satellite data does better than a standard ImageNet model, and these results can be improved with modifications to the set of augmentations. The classification scores are micro-averaged AUROC and micro-averaged F1-scores (higher is better). The wave height scores are the RMSE and MAE (lower is better). Models that outperform baseline SimCLR are in bold.

Eval. Method	Model	Classification		Wave height (m)		Air temperature (°C)	
		AUROC↑	F1-Score↑	MAE↓	RMSE↓	MAE↓	RMSE↓
kNN	ImageNet	0.925	0.675	-	-	-	-
	BigEarthNet-S1	0.894	0.599	-	-	-	-
	Baseline SimCLR	0.925	0.669	-	-	-	-
	WV-Net (ours)	**0.936**	**0.697**	-	-	-	-
Linear	ImageNet	0.952	0.730	0.447	0.601	0.682	0.974
	BigEarthNet-S1	0.929	0.650	-	-	-	-
	Baseline SimCLR	0.954	0.739	0.395	0.532	0.655	0.920
	WV-Net (ours)	**0.958**	**0.754**	**0.370**	**0.500**	**0.637**	**0.902**
MLP	ImageNet	0.931	0.715	0.479	0.656	0.702	0.996
	BigEarthNet-S1	0.906	0.630	-	-	-	-
	Baseline SimCLR	0.930	0.716	0.355	0.491	**0.691**	**0.960**
	WV-Net (ours)	**0.948**	**0.744**	**0.335**	**0.459**	0.763	1.01
Fine-tuned	ImageNet	0.931	0.759	2.696	3.001	0.661	0.964
	Baseline SimCLR	0.934	0.760	0.418	0.586	**0.623**	**0.902**
	WV-Net (ours)	**0.939**	**0.777**	**0.377**	**0.530**	0.635	0.923

Table 3.3: Comparison of final model performances. AUROC and F1 scores correspond to the image classification, and MAE and RMSE are used for the regression tasks to estimate significant wave heights and air-sea temperature differences. The best score for each task under different evaluation scenarios is highlighted in bold.

3.3.3 Transfer learning

Based on the optimization experiments above, we selected four augmentations to add to the baseline SimCLR augmentation pool: mixup, random color inversion, random rotation, and a random sharpness transform. The parameterization of these transforms remains unchanged. These are used to train the final model, called WV-Net.

In experiments, WV-Net was then compared to the baseline SimCLR model trained on WV images (without additional augmentations), the ImageNet model trained using supervised learning, and BigEarthNet-S1, a model pre-trained on the Sentinel-1 landcover dataset by Clasen et al. [158]. The two SSL models (WV-Net and baseline SimCLR) are pre-trained for 200 epochs with a global batch size of 1024 (and accordingly a learning rate of 1.2) on 8 V100-32 GB GPUs, using 400 GB of RAM and 36 CPU cores. Training takes about 12 days to complete.

Table 3.3 compares the performance of the models on three supervised learning tasks using four protocols. WV-Net outperforms the other models on most tasks under most evaluation scenarios. The BigEarthNet-S1 model performs especially poorly, illustrating the shortcomings of supervised

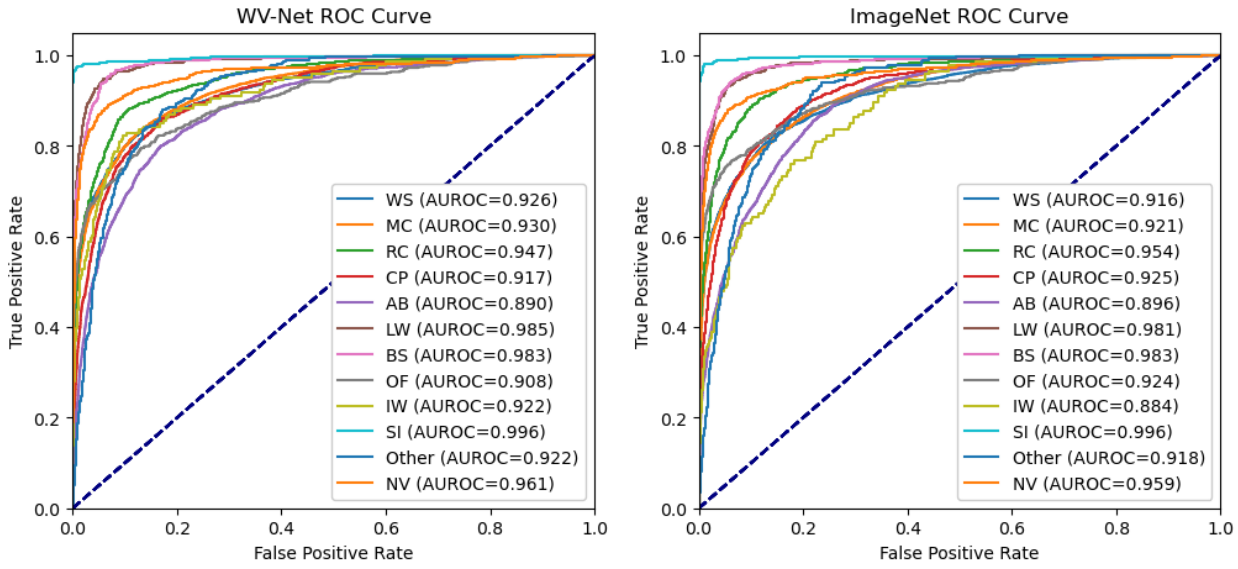


Figure 3.4: ROC curve comparison for WV-Net and ImageNet models fine-tuned on multilabel classification task.

pre-training and the wide gap between SAR WV data and land-focused imagery (because performance was so much worse than even the ImageNet model on the classification task, we skipped evaluation on the regression tasks). This underscores the need for a dedicated WV foundation model as even pre-training on other imagery acquired with the same sensor can lead to suboptimal performance. The only task where other models perform better than WV-Net is the air temperature prediction task, where WV-Net performs the worst in the MLP scenario and slightly worse than the baseline SimCLR model when fine-tuned end-to-end. In general, the linear models perform the best on the classification task, while the MLP and fine-tuned models perform the best on the wave height and air temperature regression tasks respectively.

While WV-Net improved performance on each of the three transfer learning tasks, some tasks saw larger improvements than others. The wave height regression task was particularly difficult for the ImageNet model — even with model fine-tuning the model failed to extract features relevant to the task, while WV-Net had no problem (Table 3.3). This is likely due to the importance of subtle texture features that are very different from the large-scale visual features required for object classification in NIs. On the multilabel classification task, we observed more modest improvements across the individual classes, with some classes having slightly lower AUROC than with the ImageNet model, but WV-Net performs better overall (see Figure 3.4. This could indicate that performance on that task is largely saturated or that identifying geophysical structures in the images can be done sufficiently well with the object-centric, general-purpose features learned by ImageNet models.

Figure 3.5 also illustrates that performance trends hold for data-constrained settings. Even with

as few as 100 total samples, WV-Net embeddings achieve above 0.85 micro-averaged AUROCs in all transfer learning settings, consistently beating ImageNet embeddings as dataset size is increased.

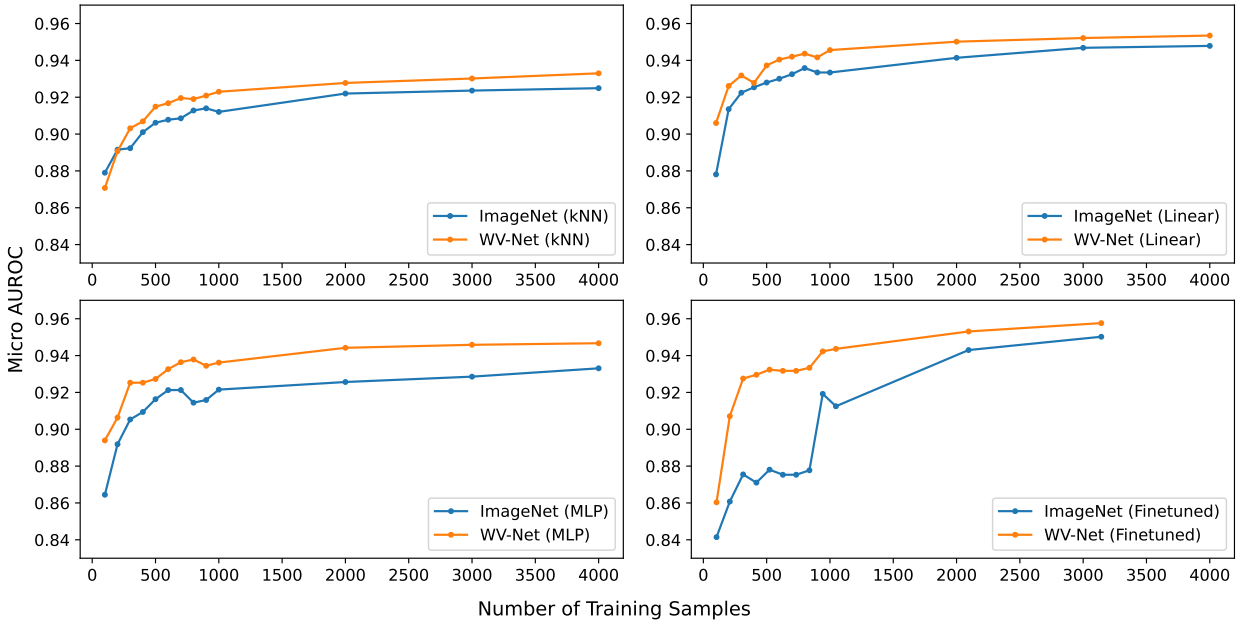


Figure 3.5: Performance of WV-Net vs. ImageNet embeddings on the downstream classification task as a function of the number of labeled training samples. WV-Net almost always outperforms ImageNet for any evaluation method and any number of labeled training samples.

3.3.4 Image retrieval

The image-retrieval task illustrates the capability of the learned embeddings from the SSL model to delineate between features of interest without any fine-tuning. WV-Net outperforms ImageNet embeddings in almost all the rare classes and remains competitive in all other cases, as detailed in Table 3.4. Because the dataset is multilabel and several classes can be present in a single image, identifying all classes from a single example can be noisy and lead to mAP scores that appear lower than for single-label datasets common in NI applications. When scoring for any class overlap between the anchor and retrieved images, mAP for both models approaches 1.0, illustrating that they can retrieve images that share some dominant characteristics. The fact that WV-Net otherwise outperforms ImageNet suggests that the SSL embeddings are more sensitive to secondary classes present in the images, allowing for more fine-grained delineation. For reference, Table 3.5 shows image retrieval performance for the remaining, non-rare classes, with WV-Net still displaying consistently higher or competitive mAP scores.

Figure 3.6 a) shows an example of the top three retrieved images for a reference atmospheric gravity waves (AW) image, or anchor. This class makes up less than 0.5% of the overall dataset.

Model	AW (N=101)	IW (N=304)	OE (N=142)	SI (N=955)	IB (N=762)	SH (N=236)	SW (N=167)
ImageNet	0.013	0.184	**0.130**	0.845	0.223	0.016	**0.024**
WV-Net (Ours)	**0.127**	**0.297**	0.119	**0.901**	**0.398**	**0.021**	0.020

Table 3.4: Comparison of image retrieval performance. mAP scores shown on rare classes in the GOALI dataset for ImageNet and WV-Net embeddings with the better-performing model for each class highlighted in bold.

Given the anchor image in Figure 3.6 a), WV-Net embeddings give an average precision of 0.95 for the top 20 retrieved images, outperforming the 0.11 average precision of ImageNet embeddings. It appears that the samples retrieved using ImageNet mostly share similar contrasts in the SAR backscatter, while WV-Net consistently identifies the correct characteristics associated with the class. However, Figure 3.6 b) illustrates that given an anchor image where the class is less obvious (AB with subtle AW signatures), WV-Net embeddings also fail to capture the relevant class characteristics. Nevertheless, Table 3.4 again shows that, on average, WV-Net is more robust to the anchor choice for the AW and most other classes. This is similar to the uncertainty that humans have when characterizing images that contain multiple features. This may also explain the overall low mAP scores shown for the SH (ship) and SW (ship wake) classes in Table 3.4, because these are generally small, isolated objects in the image where other phenomena dominate the ocean surface backscatter. More examples are provided in Figure 3.7 and Figure 3.8. Figure 3.7 shows retrieval for randomly sampled anchors of each class and Figure 3.8 shows anchors specifically sampled from the subset of images that only contain a single tagged class. These figures further illustrate that selecting "pure" anchors can improve image retrieval. Further, when multiple classes are present in the anchor, WV-Net embeddings especially seem to retrieve images from all classes present, perhaps indicating a more evenly distributed embedding space.

Model	NV (N=2243)	WS (N=11458)	MC (N=11644)	CP (N=1745)	RC (N=2174)	AB (N=3422)	LW (N=1646)	BS (N=16
ImageNet	0.475	0.648	0.672	0.318	0.477	**0.353**	0.714	**0.56**
WV-Net (Ours)	**0.561**	**0.662**	**0.752**	**0.345**	**0.510**	0.343	**0.778**	0.56

Table 3.5: Comparison of image retrieval performance for dominant classes. mAP scores for most prevalent classes in the GOALI dataset for ImageNet and WV-Net embeddings with the better-performing model for each class highlighted in bold.

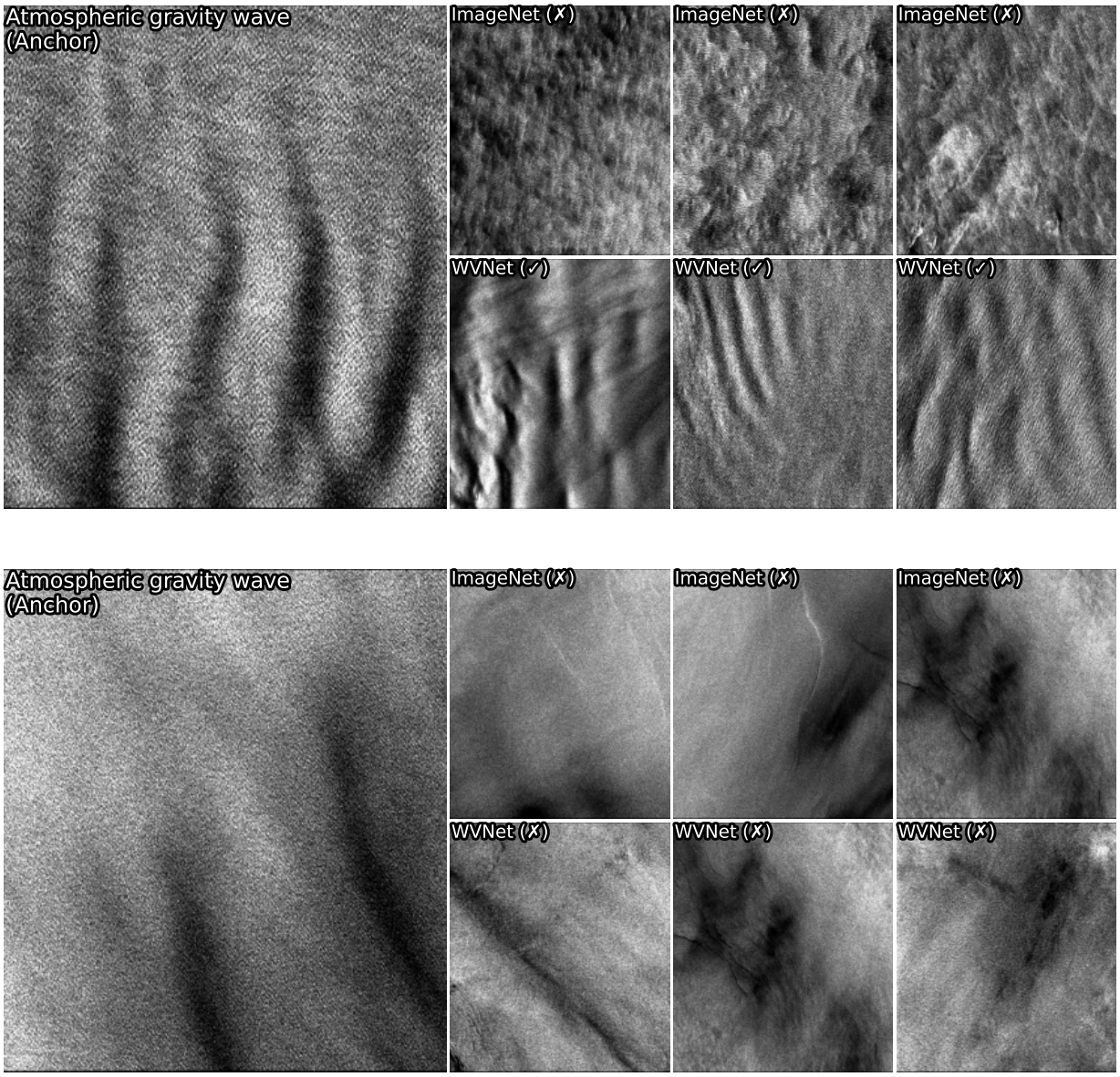


Figure 3.6: Image retrieval examples for the atmospheric gravity wave class. Anchor image (left column) is the query for kNN retrieval, and the six images to the right are the top three neighbors from ImageNet (top row) and WV-Net (bottom row) embeddings. The top panel (a) shows successful image retrieval with the class present in the lower half of the anchor image. The bottom panel (b) shows unsuccessful retrieval with the class difficult to discern in the anchor image. This sample illustrates an anchor for which both architectures have uncertainty since the target class in the anchor image is not well pronounced.

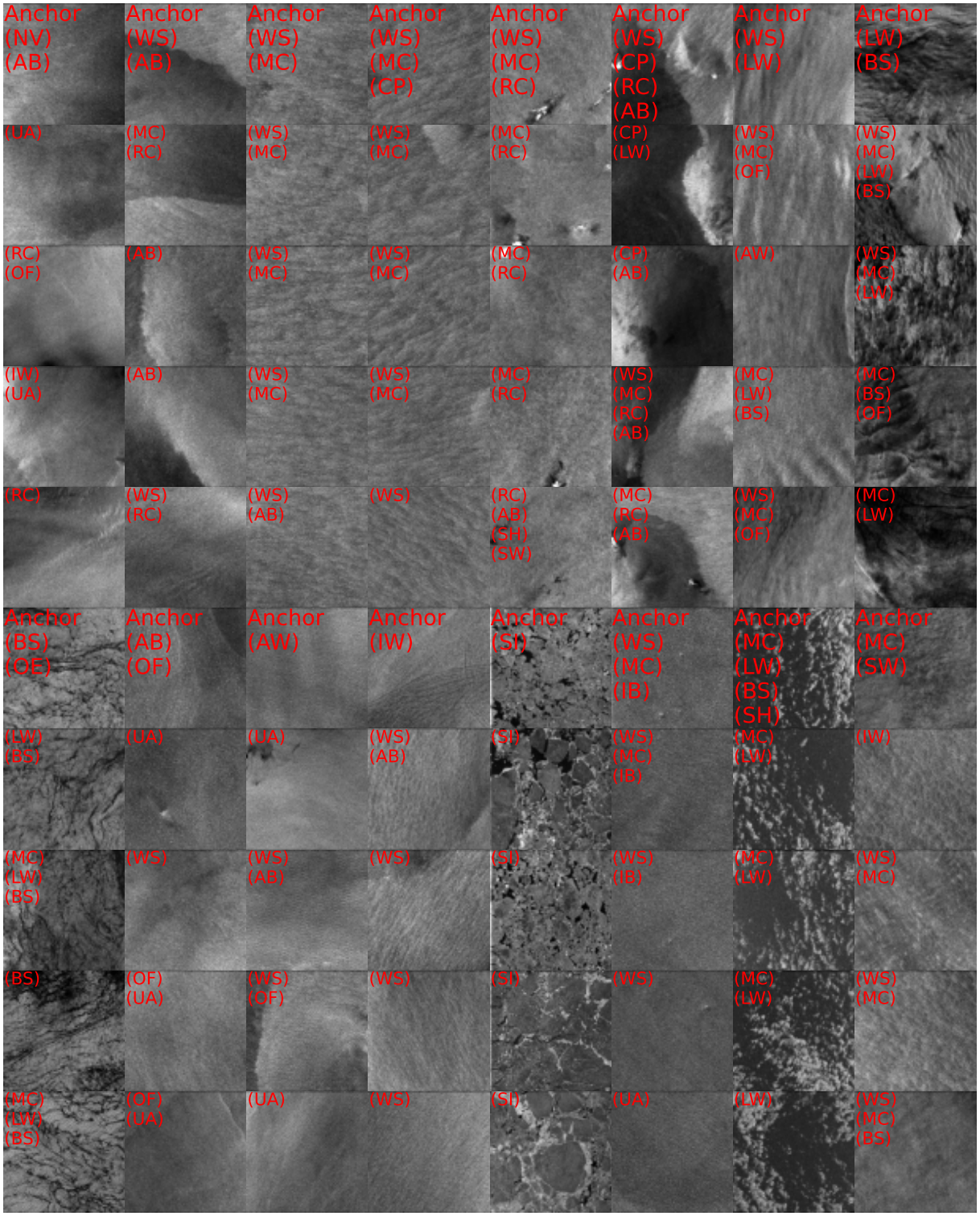


Figure 3.7: WV-Net image retrieval examples for random anchors. Besides ensuring that every class is represented at least once, anchors are sampled completely randomly.

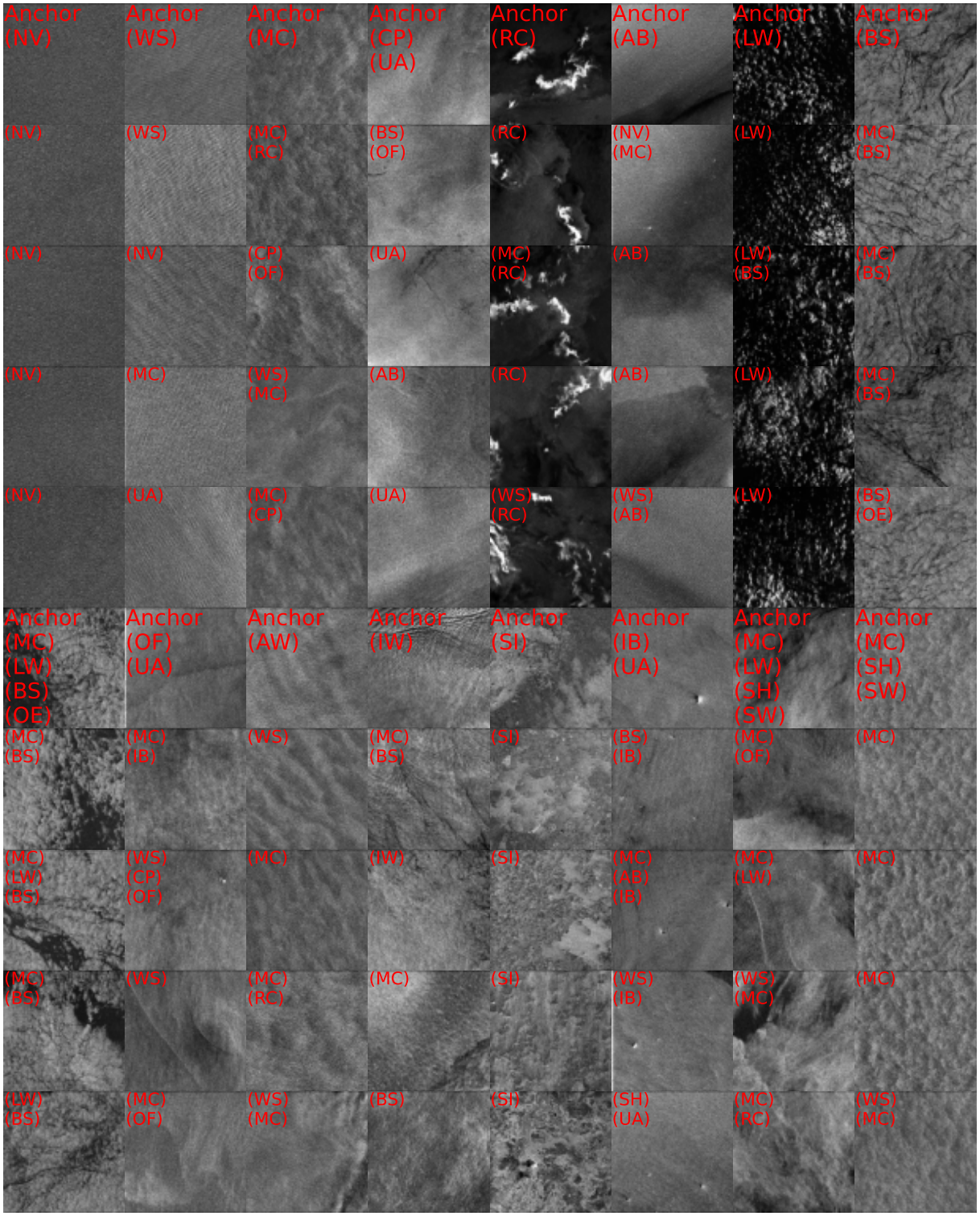


Figure 3.8: WV-Net image retrieval examples for random anchors. Anchors are randomly sampled for each class from examples that only contain that class where possible.

3.4 Discussion and Limitations

One limitation of this work is the computational cost of pre-training. Model performance could likely be improved with larger models, longer training, and more extensive hyperparameter optimization. Carefully tuning the temperature parameter during pre-training can impact task performance, especially for small batch sizes (relative to other contrastive models), such as those employed here [16]. Contrastive SSL models have been shown to scale effectively with model capacity [3, 21], thus we expect that training a larger model such as ResNet152(x2, x4) with our setup would result in even better performance on downstream tasks.

Similarly, masked-image modeling and vision transformers (ViTs) have been shown to outperform convolutional architectures given enough training time and data [4, 160]. While ViTs were included in the initial model analysis, the models were relatively small. It is possible that given a larger ViT model and enough training time this could be a competitive approach to the one presented here.

The downstream tasks presented are only a small subset of potential applications. For example, models could be trained to detect the organized large-scale eddies or lack of them (NV, WS, and MC) which are present in nearly 85% of all images. Supplementing the results with a dense prediction task like detecting organized large-scale eddies could provide further insights into the behavior of WV-Net. Previous works such as [161, 162, 163] have based their analysis of the physical dynamics associated with the marine atmospheric boundary layer (MABL) on hundreds of SAR images. WV-Net could help change the study of the MABL by systematically mapping millions of observations in time and space, enabling studies like Stopa et al. [142] to broaden their scope to the entire S-1 archive, changing the field from data-poor to data-rich. Even rarely occurring observations such as small-scale eddies (< 100 m), atmospheric gravity waves in the open ocean, or lines in the sea [164] can be well-detected by WV-Net with minimal additional annotations.

3.5 Conclusion

Using self-supervised contrastive learning on almost 10 million images, we have created WV-Net, the first foundation model for S-1 WV imagery. Experiments on downstream classification, regression, and image-retrieval tasks support the two hypotheses: (1) a model pre-trained with self-supervised contrastive learning on unannotated domain-specific imagery outperforms models pre-trained with supervised learning on NIs, and (2) self-supervised contrastive models can further be improved for non-natural-image tasks by carefully selecting pretext tasks, or augmentations. However, we found that the best augmentation strategies were not necessarily the ones that leveraged any particular domain knowledge, such as random notch filtering. Instead, we found that the best augmentations were the original SimCLR augmentations plus mixup, rotations, color inversions, and sharpness transforms.

WV-Net outperforms models pre-trained on NIs and even land-cover S-1 data from the same satellite platform. While the performance improvement of WV-Net over ImageNet models is small for some tasks, the advantage is consistent across tasks. Of the three supervised learning tasks, the largest performance improvement is observed for the wave height prediction task, which requires extracting fine-scale features that may be washed out in the ImageNet model, where final layers primarily capture high-level semantic concepts rather than low-level texture features [125]. Furthermore, experiments demonstrate that WV-Net embeddings can yield state-of-the-art performance without the need for end-to-end fine-tuning, drastically reducing the need for computational resources and time. In fine-tuning settings, WV-Net also demonstrates greater robustness to hyperparameter choices, alleviating the need for broad hyperparameter sweeps. WV-Net is also more data-efficient than competing approaches, requiring less labeled data and even displaying strong image retrieval performance with no labeled data at all. Together, these properties make WV-Net a valuable tool for the remote sensing research community. WV-Net weights and code to run the model will be made available at https://github.com/hawaii-ai/WVNet/.

More generally, this work demonstrates the value of designing domain-specific foundation models. While WV-Net is designed specifically for WV-mode images from the Sentinel-1 mission, our approach can be applied to other remote sensing imaging technologies with different physical scales. These include other important ocean monitoring technologies like Surface Water and Ocean Topography (SWOT), other SAR modes, or scatterometers. Our experiments show the value of designing a pretext task that is appropriate for the domain, highlighting the value of close collaboration between machine learning and domain scientists.

CHAPTER 4
A DXA BODY COMPOSITION FOUNDATION MODEL

4.1 Introduction

Body composition is well-known to be associated with many critical health outcomes as well as quality of life in general. In older cohorts, low strength, excessive weight loss, and especially decreasing lean mass have all been shown to be associated with all-cause mortality [165, 166]. Equally in older cohorts, bone density, another aspect of body composition, has been shown to be highly predictive of hip fracture along with hip geometry. With hip fracture incidence being repeatedly projected to rise [167, 168] by up to 50% by 2050, assessing and treating risk early can have major quality-of-life and economic impacts. Further, metabolic changes stemming from excess adiposity are associated with an increased risk of developing cardiovascular disease (CVD) and cancer [169] and while deliberate weight loss can act preventative [170], being underweight can risk of dying of increase non-cancer and non-CVD related causes [171]. As obesity prevalence keeps increasing [172, 173, 174], managing weight and body composition effectively, as well as understanding pathology and potential interventions is critical. Especially with the emergence of glucagon-like peptide-1 as a weight loss treatment and its effects on relative body composition still being studied [175], accurate monitoring is critical.

DXA is a low-dose X-ray technology commonly used to measure body composition, bone density, and hip geometry. DXA is a criterion measurement for body composition [176] as well as primary osteoporosis diagnosis tool [177]. As such, DXA measurements have been collected as part of several large NIH studies, such as the Healthy Aging and Body Composition study (HealthABC) [178], the The Healthy Aging in Neighborhoods of Diversity across the Life Span (HANDLS) study [179], The Osteoporotic Fractures in Men (MrOS) study [180] all of which are longitudinal studies with follow-up periods exceeding a decade, and the cross-sectional ShapeUp! study [181], which has the most diverse population out of these studies. This has resulted in a large repository of available NIH DXA data, spanning proximal femur, spine, and whole-body scans, along with various other variables that have been collected as part of these studies. However, these studies do not collect identical variables and have limited follow-up time; as such, metadata and label information are limited, making it difficult to do modeling that takes advantage of the data from all of these studies.

Self-supervised learning presents a viable solution for this by providing a method to learn an embedding model that can take advantage of images from all of these studies and has higher specificity than an embedding model pre-trained on NI data. Further, using self-supervised learning, a model can be pretrained using multiple anatomical regions. This model can then be combined with more limited, study-specific metadata to solve problems where otherwise training a deep learning model from scratch might not be viable.

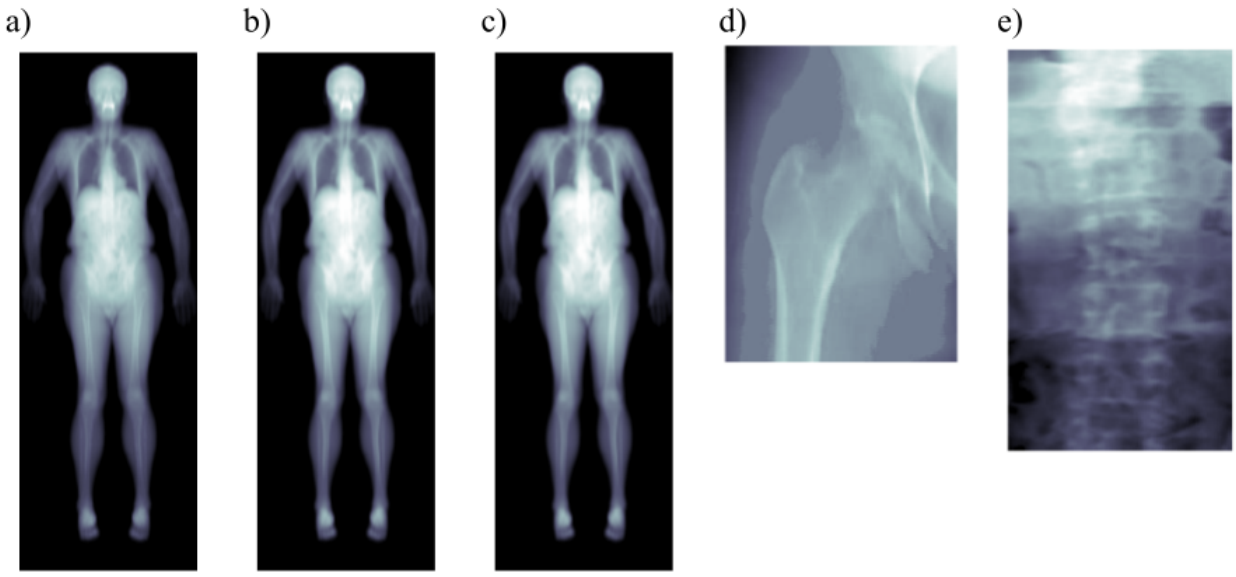


Figure 4.1: Example DXA scans. a)-c) Whole body high energy, low energy, and ratio channels, respectively. d) Proximal femur low-energy channel. e) Lumbar spine low-energy channel.

The goal of this work is to build a DXA foundation model. This model will be trained using self-supervised learning on three anatomical regions that constitute the most common DXA scan sites. We hypothesize that pre-training using a standard self-supervised framework will result in a better embedding model for transfer learning than the standard approach of pre-training on ImageNet [10] data. We further hypothesize that this model can be improved by introducing modality-specific adaptations to the self-supervised framework making it more suited to DXA domain data. To evaluate the efficacy of this method, the models will be tested on two applications with significant real-world utility: all-cause mortality prediction and hip fracture prediction. The model will be tested against current state-of-the-art approaches from the literature and a standard pre-training baseline. This paper will serve to introduce the resulting model and demonstrate its robustness and versatility. Follow-up work will focus on task-specific characteristics of the model. The foundation model will be made publicly available for researchers to use on their DXA data, alleviating the data bottleneck preventing many researchers from including medical imaging data directly into their models.

4.2 Methods

4.2.1 Dataset details

The DXA dataset is composed of data collected from four separate National Institute of Health (NIH) studies. Extending results from [97], three different DXA scan types will be combined,

proximal femur (henceforth referred to as "hip"), lumbar spine, and whole-body, for pre-training. For all studies, minimal data cleaning will be done before pre-training, removing phantoms and visibly erroneous scans (e.g., because of heavy artifacts or significantly limited scan region) from the dataset while preserving the largest number of pre-training samples possible. During all pre-training, 10% of the data, stratified by study and anatomical site, will be held out for performance monitoring. Further, 20% of the HealthABC whole-body dataset and the full HANDLS whole-body dataset are also held out for model testing. Overall, 82,983 DXA scans are used for self-supervised pre-training, consisting of 30,026 whole-body scans, 30,036 hip scans, and 20,842 spine scans.

HANDLS: The Healthy Aging in Neighborhoods of Diversity across the Life Span (HANDLS) study is a community-based longitudinal study that started collecting data in 2004 and was initially set for a 20 year follow-up period with repeat assessment every three to four years [179]. The cohort consists of 3,720 Black and White adults aged 30-64, recruited from neighborhoods in Baltimore. Designed to evaluate health disparities stemming from socioeconomic status (SES) and race, sampling was done in a 4-way factorial design between age, sex, race, and SES. For this work, 25,162 total scans from the HANDLS study are used (8283 hip, 8419 spine, 8460 whole-body), spanning the first 4 waves of data collection. Additionally, demographics and general indicators of fitness associated with mortality are available.

HealthABC: The Health, Aging, and Body Composition (HealthABC) study is a prospective cohort study of 3,075 individuals in metropolitan areas surrounding Pittsburgh, Pennsylvania, and Memphis, Tennessee [178]. The cohort consists of 48.4% men and 51.6% women aged 70 to 79 years at the time of recruitment, 41.6% of whom are Black with the remaining 58.4% being non-Hispanic White. The follow-up period for the study was 16 years with regular check-ins comprised of questionnaires and exam measures. 21,776 total scans, consisting of 10,920 hip scans and 17,553 whole-body scans, will be analyzed, as well as demographic data, indicators of fitness, and blood markers.

MrOS: The Osteoporotic Fractures in Men (MrOS) study is a longitudinal study designed to analyze healthy aging with an initial focus on fractures and osteoporosis in men over the age of 65 and is conducted across six clinical centers in the US (Birmingham, AL; Minneapolis, MN; Palo Alto, CA; the Monongahela Valley near Pittsburgh, PA; Portland, OR; and San Diego, CA) [180]. The cohort is comprised of 5,994 males at baseline visit (2000-2002) and includes five visits total spaced roughly five years apart. The total dataset consists of 17,496 whole-body, 16,437 hip, and 14,761 spine DXA scans along with demographic data, indicators of fitness, and blood markers.

Shape Up!: The Shape Up! study is an ongoing stratified cross-sectional observational study designed to analyze the relationship between body shape and health [181]. The study population

is stratified across sex, age (18-40 years, 40-60 years, >60 years), ethnicity (non-Hispanic White, non-Hispanic Black, Hispanic, Asian, and native Hawaiian or Pacific Islander), body mass index (BMI) (in kg/m^2; <18, 18-25, 25-30, >30), and location (San Francisco, CA; Baton Rouge, LA; or Honolulu, HI). Since this is a cross-sectional study, it is only be used for pre-training purposes, contributing 2247 whole body scans.

For all DXA data, custom software was used to extract raw low- and high-energy X-ray attenuation values directly from the DXA scan file. Additional processing is done for whole-body scans, where raw attenuation images have a reduced initial resolution (width and height) of 109 $\times$ 150 pixels at 16-bit pixel depth with a spatial resolution of 2 mm $\times$ 12.76 mm per pixel. All images are first upscaled by a factor of two to a resolution of 218$\times$300 using bicubic interpolation. Bone and soft tissue calibration phases in the scan file are then used to restore the high and low images to their full resolution of 654 pixels. Image pixels are subsequently squared (2 mm $\times$ 2 mm) via bicubic interpolation by a factor of 6.38 in the y-direction for a final 654 $\times$ 1914 image. During training, images are scaled to be within the range zero to one but not normalized. Unless otherwise specified, the third "air" channel generated by the software along with low and high attenuation channels is left for compatibility with ImageNet models, which expect 3 color channels.

After pre-training, all models are evaluated on two real-world tasks: hip fracture prediction and all-cause mortality prediction. For mortality prediction, the HealthABC data is be split according to [50], who split the data into 70% training data 10% validation data for hyperparameter tuning and early stopping and 20% held-out testing data stratified by participant and outcome (death or no death before end of follow-up). This work will serve as basis for comparison, as it represents the current best model for mortality prediction from DXA given the datasets. In addition to the DXA scans, these models will also be trained with matched risk factor inputs to the original work. These risk factors consist of demographics and anthropometric measurements (race, sex, age, height, weight, and BMI); blood markers (blood glucose, fasting glucose, blood insulin, fasting insulin, hemoglobin A1c, and interleukin 6); general indicators of fitness (walking speed over 3/4/6 m, 20 m, and 400 m; grip strength); and self-reported questionnaire answers (disability status for walking, climbing stairs, and activities of daily living; whether the participant had any recent falls and if so how many).

Similarly, for hip fracture the split from [182] is be adopted. The splitting strategy is the same, and again, the model from this work will serve as a comparison as it is, to our knowledge, the best hip fracture prediction model on the HealthABC dataset. Demographics for both supervised datasets can be found in Table 4.1 Both problems are classification problems; as such, models will be evaluated based on AUROC scores in three standard evaluation settings: kNN, logistic regression, and fitting an MLP head.

	Hip fracture dataset	Whole-body mortality dataset
Female scans, N	5,743	8,989
Male scans, N	5,177	8,229
Mean age (std), y	77.6 (4.22)	75.3 (3.0)
Mean height (std), cm	166.2 (10.0)	165.9 (9.8)
Mean weight (std), kg	74.3 (15.0)	74.9 (15.0)
Mean BMI (std), kg/m^2	27.2 (4.8)	27.1 (4.8)
Cases, N	739	6,911

Table 4.1: Population table for supervised DXA tasks. Cases means hip fracture for the hip fracture dataset and death for the mortality dataset.

4.2.2 Model details

Three self-supervised frameworks are evaluated in this work. SimCLR [3], a contrastive self-supervised framework, and BYOL [44], a non-contrastive joint embedding framework, are both trained with a standard joint embedding learning pretext task. For a random batch of images, each image is randomly augmented twice to generate two *views*. The goal of BYOL is only to maximize similarity between embeddings for the two views (Figure 2.2 a)) while SimCLR jointly aims to maximize similarity for views from the same image with dissimilarity between views of different images in the batch (Figure 2.2 b)). BYOL relies on an asymmetric architecture to prevent representational collapse. Both Siamese arms consist of a large vision backbone and an MLP projection head (similar to SimCLR). One arm (the one whose backbone is used for downstream transfer) has an additional, small prediction head. The other arm's parameters are not being trained using gradient descent; instead, it is an exponential moving average of the transfer arm and has no prediction head. This asymmetric setup prevents the network from collapsing into a trivial solution, which is achieved through the contrastive objective in SimCLR. The last self-supervised framework is a standard masked autoencoder (MAE) (see Figure 2.1). This framework works in tandem with a ViT vision backbone, where the input image is first divided into a sequence of patches and then processed by the framework. 75% of the patches will be masked before encoding through the ViT backbone, with the objective being for a lightweight decoder to be able to reconstruct the masked patches.

Except for the MAE, which is trained with a ViT-S/16 [63] backbone, BYOL and SimCLR are trained with a ResNet50 [147] encoder. The networks are chosen due to their similar number of trainable parameters (∼30M) and practicality, both for training and downstream inference. The resulting models are also compared to a ResNet50 initialized with ImageNet weights to determine the best framework for this task. Next, for the best framework, three common vision backbones (ResNet50, ConvNeXt-T [159], ViT-S/16) will be fit to find the best framework-backbone combination.

Subsequently, the best resulting combination will be retrained with modifications to make it more well-suited to DXA data. Broadly, three types of modifications will be explored: modifying the pool of augmentations, feature engineering, and modifying the loss, with details provided below. Each modification will be introduced individually, and the final model will be trained with a combination of the most successful adaptations.

Augmentation pool modifications: Two new augmentations will be introduced — Cutout [148] and MixUp [150]. Cutout somewhat mirrors the MIM objective of occluding a significant portion of the image; while this is not specific to DXA data, it should push the network toward learning more robust, global features without overfitting to any particular region. Especially for hip and spine scans, which are smaller and more uniform than whole-body scans, this could help prevent short-cutting or overfitting. Cutout is applied up to three times, with independent probabilities of $p = 0.5$ each, and can each time zero out between 2% and 30% of the image with a random aspect ratio between 0.3 and 3.33.

MixUp has repeatedly been shown to be an effective augmentation for self-supervised learning [152, 154] and while there is no intuitive connection to DXA domain data, it can be viewed as introducing additional in-domain noise similar to adding more general Gaussian noise [154]. Additionally, experiments will be done with removing transforms from the augmentation pool. Random grayscaling replaces all color channels with the same weighted average of the original channels while color jittering applies brightness and intensity transforms derived on NIs. While this may force the network to not overly rely on pixel-intensity features, since pixel values are meaningful in this modality and relative intensity between channels corresponds directly to meaningful physical values, this could also introduce unwanted invariance in the network.

Feature engineering: Significant work has been done to map the ratio between the observed attenuation at low energy to the attenuation at high energy to chemical elements [183, 184] and has previously been used to restrict pixel values in autoencoders to physically meaningful ranges [126]. While this information is available from just the high and the low-energy channels, directly introducing it as a third channel may be beneficial to the learning process.

Additionally, the efficacy of combining the three anatomical regions' scans into the same model will be tested here. While results from [97] suggest the additional training data resulting from combining all the DXA scan should be beneficial, it could also come at the cost of specificity. Since whole-body scans are the most abundant in the dataset, a model will be trained utilizing only whole-body scans to test whether the increased specificity is beneficial to the model for the mortality task, which only utilizes whole-body data.

Loss modification: The loss introduced in [185] is a slight variation on the NT-XEnt loss introduced in SimCLR (Eq. 4.1) that removes the similarity between positive views of the same anchor

from the denominator:

$$l_{i,j} = -\log \frac{\exp(sim(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{I}_{[k \neq i \wedge (i,k) \neq \mathbb{P}]} \exp(sim(\mathbf{z}_i, \mathbf{z}_k)/\tau)} \tag{4.1}$$

This has been shown to be especially beneficial in scenarios with smaller datasets or limited batch sizes [16, 79], both of which apply with the DXA dataset.

Another approach to modifying the loss function is by including class information either directly [186] or by the selection of positive pairs, such as [187] do by selecting MRI scans from patients with similar ages as positive samples. This work will explore a similar approach, since DXA DICOM headers usually contain demographic information; this can be done without introducing the need for additional data collection, even for the self-supervised pre-training. Four demographic variables are used (height, weight, sex, and bmi) to calculate a weight for each pair of views depending on the demographics of the people they originated from as such:

$w_{i,j} = \max(0.25, \tanh(\sum_{v=0}^{V} \mathbb{I}_{i=j} |d_{iv} - d_{jv}|))$

This weight gets multiplied by the pair's cosine similarly (Equation 3.4) to down-weigh the loss for similar negative pairs if they come from people with similar demographics.

All models are trained for 200 epochs with hyperparameter configurations as described in the original papers. Training is done across 4 A100 80 GB GPUs with a global batch size of 512 and learning rate scaled accordingly.

4.3 Results

4.3.1 Framework selection

Table 4.2 a) shows results for the initial framework selection runs. SimCLR clearly outperforms all other frameworks on both tasks. Both SimCLR and BYOL also outperform the ImageNet pre-trained model. When training several vision backbones with the standard SimCLR framework, the ViT backbone ultimately performs the best. Noticeably, the same backbone trained with the MAE objective performs significantly worse, indicating that the right choice of framework-backbone combination is crucial.

4.3.2 Framework modifications

Exploring augmentations, it quickly becomes apparent that the hip fracture task is more sensitive to framework modifications than the mortality one. Table 4.3 shows that for the mortality task, the baseline SimCLR model actually displays the best performance out of any variation. The Cutout model is very close in performance but still remains behind for each evaluation setting. Overall, there also is little variation between the best and worst models, with most being within a 0.02 AUROC score range.

	Framework	Backend	Mortality (AUROC)	Hip fracture (AUROC)
a)	ImageNet	ResNet50	0.667	0.588
	BYOL	ResNet50	0.676	0.683
	MAE	ViT-S/16	0.646	0.533
	SimCLR	ResNet50	**0.681**	**0.717**
b)		ResNet50	0.681	0.717
	SimCLR	ConvNeXt-T	0.690	0.735
		ViT-S/16	**0.701**	**0.742**

Table 4.2: Framework and backbone comparison. Best a) framework and b) vision backbone for each task are highlighted.

For the hip fracture task, on the other hand, the models show more variation in performance. All policies outperform baseline SimCLR in the majority of evaluation settings, but only the addition of the Cutout policy outperforms for all settings. Together with this policy also remaining almost on par with the baseline SimCLR model, Cutout will be included in the final model training.

The results for the feature engineering experiments in Table 4.4 clearly illustrate that restricting the training data, even for the sake of specificity, is not beneficial for the model. Training only on whole-body DXA scans, the model under-performs the baseline SimCLR model in the all-cause mortality task, indicating no performance gain from training more specific models at the cost of a reduced overall dataset size. However, adding the ratio channel has the opposite effect, boosting performance above the baseline SimCLR framework in every task for almost every evaluation setting. This is also the only modification achieving a performance boost over the baseline for the mortality task in every setting except for the logistic regression model. The ratio channel is thus a clear inclusion in the final model.

Again, for the modified losses, only the hip fracture task seems really sensitive to any changes of the framework, while the baseline SimCLR model outperforms every other model on the mortality task. The demographic loss, while getting the best logistic regression performance on the hip fracture task, overall seems not beneficial for the model, performing worse in every other evaluation setting. However, the model trained with the DCL loss function does outperform the baseline in every hip fracture setting and, again, is close in performance for the all-cause mortality task; as such, it will also be included in the final model.

Based on these results, the final model will be a ViT trained using the SimCLR framework with the DCL loss, Cutout added to the pool of augmentation functions, and a high-low ratio channel included as an input to the model.

Modification	Evaluation method	Mortality (AUROC)	Hip fracture (AUROC)
Baseline	kNN	**0.601**	0.631
	Logistic regression	**0.701**	0.742
	MLP	**0.692**	0.722
Baseline + Cutout	kNN	0.596	**0.671**
	Logistic regression	0.699	**0.783**
	MLP	0.683	**0.722**
Baseline + MixUp	kNN	0.584	0.602
	Logistic regression	0.684	**0.786**
	MLP	0.667	**0.722**
Baseline - random grayscale	kNN	0.587	**0.642**
	Logistic regression	0.690	**0.777**
	MLP	0.678	0.698
Baseline - color jitter	kNN	0.600	**0.648**
	Logistic regression	0.682	**0.746**
	MLP	0.681	0.710

Table 4.3: Augmentation study. Baseline SimCLR framework compared to versions with added or removed augmentation policies. Models outperforming or performing the same as the baseline SimCLR model are highlighted in bold.

Modification	Evaluation method	Mortality (AUROC)	Hip fracture (AUROC)
Baseline	kNN	0.601	0.631
	Logistic regression	**0.701**	0.742
	MLP	0.692	0.722
Whole-body only	kNN	0.597	-
	Logistic regression	0.690	-
	MLP	0.687	-
Baseline + ratio channel	kNN	**0.602**	**0.673**
	Logistic regression	0.693	**0.779**
	MLP	**0.695**	**0.714**

Table 4.4: Feature engineering results. Whole-body only indicates a model trained only using whole-body DXA scans. Models outperforming or performing the same as the baseline SimCLR model are highlighted in bold.

Modification	Evaluation method	Mortality (AUROC)	Hip fracture (AUROC)
Baseline	kNN	**0.601**	0.631
	Logistic regression	**0.701**	0.742
	MLP	**0.692**	0.722
DCL loss	kNN	0.591	**0.650**
	Logistic regression	0.690	**0.775**
	MLP	0.678	**0.753**
Demographic loss	kNN	0.587	0.590
	Logistic regression	0.681	**0.747**
	MLP	0.674	0.715

Table 4.5: Results of modifying SimCLR loss function during training. Models outperforming or performing the same as the baseline SimCLR model are highlighted in bold.

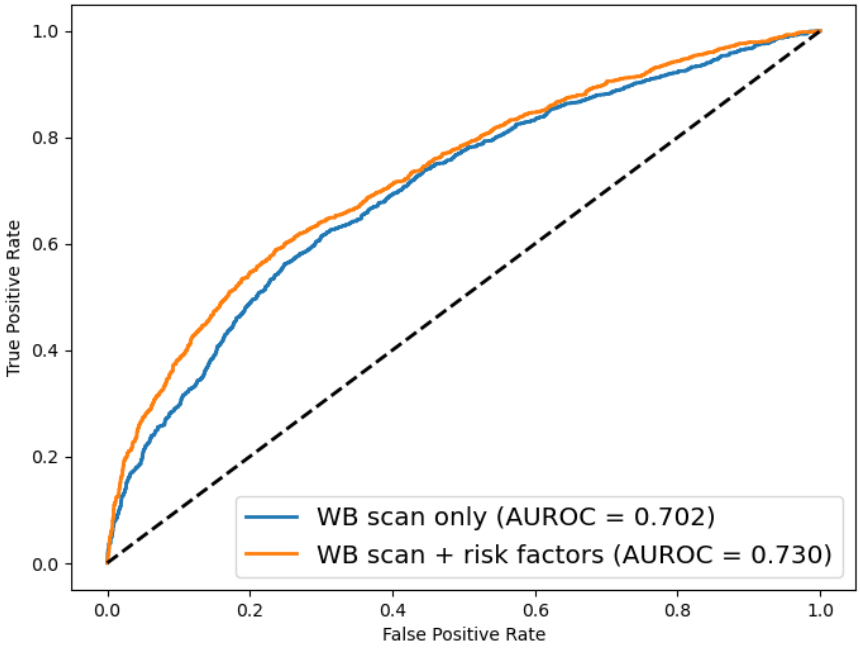
4.3.3 Final model performance

The final model achieves a 0.702 AUROC on all-cause mortality prediction from just whole-body DXA scans. Previously, the best models for this task were presented in [50] which achieved a 0.63 AUROC using only images and 0.71 AUROC combining images and risk factors. The present model performs almost the same without access to any risk factors, and with access to risk factors, as can be seen in Figure 4.2 a), the present model sets a new state-of-the-art without requiring any finetuning.

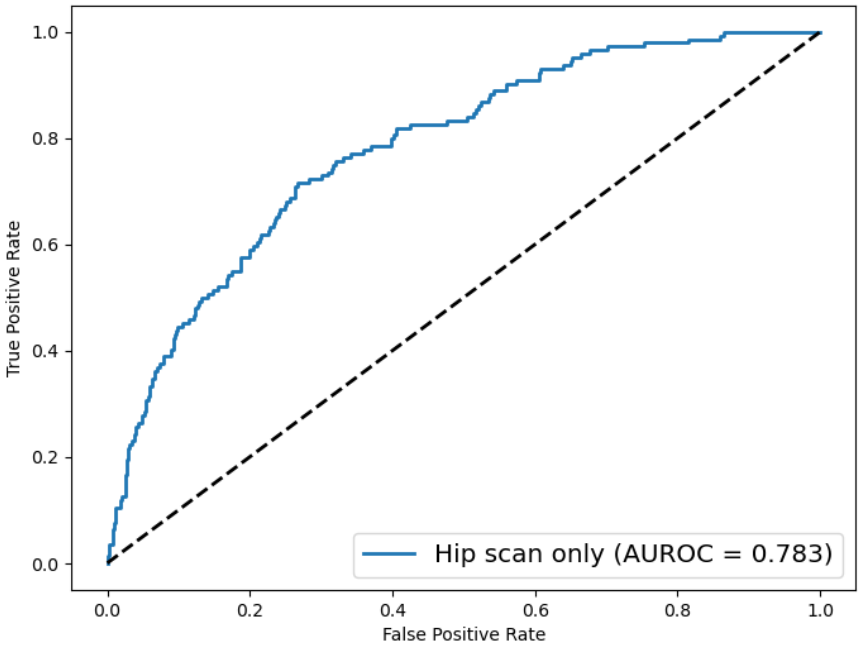
Similarly the model also sets a new state of the art for hip fracture prediction, outperforming the previous best score of a 0.75 AUROC [182] without finetuning and without access to any other risk factors. This is a significant accomplishment considering the previous model used well-established fracture risk factors like demographics, previous fractures, and a direct bone mineral density measure to derive their score.

Compared to models from previous sections, the inclusion of additional components to the framework has not set the model apart on either individual task, performing the same as baseline SimCLR on mortality prediction and slightly worse than some of the best models on hip fracture predictions; however, it has stabilized the performance across both tasks with this being the all-around best-performing model.

Table 4.6 illustrates the model's performance on various subpopulations. What stands out is the better performance for the white and male subgroups compared to their counterparts on mortality prediction. Hip fracture prediction however, seems more robust to participant race. Performances for CVD and cancer death are also both relatively low; only the model combining imaging and traditional risk factors really performs well for the cancer deaths, indicating that the inclusion of risk factors is especially beneficial for causes of death that may not be apparent in the images.



(a)



(b)

Figure 4.2: AUROC performance plots for the final model for a) 10-year all-cause mortality and b) hip fracture.

Task	Participant race		Participant sex		Cause of death	
	Black	White	Female	Male	Cardiovascular	Cancer
10-year hip fracture	0.777 (N=834)	0.773 (N=1295)	0.753 (N=1097)	0.809 (N=1032)	- -	- -
10-year mortality (WB scan only)	0.673 (N=1321)	0.721 (N=2281)	0.660 (N=1809)	0.713 (N=1793)	0.626 (N=828)	0.686 (N=562)
10-year mortality (WB scan + risk factors)	0.709 (N=1321)	0.739 (N=2281)	0.700 (N=1809)	0.741 (N=1793)	0.656 (N=828)	0.726 (N=562)

Table 4.6: Performance of the final model on demographic subpopulations and prevalent causes of death.

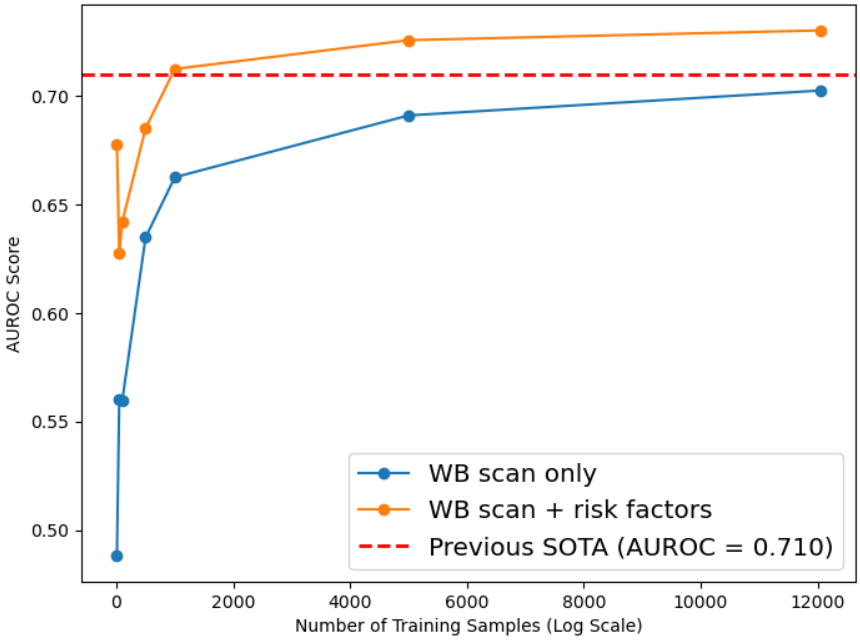
In terms of sample efficiency, the model achieves respectable scores on both tasks, requiring less than 100 samples to outperform the previous image-only best models. And while Figure 4.3 indicates that these models still require in the thousands of samples to surpass the previous state of the art, this is only when comparing to models that have access to established risk factors for both tasks.

Lastly, when performing sensitivity analysis by swapping anatomical scan sites for the tasks, model performance degrades significantly. Using hip scans to predict all-cause mortality achieves a 0.64 AUROC, which is slightly better than the previous best image-only model for this task but still significantly below what this embedding model can achieve with whole-body scans. The performance drop-off for hip fracture from whole-body scans is even worse, degrading to a 0.63 AUROC from a 0.78 when using hip scans.

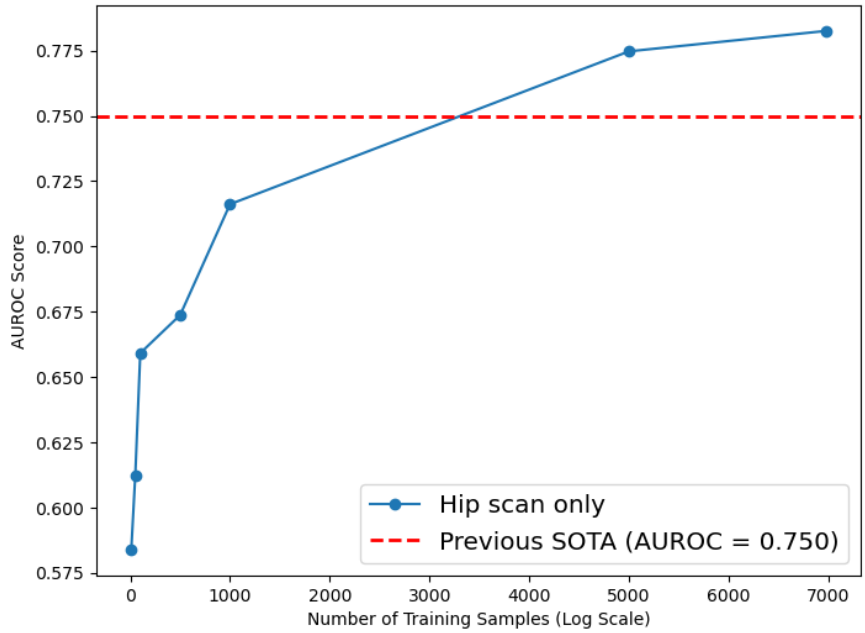
4.4 Discussion

Overall, the results confirm both initial hypotheses. Training a dedicated, self-supervised model for DXA on a large and diverse population of DXA scans results in a superior transfer learning performance. Further, this performance can be boosted, especially for hip fracture prediction, by adapting the self-supervised framework to the modality with a combination of domain-knowledge-informed modifications and changes based on the self-supervised literature.

Importantly, the self-supervised method also enables us to easily combine DXA scans from different anatomical sites, which has a tangible performance benefit for downstream tasks that rely solely on a single site.

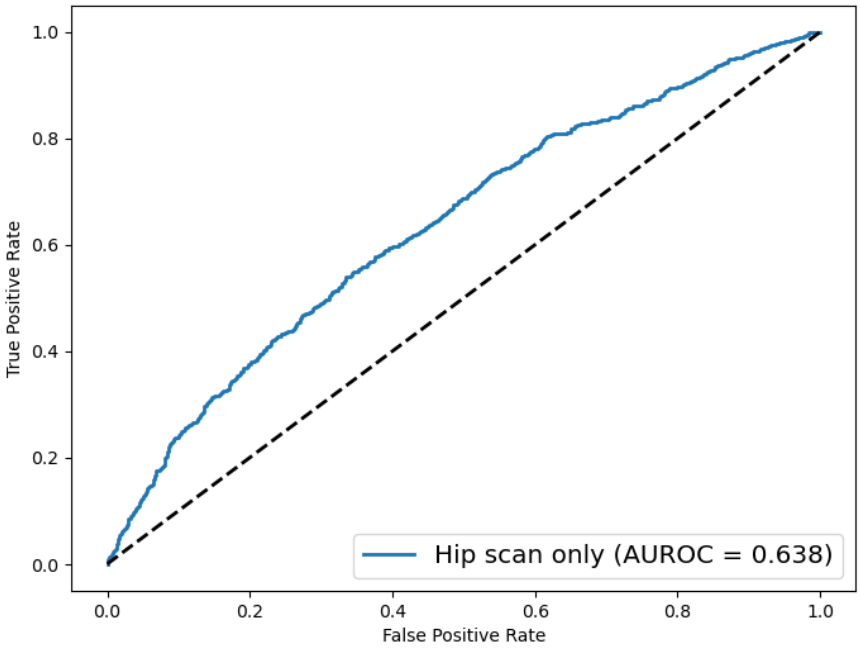


(a)

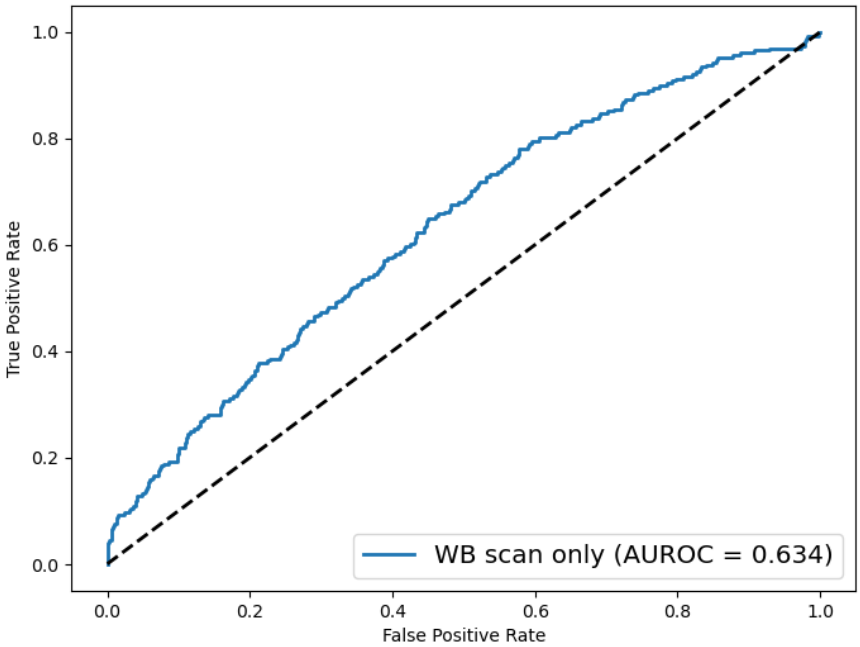


(b)

Figure 4.3: AUROC performance plots for the main model for a) 10-year all-cause mortality and b) hip fracture as training set size increases.



(a)



(b)

Figure 4.4: AUROC performance plots for the main model for a) 10-year all-cause mortality and b) hip fracture with scan sites reversed. Mortality is evaluated on hip scans, and fractures are predicted from whole-body scans.

4.4.1 Framework selection

The result that stands out when evaluating different frameworks on DXA data is the performance difference of a ViT pretrained using an MAE versus one pretrained using SimCLR. The stark difference seems to indicate that the MIM objective is not well-suited to the task. Intuitively, this could be connected to the fact that DXA scans are generally registered with little variation in general position and orientation. The MAE might converge to a simple local minimum of predicting the mean scan since there is not enough variation to produce a large loss. On the other hand, the reconstruction objective could also overly emphasize small details in the image that are not relevant to high-level semantic prediction tasks such as all-cause mortality and hip fracture risk. Lastly, while MAEs scale extremely well with large amounts of data, it is unclear how well they do in more data-limited scenarios such as this one, which could also be a reason why this approach failed to produce good results.

Regardless, the results using the SimCLR framework to train different vision backbones also illustrate that while there is some variation between the backbones, the choice of self-supervised framework seems more impactful.

4.4.2 Framework modifications

There are several takeaways from the framework modification experiments. Most obviously, the hip fracture prediction task seems significantly more sensitive to modifications to the framework than the mortality task. This illustrates how important it can be to evaluate pre-trained models on a range of tasks to prevent overfitting to a single one. Importantly, the best-performing hip models often did not overlap with the best-performing mortality models, indicating that there is some divergence in what features are useful for these tasks and generality is not guaranteed. Fruther, the mortality task overall seems to be more saturated, with performance of the baseline framework being difficult to beat for most modifications. This could potentially be due to how broad 10-year all-cause mortality prediction is as a task. Many factors that will not be captured in DXA scans can play a role, such as disease, random accidents, and lifestyle factors, and as such the 0.7 AUROC score may be close to the saturation point for this particular task.

In terms of actually beneficial modifications, the only three that stuck out in performance were Cutout, the inclusion of the ratio channel, and the DCL loss. The DCL loss is known to be beneficial for contrastive models in settings with relatively smaller batch sizes [16] which aligns well with our results. The ratio channel is a good illustration of domain expertise being included in the modeling process. While it does not provide new information to the model, it makes relationships between the low- and high-energy attenuation channels more explicit, helping with the learning process. Cutout is an augmentation that was also explored in the original SimCLR framework but ultimately omitted from that formulation [3], however, it is still well known as a beneficial augmentation with potential for self-supervised learning by, at least, preventing overfitting to the pretext task. It could also

further disincentivize the framework overfitting to characteristics of specific regions in the DXA scans. Since all scan regions are registered and views of the same area in any given scan generally contain similar anatomical markers, randomly removing that information may force the model to fit to more complex, diffuse information contained in the images.

4.4.3 Final model performance

The final model performs roughly similarly on both tasks to the best individual models from the experiments on framework modifications, indicating perhaps a more balanced embedding space than some previous models. The model also exhibits only a small performance increase from the inclusion of risk factor metadata for the mortality prediction task. This is remarkable as these are well-established mortality risk factors including, for example, blood markers that can indicate high levels of inflammation or diabetes. The fact that the model achieves almost the same performance with only the whole-body DXA scan seems to suggest that much of this information as it pertains to mortality is also, directly or indirectly, present in the DXA scans. Self-supervised pretraining almost fully closing the performance gap between the previous state-of-the-art image-only model (0.63 AUROC) and previous best overall model (0.71 AUROC) really illustrates the potency of this method for learning useful, domain-specific embedding models.

It does exhibit some clear performance biases, especially for the 10-year mortality task. To some extent, this can likely be explained by the over-representation of these groups in the training dataset. Especially men are overrepresented due to the MrOS study including no women. The fact that the model performs better on hip fracture prediction for men stands out since they are less likely to fracture than the women, who are post-menopausal and make up about 70% of all hip fractures annually [188] and 470 out of the 739 fractures in the HealthABC cohort.

The final model also exhibits greater labeled-sample-efficiency, reaching state-of-the-art performance thresholds with 1,000 samples or less for both tasks, an order of magnitude below what previous models used. This requires no careful finetuning either; simply doing logistic regression on the scan embeddings is sufficient.

However, while this model is shown to perform well on a hip DXA task and a whole-body DXA task, it is limited by the information present in the scans. The sensitivity analysis in Figure 4.4 clearly shows that matching the anatomical site to the task is still required. While it seems more obvious that using a hip scan for all-cause mortality is limiting, trying to use whole-body scans for hip fracture, even though they clearly contain relevant information, also seems beyond this model. This could be a limitation in resolution, either of the scans or the learned features of the model. Regardless, it underscores the merits of combining anatomical sites for the training of a single model since switching sites does not require any retraining.

4.5 Limitations

While this work presents the training and evaluation process of our DXA foundation model, more focused investigation into each downstream task remains to be done. The model needs to be expanded to other studies to evaluate whether it can maintain the strong performance on populations with other demographic makeups. It also needs to be validated against standard tools such as the Fracture Risk Assessment Tool (FRAX) and Garvan bone fracture risk calculator[1].

Possibly the largest limitation of the model is its training on only data collected on Hologic scanners, calling into question generalizability to other manufacturers, something that has to be tested.

From a modeling perspective, computational budget is a big limitation for this work, with careful hyperparameter tuning of the SimCLR framework, loss weights, and augmentation probabilities being out of scope. While attempting to use bigger computer vision backbones may also be an avenue to explore, given the relatively limited size of the dataset for a self-supervised learning application that may lead to quick overfitting.

4.6 Conclusion

This work introduces our DXA foundation model. The model is trained on a dataset of four combined NIH studies consisting of over 80,000 DXA scans from three anatomical sites: whole body, proximal femur, and lumbar spine. The model outperforms transfer learning from models pretrained on ImageNet as well as previous state-of-the-art approaches for both 10-year hip fracture prediction and 10-year all-cause mortality prediction. The model performs on par with or better than previous state-of-the-art models without access to any of the risk factors used in those models, predicting directly from the DXA scan. The model is data efficient, requiring fewer labeled training samples and no end-to-end finetuning at all, reducing data and computation bottlenecks. Further, downstream models, especially the hip fracture model, have great potential for opportunistic screening applications. As the model requires no additional input variables besides the hip DXA scan and is relatively lightweight, it could feasibly be run along with any hip DXA scan taken for osteoporosis screening down the line. The embedding model will be made available to the community with the intent of empowering researchers to take advantage of the rich and varied information contained in DXA scans.

[1] https://www.garvan.org.au/research/bone-fracture-risk-calculator

CHAPTER 5
UNSUPERVISED REPRESENTATION QUALITY METRICS

5.1 Introduction

As the previous chapters illustrate, while self-supervised learning is a powerful tool for non-NI vision data, training and especially tuning these frameworks can be cumbersome, requiring either compromises or excessive compute budgets. This difficulty is exacerbated by the lack of good evaluation metrics during self-supervised pre-training. Unlike supervised training, where losses or target metrics on a held-out set can be monitored and used for early stopping, hyperparameter tuning, or model pruning, the losses in self-supervised training are often not well-correlated with downstream transfer task performance [79].

If self-supervision is done with a particular task in mind, linear probing, essentially training a small linear task-head on top of network features during training, is one option for gauging training success, but when building a general-purpose model or wanting to accomplish several tasks this strategy becomes infeasible. Especially for image data from domains with a large domain gap to NIs, this presents a problem as it is well established that self-supervised frameworks are overfit to NI data [16] and benefit from task-specific tuning [29].

Several metrics have emerged to solve this issue. These metrics are not reliant on labels and assign a score calculated only based on raw embedding matrices. Generally, unsupervised embedding space or representation quality metrics are understood from a broadly information-theoretic perspective, trying to estimate whether embeddings produced by a model contain as much information as their size would allow. From this perspective, these metrics are focused on dimensional collapse of the embeddings space, where some dimensions become non-meaningful [22] or embeddings collapse into some highly anisotropic space [114].

Unfortunately, these metrics may suffer from the same limitation as self-supervised frameworks: they are developed and tested primarily on NI vision data. Some metrics are not even developed for vision data at all, originating from NLP research [114]. Further, when these metrics are evaluated, they are primarily tested for correlation with classification problems. While no research has directly refuted the usefulness of these metrics for domains outside what they have been tested on and settings beyond classification, generalizability remains unclear, and since there is no incentive and not enough trust for self-supervised practitioners to use these scores beyond well-tested domains, they remain under-utilized and under-explored.

This chapter will use the models trained in previous chapters of this dissertation and evaluate unsupervised representation quality metrics from the literature against downstream task performance, thus providing some insight into these metrics' behavior. The goal is to establish whether any metrics can be relied upon even outside commonly tested domains and applications. Given the

relatively large scope of models explored in this work and diverse downstream tasks, it is a good benchmark for the reliability of these scores. If metrics generalize well here, they can more reliably be used for model selection, hyperparameter tuning, pretext task selection, and other framework modifications. This can have a significant impact on the workflow of practitioners as all of these tasks currently require training a self-supervised model to completion, which can take hundreds of hours, and then having access to enough labeled data to evaluate the model on a downstream task. This is computationally expensive, slow, and presents a significant barrier to entry for employing self-supervised learning, a tool that could otherwise benefit many under-served domains.

5.2 Methods

5.2.1 Unsupervised embedding space metrics

As mentioned before, most representation quality metrics come from a perspective of embedding space utilization. α-ReQ [23] emerged from the observation of [116] that, empirically, the singular values of embedding matrices decay according to a power law, $\lambda_i \propto i^{-\alpha}$. If the decay coefficient, α, is around 1 this is correlated with strong downstream task performance.

Similarly, the normalized eigenvalue sum (NESum) introduced by [115] quantifies how balanced the eigenspectrum of the covariance matrix M of representations is:

$$NESum(M) = \frac{\sum_i \lambda_i e_i}{\sum_i \lambda_i} \tag{5.1}$$

RankMe[22] approaches the problem from a different angle, estimating the effective rank of the representation matrix as the entropy of the normalized singular values of an embedding matrix M:

$$RankMe(M) = -\sum_i p_i log\ p_i, \quad \text{where} \quad p_i = \frac{\sigma_i}{||\Sigma||_1} \tag{5.2}$$

[24] provide a good review over most of these metrics, proposing three broad categories to expand beyond the information-maximization metrics.

The linear classifier perspective tries to quantify how hard it is for a given matrix of embeddings to find a linear transformation to downstream task targets. The authors propose coherence (μ_0-incoherence) to characterize how aligned the singular vectors of an embedding matrix are to the standard basis. Higher coherence indicates better potential performance.

The high-dimensional probability perspective evaluates embedding quality by comparing the distribution of the n embeddings to a random uniform distribution on a d-dimensional sphere. The authors introduce the SelfCluster metric based on this:

$$\text{SelfCluster}(M) = \frac{d||MM^\top||_F - n(n-1)}{(d+n-1)(n-1)n} \tag{5.3}$$

The numerical linear algebra perspective attempts to use tools from numerical linear algebra to measure the stability of learned representations. More stable representations mean a less sensitive a linear system to small changes in the input. The pseudo condition number (κ_2) is proposed here, calculating the ratio of the largest and smallest singular values to estimate stability $\kappa_2(M) = \frac{\sigma_1}{\sigma_n}$.

Lastly, [114] propose IsoScore to quantify how uniformly a point cloud utilizes its ambient vector space. It works by first reorienting the data using PCA and then calculating an "isotropy defect," which measures the Euclidean distance between the data's normalized variance vector and a perfectly uniform vector. This defect is then transformed into a final score between 0 (perfectly anisotropic, like a line) and 1 (perfectly isotropic, like a sphere). IsoScore is calculated as

$$\iota(M) := \frac{(n - \delta(M)^2(n - \sqrt{n}))^2 - n}{n(n-1)} \tag{5.4}$$

where isotropy defect $\delta(M)$ is

$$\delta(M) := \frac{||\hat{\Sigma}_D - 1||}{\sqrt{2(n - \sqrt{n})}} \tag{5.5}$$

5.2.2 Experiments

All the embedding space metrics will be evaluated for all self-supervised models trained in Chapters 2 and 3 broadly following the protocol of [24]. For each data domain, SAR WV mode imagery and DXA scans, a sample of 10,000 images will be embedded by the model after each 10 epochs of training are completed. Embedding space metrics will be calculated for the resulting matrix and correlated with the final downstream task performance of that model. For the analysis in this dissertation, the α-ReQ score will be calculated as the absolute difference between the calculated decay coefficient α and 1, ($\alpha ReQ =\mid 1 - \alpha \mid$), with small difference being desirable. For WV imagery, downstream performance will consist of GOALI multiclass classification and wave height regression tasks, while for DXA, performance will be on the fracture and mortality prediction.

First, embedding quality at the final epoch will be evaluated against downstream task performance. Provided that any metrics perform well on this, secondarily correlation at each tenth epoch will be calculated for those metrics to test for reliability and stability. Correlation to downstream performance will be evaluated using Spearman rank correlation ρ.

5.3 Results

5.3.1 Final embedding quality

For both tasks, most embedding quality metrics fail. On the WV data (Figure 5.1) RankMe and α-ReQ perform best on both tasks. RankMe achieves a 0.65 Spearman correlation on the classification task and -0.74 correlation for the regression task, where negative correlation is desirable. Similarly,

α-ReQ achieves a -0.65 and 0.71 correlation on the classification and regression tasks, respectively, with the direction of correlation being inverted compared to RankMe, as is indicative of a good score. The condition number is also shown to be adequate at predicting downstream task performance.

Equally, for the DXA data, RankMe again performs well, achieving a 0.69 correlation on the mortality task and a 0.81 correlation on fracture prediction. On this data domain however, IsoScore also performs equally well, achieving strong positive correlations on both tasks and even outperforming RankMe a little on the fracture task (see Figure 5.2). Again, the condition number is moderately predictive but under-performing RankMe and IsoScore. Interestingly, while α-ReQ gets a moderate 0.4 Spearman correlation for the fracture task, it fails to capture any correlation to downstream performance for the mortality task.

Overall, this leaves RankMe and, to a lesser extent, the condition number as the two metrics that seem to be at least adequately correlated with downstream task performance across both domains and all three evaluation settings, regression, binary classification, and multi-label classification. As such, these will be analyzed more closely in the next section.

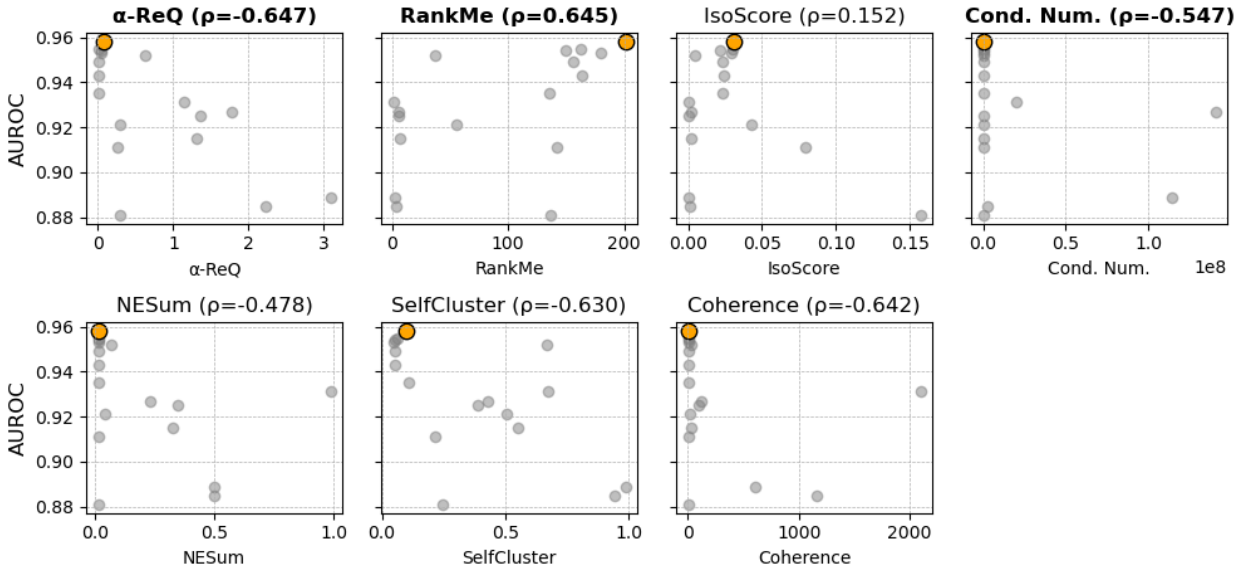
5.3.2 Embedding quality progression

When observing these scores over time, a drastically different behavior emerges. Embedding quality scores stay relatively consistent for both WV tasks, with RankMe only slightly moving in the right direction as training progresses. However, for the DXA data, we observe a more linearly increasing trend across both scores. RankMe and the condition number start with weak initial correlations for both tasks, but especially the hip fracture task, but steadily increase over the course of training. Especially for the hip fracture task, condition number initially shows no correlation at all but toward the end of training is close to caught up with RankMe.

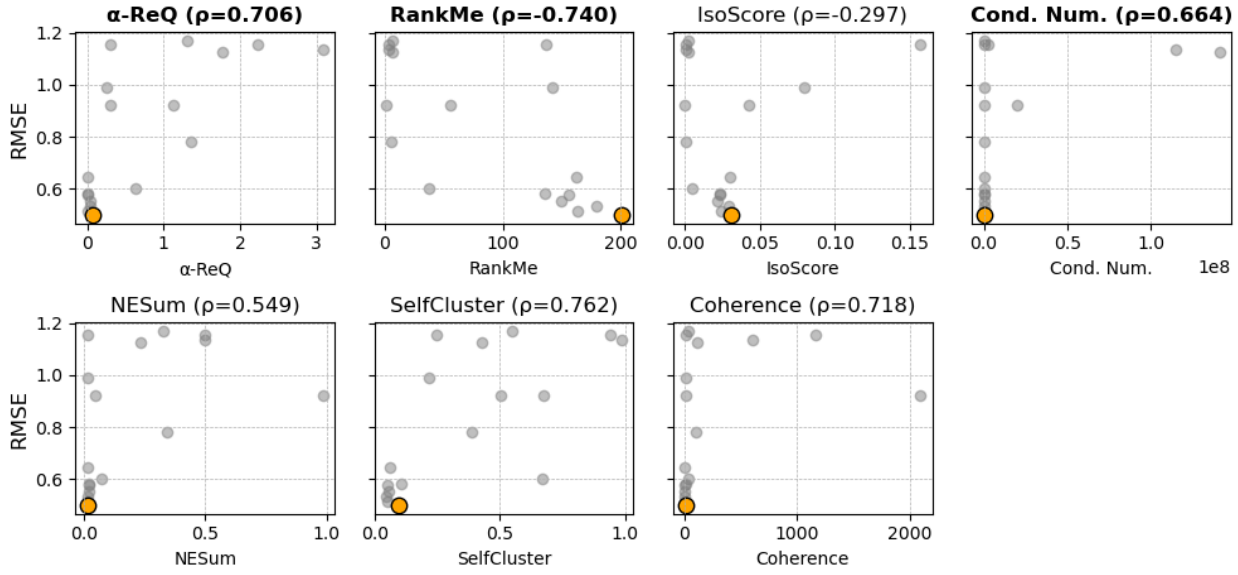
5.4 Discussion and Limitation

5.4.1 Final embedding quality

Overall, many of the failing scores are not entirely unsurprising. While α-ReQ has some amount of literature backing up its efficacy and is well rooted in empirical observations, many of the other scores are barely tested beyond their original work. α-ReQ was shown to be correlated to most tasks' downstream performance, but only weakly or not at all for the DXA data. This could indicate sensitivity to the data modality, pulling into question whether the empirical observations from the literature underpinning this metric hold for non-NI domains. IsoScore seems like it could be suffering from some outlying models in the WV tasks resulting in the bad performance compared to the DXA data, again potentially indicating that desirable embedding behavior in one domain (isotropy) may not translate to other domains.

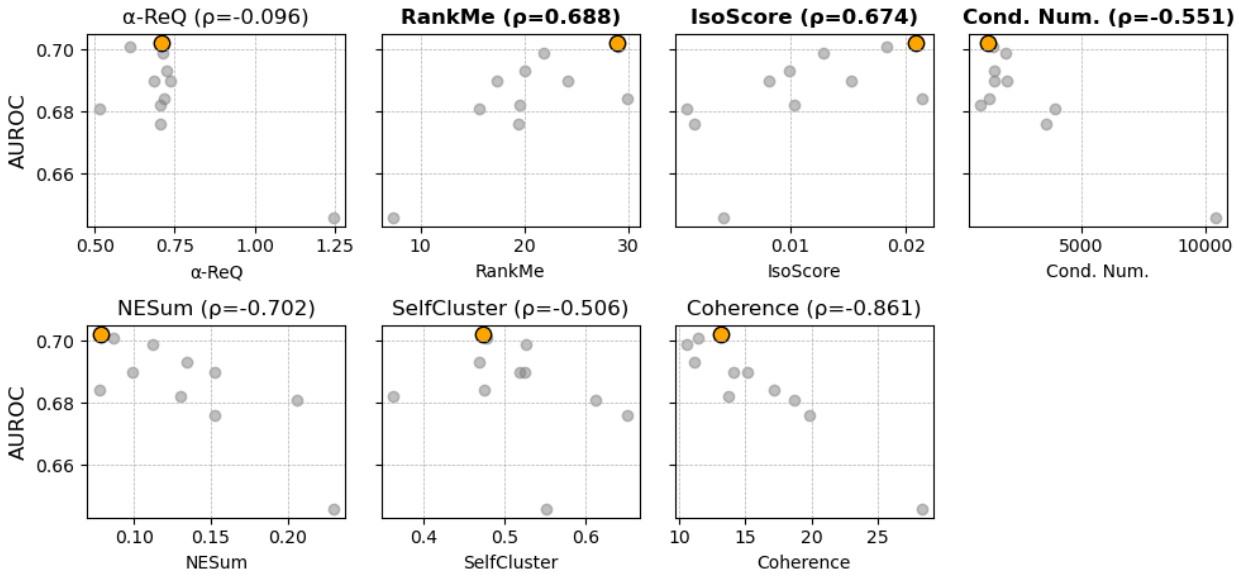


(a)

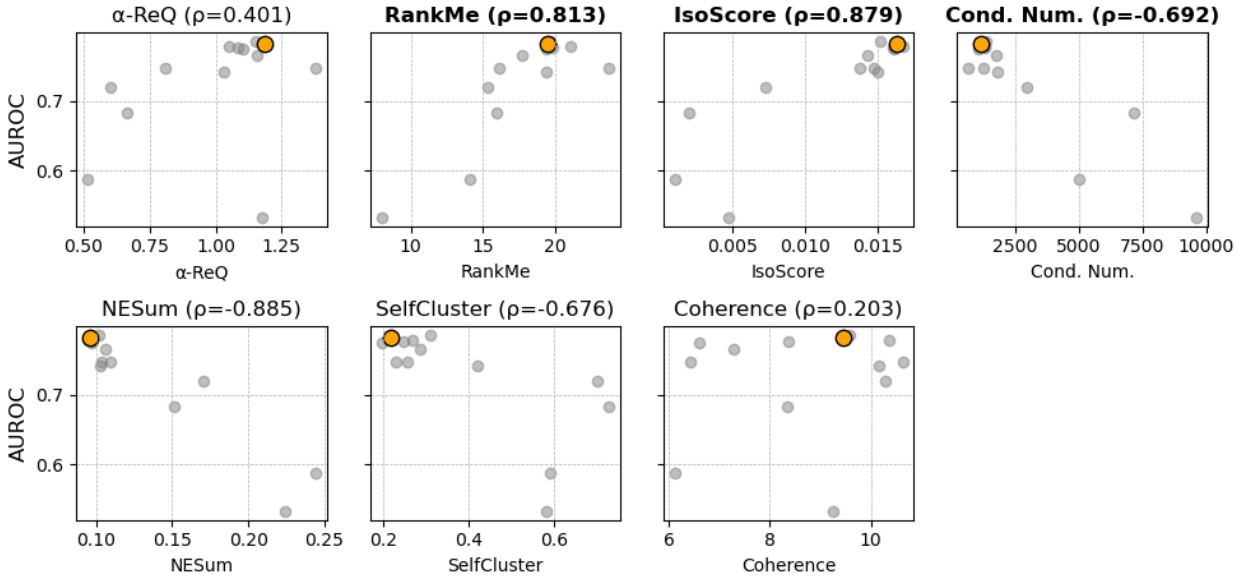


(b)

Figure 5.1: Embedding quality metric comparison on SAR WV data with the final model highlighted and well-performing quality metrics bolded. a) Correlation of different models' embedding scores and their final downstream task performance for GOALI multiclass classification. b) Wave height regression (negative correlations are better here).

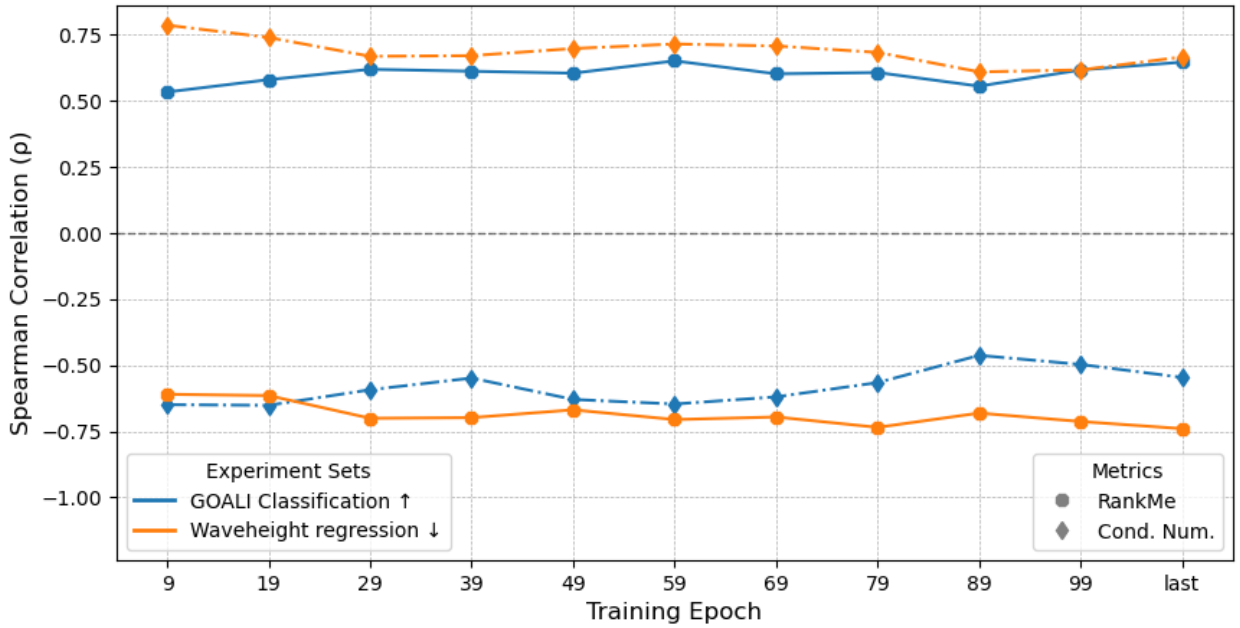


(a)

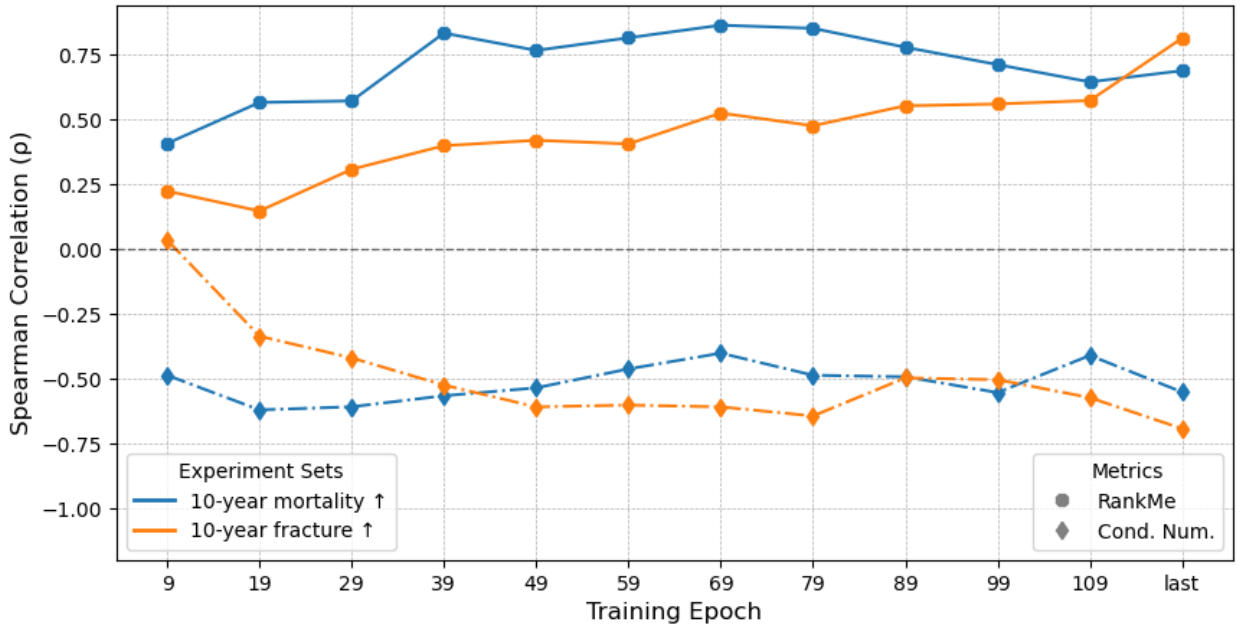


(b)

Figure 5.2: Embedding quality metric comparison on DXA data with the final model highlighted and well-performing quality metrics bolded. a) Correlation of different models' embedding scores and their final downstream task performance for mortality prediction. b) Hip fracture prediction.



(a)



(b)

Figure 5.3: Progression of correlation between embedding quality score and final model performance over the course of training for condition number and RankMe. a) shows WV data and b) DXA data

Since RankMe performed very well, it is also worth pointing out that the effective rank indicated by the score is very low, especially for the mortality models, utilizing less than 10% of the potential feature space; however, performance does not seem to suffer, and relative rank to other models is still highly indicative of comparative downstream performance. While the condition number also showed consistent correlation with downstream task performance, it was only weak and may not be enough to guide hyperparameter or modeling choices by itself. However, there is potential for further research into the notion of embedding stability as an indicator of quality.

Most interesting is the different behavior between the two data domains, with the WV data's embedding metrics staying very consistent throughout training and the DXA metrics displaying a near-linear upwards trend. There are several potential reasons for this. The WV dataset is evaluated on two non-standard tasks — regression and multilabel classification. Since these are uncommon evaluation settings these metrics might be less sensitive compared to classification, which is more commonly used when developing these metrics and would also explain the DXA data results. Most likely, however, is that this is indicative of instability of these metrics very early on in training. Since the DXA dataset is several orders of magnitude smaller than the WV dataset, one epoch for each model translates into a vastly different number of individual update steps. As such it could be that embedding quality metrics are more volatile early during training but eventually stabilize as training progresses.

Regardless, across both domains and all evaluation settings, RankMe does seem like a solid, dependable choice for evaluating embedding quality during training.

5.4.2 Limitations

This analysis could further benefit from evaluation on more downstream tasks. Training models for toy regression problems with the DXA data (e.g., weight regression) could help disentangle whether the downstream task or different dataset sizes are responsible for the divergent behavior during training.

Additionally, more frequent evaluation of the models and doing it on a schedule based on the number of update steps instead of epochs could further shed light on metric behavior and stability.

5.5 Conclusion

While most metrics exhibited no correlation with downstream performance at all, RankMe especially seems to robustly indicate embedding quality. While it might require some initial training period for embeddings to stabilize, it works across domains and downstream task settings. This makes it a viable candidate for the inclusion in self-supervised workflows. Based on the low effective rank estimates observed here, RankMe may best be employed in a comparative setting, such as when tuning a batch of models simultaneously or deciding between vision backbones, where relative

rank is a useful proxy for relative downstream performance.

CHAPTER 6
CONCLUSION

This dissertation studies the adaptation of self-supervised learning to medical imaging and remote sensing. Specifically, the dissertation focuses on SAR WV open-ocean satellite data and whole-body, proximal femur, and lumbar spine DXA scans. While self-supervised learning has been shown to be effective on natural images and beyond, adapting these frameworks to new domains can be a difficult process. Self-supervised learning frameworks are developed and tuned on natural images, images that have three color channels, perspective, are often object-centric, and are of objects seen in the natural world. These frameworks rely heavily on implicit biases of the data; therefore, successfully adapting them to domain data that does not share most of these characteristics can be challenging.

This work set out to answer, first, whether self-supervised learning can be used effectively for SAR WV data and DXA scans, improving over traditional pre-training approaches that rely on labeled natural image datasets. The next question is whether self-supervised frameworks, once successfully applied to these domains, can further be finetuned to better fit characteristics of the domain and downstream applications. To answer the first question, a pre-trained ImageNet model was compared with models pre-trained using SimCLR, BYOL, and MAE, three of the most common self-supervised pre-training frameworks in computer vision. In both domains, a marked improvement was observable, showing that self-supervised learning, out of the box, is indeed capable of improving over transfer learning from ImageNet pre-trained models. To answer the second question, the best framework and backbone combination for both tasks was then adapted. In the case of the SAR model, several augmentation strategies were introduced and removed from the augmentation pool, resulting in the WV-Net configuration, with two added augmentation policies that yield another improvement over the new self-supervised baseline. For DXA, experiments were conducted with augmentation policies, feature engineering, and modified loss functions. Ultimately, one domain-inspired additional feature in the ratio channel, as well as a modified loss function and an extra augmentation were added to, again, improve over the new self-supervised baseline.

In both cases, the resulting models outperformed any previous benchmarks on downstream tasks. Further, both models can serve as foundation models in their respective domains — they are more labeled-data-efficient than supervised counterparts parts, more robust to hyperparameter choices, can be adapted without any end-to-end finetuning, and are highly generalizable to downstream tasks. These results clearly illustrate the power of self-supervised learning, especially for domains with highly specialized data. However, the tuning process also illustrates a difficulty with self-supervised learning — even in the presence of a large enough dataset to make this a viable approach, it often requires prohibitive time and compute resources. Having to train every model to completion to evaluate it on downstream tasks is highly inefficient but unfortunately a necessity in

the absence of good performance indicators. This results in models being under-tuned and performance being left suboptimal because of time or compute restraints. This is why this dissertation also analyzed a third question: Are there any unsupervised embedding methods that exhibit good generalization to uncommon downstream tasks such as regression and multiclass classification and robustness to image domains with unusual characteristics? To that end, embedding quality metrics from the literature were evaluated on the models trained for the SAR and DXA tasks. Out of 7 evaluated metrics, only RankMe showed robustness to both data domain and downstream tasks. Higher RankMe scores were consistently associated with better final model performance across highly variable training setups for both tasks. As such it seems like a viable choice for model selection and hyperparameter tuning for similar future efforts.

Overall this dissertation answered the questions it set out to answer, demonstrating that self-supervised learning is the most promising path toward generalizable, domain-specific embedding models that have direct implications on how effectively science can be done in downstream applications.

6.1 Contributions

The two primary contributions of this work are the pretrained models for either domain. While the SAR WV-Net is already publicly available, the DXA model will similarly be made available with the publication of the associated work. The goal is for this work to empower researchers and practitioners by giving them tools to quickly analyze large backlogs of data, train powerful models with little compute and labeling budget, and maybe build on these models to drive the field forward.

Further, the downstream task models also have direct implications on their respective fields. The GOALI dataset and accompanying multiclass classification model will be the subject of a future manuscript and both, the hip fracture and mortality models have implications for opportunistic screening that are worth pursuing.

Additionally, the results presented in this work have more indirect implications. RankMe, already one of the most prevalent embedding space metrics, now has further evidence for its domain-spanning utility as an embedding quality estimator. As such, future self-supervised projects in related domains, such as other medical imaging modalities or SAR satellite images should feel more confident in using RankMe to improve their workflow. Having a reliable performance indicator early during training can significantly improve efficiency and can help researchers make better choices about what research avenues to pursue. Similarly, the workflow for training self-supervised models is also transferable. First, optimizing the backbone and framework combination against an ImageNet baseline gives a good idea of performance capabilities while narrowing down future modeling choices. Afterward, altering augmentation policies, framework parameters, and other model characteristics separately reduces the overall search space while evidently still giving a good

indicator of performance. As such, maybe the strategy employed in this dissertation can be followed by future researchers who are also under resource constraints.

6.2 Future Work

This dissertation is heavily focused on the work associated with creating the foundation models, but future work will be focused on their downstream applications. Classifying atmospheric phenomena and reanalyzing over a decade of global backlog with the GOALI model is an immediate next step. More deliberate finetuning for all the downstream tasks is also a viable area of research as this was omitted for this work due to compute constraints. Since the model is already trained to a large degree exploratory analysis of downstream applications is relatively easy as long as a minimal amount of labeled data is available. Moreover, expanding this work to other satellite platforms such as the NASA Surface Water and Ocean Topography (SWOT) platform is a promising research direction.

Similarly, for the DXA data, the focus will shift to more fully exploring downstream applications. Beyond the applications explored in this model, there is also considerable interest in biological markers of aging. While these are often associated with epigenetic indicators, a body-composition-based marker of aging is also an interesting research direction given that the model is already very capable at predicting all-cause mortality.

BIBLIOGRAPHY

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. 1, 10, 17

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1, 17

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020. 1, 4, 10, 17, 19, 22, 23, 25, 26, 36, 42, 52

[4] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 1, 4, 9, 22, 36

[5] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193. 1, 12

[6] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 1, 9

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1, 11

[8] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen,

Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2022. URL https://arxiv.org/abs/2108.07258. 1

[9] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer, 2022. URL https://arxiv.org/abs/2111.07832. 1, 12, 17

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 4, 12, 16, 39

[11] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images, 2009. 1

[12] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop, 2016. URL https://arxiv.org/abs/1506.03365. 1

[13] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017. 1, 12

[14] Priya Goyal, Mathilde Caron, Benjamin Lefaudeux, Min Xu, Pengchao Wang, Vivek Pai, Mannat Singh, Vitaliy Liptchinsky, Ishan Misra, Armand Joulin, and Piotr Bojanowski. Self-

supervised pretraining of visual features in the wild, 2021. URL https://arxiv.org/abs/2103.01988. 1, 2, 12

[15] Elijah Cole, Xuan Yang, Kimberly Wilber, Oisin Mac Aodha, and Serge Belongie. When does contrastive visual representation learning work? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14755–14764, 2022. 1, 13

[16] Florian Bordes, Randall Balestriero, and Pascal Vincent. Towards democratizing joint-embedding self-supervised learning, 2023. URL https://arxiv.org/abs/2303.01986. 1, 2, 13, 25, 36, 44, 52, 55

[17] Lambert T Leong, Michael C Wong, Yong E Liu, Yannik Glaser, Brandon K Quon, Nisa N Kelly, Devon Cataldi, Peter Sadowski, Steven B Heymsfield, and John A Shepherd. Generative deep learning furthers the understanding of local distributions of fat and muscle on body shape and health using 3d surface scans. *Communications Medicine*, 4(1):13, 2024. 1, 8

[18] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 1, 11, 17, 25

[19] Alex Tamkin, Mike Wu, and Noah Goodman. Viewmaker networks: Learning views for unsupervised representation learning, 2021. URL https://arxiv.org/abs/2010.07432. 1

[20] Priya Goyal, Quentin Duval, Isaac Seessel, Mathilde Caron, Ishan Misra, Levent Sagun, Armand Joulin, and Piotr Bojanowski. Vision models are more robust and fair when pretrained on uncurated images without supervision, 2022. URL https://arxiv.org/abs/2202.08360. 2

[21] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020. 2, 36

[22] Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann Lecun. Rankme: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *International conference on machine learning*, pages 10929–10974. PMLR, 2023. 2, 4, 14, 55, 56

[23] Kumar K Agrawal, Arnab Kumar Mondal, Arna Ghosh, and Blake Richards. α-ReQ: Assessing representation quality in self-supervised learning by measuring eigenspectrum decay. *Advances in Neural Information Processing Systems*, 35:17626–17638, 2022. 2, 4, 15, 56

[24] Anton Tsitsulin, Marina Munkhoeva, and Bryan Perozzi. Unsupervised embedding quality evaluation. In *Topological, Algebraic and Geometric Learning Workshops 2023*, pages 169–188. PMLR, 2023. 2, 15, 56, 57

[25] Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34: 16451–16467, 2021. 2

[26] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning, 2022. URL https://arxiv.org/abs/2105.04906. 2, 12

[27] Florian Bordes, Randall Balestriero, and Pascal Vincent. High fidelity visualization of what your self-supervised representation knows about, 2022. URL https://arxiv.org/abs/2112.09164. 2

[28] Florian Bordes, Randall Balestriero, Quentin Garrido, Adrien Bardes, and Pascal Vincent. Guillotine regularization: Why removing layers is needed to improve generalization in self-supervised learning, 2023. URL https://arxiv.org/abs/2206.13378. 2, 13

[29] Mark Ibrahim, David Klindt, and Randall Balestriero. Occam's razor for self supervised learning: What is sufficient to learn good representations?, 2024. URL https://arxiv.org/abs/2406.10743. 2, 11, 13, 55

[30] Raphael Gontijo-Lopes, Sylvia J. Smullin, Ekin D. Cubuk, and Ethan Dyer. Affinity and diversity: Quantifying mechanisms of data augmentation, 2020. URL https://arxiv.org/abs/2002.08973. 2

[31] Ramon Torres, Paul Snoeij, Dirk Geudtner, David Bibby, Malcolm Davidson, Evert Attema, Pierre Potin, BjÖrn Rommen, Nicolas Floury, Mike Brown, et al. Gmes sentinel-1 mission. *Remote sensing of environment*, 120:9–24, 2012. 3, 17

[32] Sanja Šćepanović, Oleg Antropov, Pekka Laurila, Yrjo Rauste, Vladimir Ignatenko, and Jaan Praks. Wide-area land cover mapping with sentinel-1 imagery using deep learning semantic segmentation models. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:10357–10374, 2021. 3

[33] Umangi Jain, Alex Wilson, and Varun Gulshan. Multimodal contrastive learning for remote sensing tasks, 2022. URL https://arxiv.org/abs/2209.02329. 3

[34] Timothy Mayer, Ate Poortinga, Biplov Bhandari, Andrea P Nicolau, Kel Markert, Nyein Soe Thwal, Amanda Markert, Arjen Haag, John Kilbride, Farrukh Chishtie, et al. Deep learning

approach for sentinel-1 surface water mapping leveraging google earth engine. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 2:100005, 2021. 3

[35] Fernando Paolo, Tsu-ting Tim Lin, Ritwik Gupta, Bryce Goodman, Nirav Patel, Daniel Kuster, David Kroodsma, and Jared Dunnmon. xview3-sar: Detecting dark fishing activity using synthetic aperture radar imagery. *Advances in Neural Information Processing Systems*, 35:37604–37616, 2022. 3

[36] European Space Agency. Overview of s1 applications, . URL https://sentiwiki.copernicus.eu/web/s1-applications#S1Applications-ShipMonitoringS1-Applications-Ship-Monitoring. 3

[37] Chen Wang, Alexis Mouche, Pierre Tandeo, Justin E Stopa, Nicolas Longépé, Guillaume Erhard, Ralph C Foster, Douglas Vandemark, and Bertrand Chapron. A labelled ocean sar imagery dataset of ten geophysical phenomena from sentinel-1 wave mode. *Geoscience Data Journal*, 6(2):105–115, 2019. 3, 7, 17, 18, 19, 24

[38] Christian Kruse, Pia Eiken, and Peter Vestergaard. Machine learning principles can improve hip fracture prediction. *Calcified tissue international*, 100(4):348–360, 2017. 3

[39] Namki Hong, Sang Wouk Cho, Sungjae Shin, Seunghyun Lee, Seol A Jang, Seunghyun Roh, Young Han Lee, Yumie Rhee, Steven R Cummings, Hwiyoung Kim, et al. Deep-learning-based detection of vertebral fracture and osteoporosis using lateral spine x-ray radiography. *Journal of Bone and Mineral Research*, 38(6):887–895, 2020. 3, 8

[40] Yannik Glaser, Peter Sadowski, Thomas Wolfgruber, Li-Yung Lui, Steven Cummings, and John Shepherd. Hip fracture risk modeling using dxa and artificial intelligence. In *Journal of Bone and Mineral Research*, volume 35, pages 200–200. John Wiley and Sons and The American Society for Bone and Mineral Research, 2020. 3, 8

[41] John A Shepherd, Bennett K Ng, Markus J Sommer, and Steven B Heymsfield. Body composition by dxa. *Bone*, 104:101–105, 2017. 3, 8

[42] Thomas J Beck. Extending dxa beyond bone mineral density: understanding hip structure analysis. *Current osteoporosis reports*, 5(2):49–55, 2007. 3

[43] John A Shepherd, Bo Fan, Ying Lu, Xiao P Wu, Wynn K Wacker, David L Ergun, and Michael A Levine. A multinational study to develop universal standardization of whole-body bone density and composition using ge healthcare lunar and hologic dxa systems. *Journal of Bone and Mineral Research*, 27(10):2208–2216, 2012. 3

[44] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 4, 10, 11, 26, 42

[45] European Space Agency. Overview of sentinel-1 mission, . URL https://sentiwiki.copernicus.eu/web/s1-mission#S1Mission-WaveS1-Mission-Wave. 7

[46] Glen M Blake and Ignac Fogelman. The role of dxa bone density scans in the diagnosis and treatment of osteoporosis. *Postgraduate medical journal*, 83(982):509–517, 2007. 7, 8

[47] JCK Wells and MS Fewtrell. Measuring body composition. *Archives of disease in childhood*, 91(7):612–617, 2006. 8

[48] Kristine E Ensrud, John T Schousboe, Carolyn J Crandall, William D Leslie, Howard A Fink, Peggy M Cawthon, Deborah M Kado, Nancy E Lane, Jane A Cauley, and Lisa Langsetmo. Hip fracture risk assessment tools for adults aged 80 years and older. *JAMA Network Open*, 7(6):e2418612–e2418612, 2024. 8

[49] WF Lems, J Paccou, Jean Zhang, NR Fuggle, M Chandran, NC Harvey, Cyrus Cooper, K Javaid, S Ferrari, Kristina E Akesson, et al. Vertebral fracture: epidemiology, impact and use of dxa vertebral fracture assessment in fracture liaison services. *Osteoporosis international*, 32:399–411, 2021. 8

[50] Yannik Glaser, John Shepherd, Lambert Leong, Thomas Wolfgruber, Li-Yung Lui, Peter Sadowski, and Steven R Cummings. Deep learning predicts all-cause mortality from longitudinal total-body dxa imaging. *Communications medicine*, 2(1):102, 2022. 8, 41, 47

[51] Erik L. Ridley. Ai, whole-body dexa scans reveal patient age, mortality risk, Dec 2021. URL https://www.auntminnie.com/imaging-informatics/advanced-visualization/image-processing/article/15630064/ai-wholebody-dexa-scans-reveal-patient-age-mortality-risk. 8

[52] Judith E Adams. Opportunistic identification of vertebral fractures. *Journal of Clinical Densitometry*, 19(1):54–62, 2016. 8

[53] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. *Advances in neural information processing systems*, 19, 2006. 9

[54] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006. 9

[55] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008. 9

[56] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010. 9

[57] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. URL https://arxiv.org/abs/1312.6114. 9

[58] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *International conference on machine learning*, pages 1083–1092. PMLR, 2015. 9, 11

[59] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1058–1067, 2017. 9

[60] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 577–593. Springer, 2016. 9

[61] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 649–666. Springer, 2016. 9

[62] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 9

[63] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929. 9, 26, 42

[64] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers, 2022. URL https://arxiv.org/abs/2106.08254. 9

[65] Pulkit Agrawal, Joao Carreira, and Jitendra Malik. Learning to see by moving. In *Proceedings of the IEEE international conference on computer vision*, pages 37–45, 2015. 10

[66] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, 6, 1993. 10

[67] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005. 10

[68] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006. 10

[69] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2015. 10

[70] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016. 10

[71] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 10, 24

[72] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021. 10, 11

[73] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32, 2019. 10

[74] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 11

[75] Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsiavash. Mean shift for self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10326–10335, 2021. 11

[76] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149, 2018. 11

[77] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning, 2020. URL https://arxiv.org/abs/1911.05371. 11

[78] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020. 11, 26

[79] Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, Avi Schwarzschild, Andrew Gordon Wilson, Jonas Geiping, Quentin Garrido, Pierre Fernandez, Amir Bar, Hamed Pirsiavash, Yann LeCun, and Micah Goldblum. A cookbook of self-supervised learning, 2023. URL https://arxiv.org/abs/2304.12210. 11, 13, 26, 44, 55

[80] Harold Hotelling. Relations between two sets of variates. In *Breakthroughs in statistics: methodology and distribution*, pages 162–190. Springer, 1992. 11

[81] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pages 12310–12320. PMLR, 2021. 11

[82] Randall Balestriero and Yann LeCun. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. *Advances in Neural Information Processing Systems*, 35:26671–26685, 2022. 12

[83] Quentin Garrido, Yubei Chen, Adrien Bardes, Laurent Najman, and Yann Lecun. On the duality between contrastive and non-contrastive self-supervised learning. 2022. URL https://arxiv.org/abs/2206.02574. 12

[84] Chenxin Tao, Honghui Wang, Xizhou Zhu, Jiahua Dong, Shiji Song, Gao Huang, and Jifeng Dai. Exploring the equivalence of siamese self-supervised learning via a unified gradient framework. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14431–14440, 2022. 12

[85] Manu Srinath Halvagal, Axel Laborieux, and Friedemann Zenke. Implicit variance regularization in non-contrastive ssl. *Advances in Neural Information Processing Systems*, 36: 63409–63436, 2023. 12

[86] Weiran Huang, Mingyang Yi, Xuyang Zhao, and Zihao Jiang. Towards the generalization of contrastive self-supervised learning, 2023. URL https://arxiv.org/abs/2111.00743. 12

[87] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2661–2671, 2019. 12

[88] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE international conference on computer vision*, pages 2938–2946, 2015. 12

[89] Alexander Mathis, Pranav Mamidanna, Kevin M Cury, Taiga Abe, Venkatesh N Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature neuroscience*, 21(9):1281–1289, 2018. 12

[90] Yuan Liu, Ayush Jain, Clara Eng, David H Way, Kang Lee, Peggy Bui, Kimberly Kanada, Guilherme de Oliveira Marinho, Jessica Gallegos, Sara Gabriele, et al. A deep learning system for differential diagnosis of skin diseases. *Nature medicine*, 26(6):900–908, 2020. 12

[91] Neal Jean, Marshall Burke, Michael Xie, W Matthew Alampay Davis, David B Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016. 12

[92] Yunhui Guo, Noel C Codella, Leonid Karlinsky, James V Codella, John R Smith, Kate Saenko, Tajana Rosing, and Rogerio Feris. A broader study of cross-domain few-shot learning. In *European conference on computer vision*, pages 124–141. Springer, 2020. 12, 13, 16

[93] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019. 13

[94] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019. 13

[95] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do cifar-10 classifiers generalize to cifar-10?, 2018. URL https://arxiv.org/abs/1806.00451. 13

[96] Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, et al. Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3478–3488, 2021. 13

[97] Shekoofeh Azizi, Laura Culp, Jan Freyberg, Basil Mustafa, Sebastien Baur, Simon Kornblith, Ting Chen, Nenad Tomasev, Jovana Mitrović, Patricia Strachan, et al. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nature Biomedical Engineering*, 7(6):756–779, 2023. 13, 39, 43

[98] Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature biomedical engineering*, 6(12):1399–1406, 2022. 13

[99] Zhixiu Lu, Hailong Li, Nehal A Parikh, Jonathan R Dillman, and Lili He. Radclip: Enhancing radiologic image analysis through contrastive language–image pretraining. *IEEE Transactions on Neural Networks and Learning Systems*, 2025. 13

[100] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, 2025. URL https://arxiv.org/abs/2303.00915. 13

[101] Shih-Cheng Huang, Anuj Pareek, Malte Jensen, Matthew P Lungren, Serena Yeung, and Akshay S Chaudhari. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *NPJ Digital Medicine*, 6(1):74, 2023. 13

[102] Rhydian Windsor, Amir Jamaludin, Timor Kadir, and Andrew Zisserman. Self-supervised multi-modal alignment for whole body medical imaging. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, pages 90–101. Springer, 2021. 13

[103] Ung Hwang, Chang-Hun Lee, and Kijung Yoon. Osteoporosis prediction from hand x-ray images using segmentation-for-classification and self-supervised learning, 2024. URL https://arxiv.org/abs/2412.05345. 13

[104] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35: 197–211, 2022. 14, 17, 22

[105] Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings*

of the IEEE/CVF International Conference on Computer Vision, pages 4088–4099, 2023. 14, 17

[106] J Xavier Prochaska, Erdong Guo, Peter C Cornillon, and Christian E Buckingham. The fundamental patterns of sea surface temperature. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–19, 2023. 14

[107] Thomas Kerdreux, Alexandre Tuel, Quentin Febvre, Alexis Mouche, and Bertrand Chapron. Efficient self-supervised learning for earth observation via dynamic dataset curation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3017–3027, 2025. 14

[108] Heechul Jung, Yoonju Oh, Seongho Jeong, Chaehyeon Lee, and Taegyun Jeon. Contrastive self-supervised learning with smoothed representation for remote sensing. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021. 14

[109] Haohua Dong, Yutaro Iwamoto, Xianhua Han, Lanfen Lin, Hongjie Hu, Xiujun Cai, and Yen-Wei Chen. Case discrimination: self-supervised feature learning for the classification of focal liver lesions. In *Innovation in Medicine and Healthcare: Proceedings of 9th KES-InMed 2021*, pages 241–249. Springer, 2021. 14

[110] Shih-Cheng Huang, Anuj Pareek, Malte Jensen, Matthew P Lungren, Serena Yeung, and Akshay S Chaudhari. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *NPJ Digital Medicine*, 6(1):74, 2023. 14

[111] Burhaneddin Yaman, Seyed Amir Hossein Hosseini, Steen Moeller, Jutta Ellermann, Kâmil Uğurbil, and Mehmet Akçakaya. Self-supervised learning of physics-guided reconstruction neural networks without fully sampled reference data. *Magnetic resonance in medicine*, 84 (6):3172–3191, 2020. 14

[112] Huaijin Chen, Jinwei Gu, Orazio Gallo, Ming-Yu Liu, Ashok Veeraraghavan, and Jan Kautz. Reblur2deblur: Deblurring videos via self-supervised learning. In *2018 IEEE International Conference on Computational Photography (ICCP)*, pages 1–9. IEEE, 2018. 14

[113] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR, 2020. 14

[114] William Rudman, Nate Gillman, Taylor Rayne, and Carsten Eickhoff. Isoscore: Measuring the uniformity of embedding space utilization. In *Findings of the Association for Computational Linguistics: ACL 2022*, page 3325–3339. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-acl.262. URL http://dx.doi.org/10.18653/v1/2022.findings-acl.262. 14, 55, 57

[115] Bobby He and Mete Ozay. Exploring the gap between collapsed & whitened features in self-supervised learning. In *International Conference on Machine Learning*, pages 8613–8634. PMLR, 2022. 15, 56

[116] Arna Ghosh, Arnab Kumar Mondal, Kumar Krishna Agrawal, and Blake Richards. Investigating power laws in deep representation learning, 2022. URL https://arxiv.org/abs/2202.05808. 15, 56

[117] Andrew J Tatem, Scott J Goetz, and Simon I Hay. Fifty years of earth-observation satellites: Views from space have led to countless advances on the ground in both scientific knowledge and daily life. *American scientist*, 96(5):390–398, 2008. 16

[118] Union of Concerned Scientists, May 2022. URL https://www.ucsusa.org/resources/satellite-database. 16

[119] Ramon Torres, Paul Snoeij, Dirk Geudtner, David Bibby, Malcolm Davidson, Evert Attema, Pierre Potin, BjÖrn Rommen, Nicolas Floury, Mike Brown, et al. Gmes sentinel-1 mission. *Remote sensing of environment*, 120:9–24, 2012. 16, 18

[120] Konstantinos Topouzelis and Dimitra Kitsiou. Detection and classification of mesoscale atmospheric phenomena above sea in sar imagery. *Remote sensing of environment*, 160:263–272, 2015. 16

[121] Chen Wang, Pierre Tandeo, Alexis Mouche, Justin E Stopa, Victor Gressani, Nicolas Longepe, Douglas Vandemark, Ralph C Foster, and Bertrand Chapron. Classification of the global sentinel-1 sar vignettes for ocean surface process studies. *Remote Sensing of Environment*, 234:111457, 2019. 16

[122] Fan Hu, Gui-Song Xia, Jingwen Hu, and Liangpei Zhang. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. *Remote Sensing*, 7(11):14680–14707, 2015. 16

[123] Chen Wang, Alexis Mouche, Pierre Tandeo, Justin Stopa, Bertrand Chapron, Ralph Foster, and Douglas Vandemark. Automated geophysical classification of sentinel-1 wave mode sar images through deep-learning. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 1776–1779. IEEE, 2018. 16

[124] Keiller Nogueira, Otávio AB Penatti, and Jefersson A Dos Santos. Towards better exploiting convolutional neural networks for remote sensing scene classification. *Pattern Recognition*, 61:539–556, 2017. 16

[125] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014. 16, 37

[126] Lambert T. Leong, Michael C. Wong, Yannik Glaser, Thomas Wolfgruber, Steven B. Heymsfield, Peter Sadowski, and John A. Shepherd. Quantitative imaging principles improves medical image learning, 2022. URL https://arxiv.org/abs/2206.06663. 16, 43

[127] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. *Advances in neural information processing systems*, 32, 2019. 16

[128] Esther Rolf, Konstantin Klemmer, Caleb Robinson, and Hannah Kerner. Mission critical – satellite data is a distinct modality in machine learning, 2024. URL https://arxiv.org/abs/2402.01444. 16

[129] Kenneth Ward Church. Word2Vec. *Natural Language Engineering*, 23(1):155–162, 2017. 17

[130] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 17

[131] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training, 2018. 17

[132] Maxim Neumann, Andre Susano Pinto, Xiaohua Zhai, and Neil Houlsby. In-domain representation learning for remote sensing, 2019. URL https://arxiv.org/abs/1911.06721. 17

[133] Anthony Fuller, Koreen Millard, and James R Green. Satvit: Pretraining transformers for earth observation. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022. 17

[134] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10181–10190, 2021. 17

[135] Klaus Hasselmann, B Chapron, L Aouf, F Ardhuin, F Collard, G Engen, Susanne Hasselmann, P Heimbach, PAEM Janssen, H Johnsen, et al. The ers sar wave mode: A breakthrough in global ocean wave observations. *ESA Communications*, 2013. 17

[136] Justin E Stopa and Alexis Mouche. Significant wave heights from s entinel-1 sar: Validation and applications. *Journal of Geophysical Research: Oceans*, 122(3):1827–1848, 2017. 17

[137] Brandon Quach, Yannik Glaser, Justin Edward Stopa, Alexis Aurélien Mouche, and Peter Sadowski. Deep learning for predicting significant wave height from synthetic aperture radar. *IEEE Transactions on Geoscience and Remote Sensing*, 59(3):1859–1867, 2020. 17, 19

[138] Chris R. Jackson, John R. Apel, Pablo Clemente-Colón, William G. Pichel, Robert A. Shuchman, and Christopher C. Wackerman. Synthetic aperture radarmarine user's manual. Technical report, National Oceanic and Atmospheric Administration, September 2004. 17

[139] Nicolas Rascle, Jeroen Molemaker, Louis Marié, Frédéric Nouguier, Bertrand Chapron, Björn Lund, and Alexis Mouche. Intense deformation field at oceanic front inferred from directional sea surface roughness observations. *Geophysical Research Letters*, 44(11):5599–5608, 2017. 17

[140] George S Young, Todd D Sikora, and Nathaniel S Winstead. Inferring marine atmospheric boundary layer properties from spectral characteristics of satellite-borne sar imagery. *Monthly weather review*, 128(5):1506–1520, 2000. 17

[141] D Vandemark, PD Mourad, SA Bailey, TL Crawford, CA Vogel, J Sun, and Bertrand Chapron. Measured changes in ocean surface roughness due to atmospheric boundary layer rolls. *Journal of Geophysical Research: Oceans*, 106(C3):4639–4654, 2001. 17

[142] Justin E Stopa, Chen Wang, Doug Vandemark, Ralph Foster, Alexis Mouche, and Bertrand Chapron. Automated global classification of surface layer stratification using high-resolution sea surface roughness measurements by satellite synthetic aperture radar. *Geophysical Research Letters*, 49(12):e2022GL098686, 2022. 17, 19, 36

[143] Justin E Stopa, Fabrice Ardhuin, Bertrand Chapron, and Fabrice Collard. Estimating wave orbital velocity through the azimuth cutoff from space-borne satellites. *Journal of Geophysical Research: Oceans*, 120(11):7616–7634, 2015. 18

[144] Hans Hersbach. Comparison of c-band scatterometer cmod5. n equivalent neutral winds with ecmwf. *Journal of Atmospheric and Oceanic Technology*, 27(4):721–736, 2010. 18

[145] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly journal of the royal meteorological society*, 146(730):1999–2049, 2020. 19

[146] James B Edson, Venkata Jampana, Robert A Weller, Sebastien P Bigorre, Albert J Plueddemann, Christopher W Fairall, Scott D Miller, Larry Mahrt, Dean Vickers, and Hans Hersbach. On the exchange of momentum over the open ocean. *Journal of Physical Oceanography*, 43 (8):1589–1610, 2013. 19

[147] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 19, 26, 42

[148] Terrance DeVries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout, 2017. URL https://arxiv.org/abs/1708.04552. 22, 23, 27, 43

[149] Salman H Khan, Munawar Hayat, and Fatih Porikli. Regularization of deep neural networks with spectral dropout. *Neural Networks*, 110:82–90, 2019. 23

[150] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization, 2018. URL https://arxiv.org/abs/1710.09412. 23, 43

[151] Sungnyun Kim, Gihun Lee, Sangmin Bae, and Se-Young Yun. Mixco: Mix-up contrastive learning for visual representation, 2020. URL https://arxiv.org/abs/2010.06300. 23

[152] Zhiqiang Shen, Zechun Liu, Zhuang Liu, Marios Savvides, Trevor Darrell, and Eric Xing. Un-mix: Rethinking image mixtures for unsupervised visual representation learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 2216–2224, 2022. 23, 43

[153] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. *Advances in neural information processing systems*, 33:21798–21809, 2020. 23

[154] Vikas Verma, Thang Luong, Kenji Kawaguchi, Hieu Pham, and Quoc Le. Towards domain-agnostic contrastive learning. In *International Conference on Machine Learning*, pages 10530–10541. PMLR, 2021. 23, 43

[155] Muen Jin and Michael Heizmann. Cutout as augmentation in contrastive learning for detecting burn marks in plastic granules. *Journal of Sensors and Sensor Systems*, 13(1):63–69, 2024. 23

[156] Abien Fred Agarap. Deep learning using rectified linear units (ReLU), 2019. URL https://arxiv.org/abs/1803.08375. 25

[157] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980. 25

[158] Kai Norman Clasen, Leonard Hackel, Tom Burgert, Gencer Sumbul, Begüm Demir, and Volker Markl. reben: Refined bigearthnet dataset for remote sensing image analysis, 2025. URL https://arxiv.org/abs/2407.03653. 26, 29

[159] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 26, 42

[160] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021. 36

[161] TD Sikora, GS Young, RC Beal, and JB Edson. Use of spaceborne synthetic aperture radar imagery of the sea surface in detecting the presence and structure of the convective marine atmospheric boundary layer. *Monthly Weather Review*, 123(12):3623–3632, 1995. 36

[162] George S Young, David AR Kristovich, Mark R Hjelmfelt, and Ralph C Foster. Supplement to rolls, streets, waves, and more. *Bulletin of the American Meteorological Society*, 83(7), 2002. 36

[163] George S Young, TN Sikora, and Nathaniel S Winstead. Use of synthetic aperture radar in finescale surface analysis of synoptic-scale fronts at sea. *Weather and forecasting*, 20(3): 311–327, 2005. 36

[164] James A Yoder, Steven G Ackleson, Richard T Barber, Pierre Flament, and William M Balch. A line in the sea. *Nature*, 371(6499):689–692, 1994. 36

[165] Anne B Newman, Varant Kupelian, Marjolein Visser, Eleanor M Simonsick, Bret H Goodpaster, Stephen B Kritchevsky, Frances A Tylavsky, Susan M Rubin, and Tamara B Harris. Strength, but not muscle mass, is associated with mortality in the health, aging and body composition study cohort. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 61(1):72–77, 2006. 38

[166] Adam J Santanasto, Bret H Goodpaster, Stephen B Kritchevsky, Iva Miljkovic, Suzanne Satterfield, Ann V Schwartz, Steven R Cummings, Robert M Boudreau, Tamara B Harris, and Anne B Newman. Body composition remodeling and mortality: the health aging and body composition study. *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences*, 72(4):513–519, 2017. 38

[167] B Gullberg, O Johnell, and JA Kanis. World-wide projections for hip fracture. *Osteoporosis international*, 7(5):407–413, 1997. 38

[168] Chor-Wing Sing, Tzu-Chieh Lin, Sharon Bartholomew, J Simon Bell, Corina Bennett, Kebede Beyene, Pauline Bosco-Levy, Brian D Bradbury, Amy Hai Yan Chan, Manju Chandran, et al. Global epidemiology of hip fractures: secular trends in incidence rate, post-fracture treatment, and all-cause mortality. *Journal of Bone and Mineral Research*, 38(8):1064–1075, 2023. 38

[169] Kathleen Y Wolin, Kenneth Carson, and Graham A Colditz. Obesity and cancer. *The oncologist*, 15(6):556–565, 2010. 38

[170] Béatrice Lauby-Secretan, Chiara Scoccianti, Dana Loomis, Yann Grosse, Franca Bianchini, and Kurt Straif. Body fatness and cancer—viewpoint of the iarc working group. *New England journal of medicine*, 375(8):794–798, 2016. 38

[171] Katherine M Flegal, Barry I Graubard, David F Williamson, and Mitchell H Gail. Cause-specific excess deaths associated with underweight, overweight, and obesity. *Jama*, 298(17): 2028–2037, 2007. 38

[172] Michael C Wong, Cassidy McCarthy, Nicole Fearnbach, Shengping Yang, John Shepherd, and Steven B Heymsfield. Emergence of the obesity epidemic: 6-decade visualization with humanoid avatars. *The American journal of clinical nutrition*, 115(4):1189–1193, 2022. 38

[173] Matthias Blüher. Obesity: global epidemiology and pathogenesis. *Nature reviews endocrinology*, 15(5):288–298, 2019. 38

[174] Jonathan Bennett, Michael C Wong, Cassidy McCarthy, Nicole Fearnbach, Katie Queen, John Shepherd, and Steven B Heymsfield. Emergence of the adolescent obesity epidemic in the united states: five-decade visualization with humanoid avatars. *International Journal of Obesity*, 46(9):1587–1590, 2022. 38

[175] Ian J Neeland, Jennifer Linge, and Andreas L Birkenfeld. Changes in lean body mass with glucagon-like peptide-1-based therapies and mitigation strategies. *Diabetes, Obesity and Metabolism*, 26:16–27, 2024. 38

[176] John A Shepherd, Bennett K Ng, Markus J Sommer, and Steven B Heymsfield. Body composition by dxa. *Bone*, 104:101–105, 2017. 38

[177] Glen M Blake and Ignac Fogelman. The role of dxa bone density scans in the diagnosis and treatment of osteoporosis. *Postgraduate medical journal*, 83(982):509–517, 2007. 38

[178] Anne B Newman, Catherine L Haggerty, Bret Goodpaster, Tamara Harris, Steve Kritchevsky, Michael Nevitt, Toni P Miles, and Marjolein Visser. Strength and muscle quality in a well-functioning cohort of older adults: the health, aging and body composition study. *Journal of the American Geriatrics Society*, 51(3):323–330, 2003. 38, 40

[179] Michele K Evans, James M Lepkowski, Neil R Powe, Thomas LaVeist, Marie Fanelli Kuczmarski, and Alan B Zonderman. Healthy aging in neighborhoods of diversity across the life span (handls): overcoming barriers to implementing a longitudinal, epidemiologic, urban study of health, race, and socioeconomic status. *Ethnicity & disease*, 20(3):267, 2010. 38, 40

[180] Eric Orwoll, Janet Babich Blank, Elizabeth Barrett-Connor, Jane Cauley, Steven Cummings, Kristine Ensrud, Cora Lewis, Peggy M Cawthon, Robert Marcus, Lynn M Marshall, et al.

Design and baseline characteristics of the osteoporotic fractures in men (mros) study—a large observational study of the determinants of fracture in older men. *Contemporary clinical trials*, 26(5):569–585, 2005. 38, 40

[181] Bennett K Ng, Markus J Sommer, Michael C Wong, Ian Pagano, Yilin Nie, Bo Fan, Samantha Kennedy, Brianna Bourgeois, Nisa Kelly, Yong E Liu, et al. Detailed 3-dimensional body shape features predict body composition, blood metabolites, and functional strength: the shape up! studies. *The American journal of clinical nutrition*, 110(6):1316–1326, 2019. 38, 40

[182] Y Glaser, P Sadowski, T Wolfgruber, L Lui, S Cummings, and JA Shepherd. Hip fracture risk modeling using dxa and deep learning. In *Medical Imaging meets NeurIPS Workshop*, 2020. 41, 47

[183] John H Hubbell and Stephen M Seltzer. Tables of x-ray mass attenuation coefficients and mass energy-absorption coefficients 1 kev to 20 mev for elements z= 1 to 92 and 48 additional substances of dosimetric interest. Technical report, National Inst. of Standards and Technology-PL, Gaithersburg, MD, 1995. 43

[184] DR White, LHJ Peaple, and TJ Crosby. Measured attenuation coefficients at low photon energies (9.88-59.32 kev) for 44 materials and tissues. *Radiation research*, 84(2):239–252, 1980. 43

[185] Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, and Yann LeCun. Decoupled contrastive learning. In *European conference on computer vision*, pages 668–684. Springer, 2022. 43

[186] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 44

[187] Benoit Dufumier, Pietro Gori, Julie Victor, Antoine Grigis, Michele Wessa, Paolo Brambilla, Pauline Favre, Mircea Polosan, Colm Mcdonald, Camille Marie Piguet, et al. Contrastive learning with continuous proxy meta-data for 3d mri classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 58–68. Springer, 2021. 44

[188] Jane A Cauley, Peggy M Cawthon, Katherine E Peters, Steven R Cummings, Kristine E Ensrud, Douglas C Bauer, Brent C Taylor, James M Shikany, Andrew R Hoffman, Nancy E Lane, et al. Risk factors for hip fracture in older men: the osteoporotic fractures in men study (mros). *Journal of Bone and Mineral Research*, 31(10):1810–1819, 2016. 53