Master-Thesis

Improving Cancer Diagnosis and Prognosis with Self-Supervised Representation Learning in Medical Imaging

Oscar Gentilhomme

(9th June 2025 - 22th Decemebr 2025)

Advisor:
Prof. Dr. Valentina Boeva

Supervisors:
Prof. Dr. Christopher P. Bridge
Albert E. Kim, MD

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of Originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of Master's thesis (in block letters):

Improving Cancer Diagnosis and Prognosis with Self-Supervised Representation Learning in Medical Imaging

Authored by (in block letters):
For papers written by groups the names of all authors are required.

Name(s): GENTILHOMME

First name(s): OSCAR

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date

Boston, 12.22.2025

Signature(s)

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.

Acknowledgments

Thanks Chris, for your patience and kindness. Thanks for the Martinos Center hosting us and allowing me to monopolize the MLSC Cluster. And thank you ICE for not taking me

Contents

List of Figures 4

List of Tables 6

Abstract 7

1 Introduction 8

1.1 From Transformers to Self-Supervised Vision Foundation Models . . 9

1.2 Screening Method: Digital Breast Tomosynthesis 11

1.2.1 BI-RADS Scoring 12

1.2.2 State of the Art of AI on DBT and Mammography 13

1.3 DINO for DBT 14

2 Methods 16

2.1 Pre-training 16

2.1.1 Pre-training Dataset 16

2.1.2 Continuous Pre-Training of DINOv2 16

2.1.3 Embedding Aggregation for Density and Risk Tasks 17

2.2 Benchmark 18

2.2.1 Downstream Task Datasets 18

2.2.2 Density classification 19

2.2.3 Cancer detection 20

2.2.4 Risk assessment 21

3 Experiments 24

3.1 DINOSCAR 24

3.1.1 Checkpoint selection in DINOSCAR 24

3.2 Masking Ratio 26

3.3 Crop size 27

3.4 Loss ratio 28

3.5 Augmentation techniques 30

3.6 Multi-slice 33

3.7 Registers 34

4 Conclusion 36

Bibliography 38

List of Figures

1.1 Reconstruction of an image with 75% of its patches masked using Masked Image Modeling [1] 10
1.2 Self-distillation without labels in DINO [2] 11
1.3 Supported input resolutions in DINOv3 [3] 11
1.4 Digital breast tomosynthesis acquistion and image reconstruction, taken from [4] 12
1.5 BI-RADS breast composition categories: A) almost entirely fatty, B) scattered areas of fibroglandular density, C) heterogeneously dense, D) extremely dense [5] 13
1.6 Downstream task performance as a function of pre-training on DBT scans 14
1.7 Evolution of the cosine similarity between the patch highlighted in red and all other patches at different stages of pre-training. Each panel corresponds to a specific training epoch (indicated by the number below the image). As training progresses, the features produced by the model become less localized and the similarity maps become noisier [3]. 15

2.1 Patient flow chart for the datasets used in this study. The *Risk Dataset* and *Density Dataset* refers to the downstream tasks used to evaluate the performance of the pre-training. BCS-DBT: Breast Cancer Screening—Digital Breast Tomosynthesis dataset is used for the cancer detection task. [6]. 17
2.2 Averaging results over five independent runs reduces performance variability caused by the non-deterministic training process. 21
2.3 Learning rate tuning for the risk assessment task. 22
2.4 Comparison of result variability across datasets and subsets. The *Balanced Subset* shows lower performance than the *Unbalanced Subset*, but yields more consistent results across runs. 23

3.1 Cosine similarity of attention representations across DINOSCAR pre-training checkpoints. Red is 1, showing high similarity and blue represents very low similarity 25
3.2 Downstream task performance across DINOSCAR pre-training checkpoints, used to determine the most suitable checkpoint for each task. 26
3.3 Impact of the masking ratio on downstream task performance. . . . 27
3.4 Effect of varying the relative weighting of the DINO and iBOT losses on downstream task performance. 29
3.5 Illustration about the maximum range of the used augmentations. Center column display what is used in our 'DBT-tailored Augmentations' and in the right columns what was used in DINOv2 32

3.6 Overview of the proposed multi-slice contrastive learning strategy, where local crops are sampled from neighboring slices within the same DBT volume. 33
3.7 Learning from adjacent slices enables the model to learn the context of the slice 34

List of Tables

2.1 Dataset summary statistics for the four datasets used in the study. Information is provided as available for the particular dataset [?] . . 18

2.2 Performance comparison between the full breast density dataset and the down-sampled subset using the same backbone architecture. . . . 19

2.3 Difference in results between our setup and the paper setup. All settings are identical except for the maximum number of Optuna trials. 20

2.4 Mean sensitivity on the test set for the paper results and five independent runs using the same hyperparameters. 21

3.1 Performance metrics across training steps, two separated pre-training runs 26

3.2 Crop size changes the amount and quality of information a crop can contain and affect its learning. 28

3.3 Augmentation difference between DINOv2 and DINOSCAR 30

3.4 Augmentation difference between DINOv2 and DINOSCAR 31

3.5 Multi-slice learning improves the pre-training 34

3.6 Performance comparison of DINOv2 backbones with and without registers, including additional pre-training on DBTs. Cancer detection task was not implemented as it required modifying the model to include the extra register tokens 35

Abstract

This project explore the limitations of self-supervising representation in the medical imaging field. The goal was to understand and evaluate the assumption made to pre-train large SSL models such as DINOv2, and improve the performance of downstream tasks such as: 5 years Risk Assessment, Density Prediction and Cancer Detection. An automated benchmark was developed to optimize the evaluation of the pre-trained model performance. This work aims to bring a better understanding of self-supervised representation, to improve generalization of foundation model in the medical field.

Chapter 1

Introduction

Machine learning has become a central component of modern medical imaging, enabling automated analysis and decision support across a wide range of clinical tasks. In particular, deep learning methods have demonstrated strong performance in image classification, detection, and segmentation, often matching or exceeding expert-level accuracy in controlled settings. Despite these successes, most medical imaging systems remain highly task-specific and rely on large quantities of carefully annotated data, which limits their scalability and robustness in real-world clinical environments.

A key bottleneck in medical imaging research is the cost and availability of expert annotations. Medical images are expensive to label, require specialized domain knowledge, and are often subject to inter-observer variability. At the same time, healthcare institutions possess vast amounts of unlabeled imaging data that remain largely untapped by supervised learning approaches. As a result, models trained on narrowly curated datasets often struggle to generalize beyond their original training distribution and fail to fully leverage the diversity of available data.

Foundation models have recently emerged as a promising alternative to task-specific training paradigms. Trained on large-scale, heterogeneous datasets using self-supervised or weakly supervised objectives, these models aim to learn general-purpose representations that can be transferred across tasks, modalities, and domains. In medical AI, such models offer the potential to reduce dependence on expert annotations while improving robustness and data efficiency, particularly in low-label or zero-shot settings.

However, the adoption of foundation models in medical imaging raises several open questions. Due to privacy constraints and data governance regulations, it is often unrealistic to train or fine-tune these models on large, centralized collections of medical data. Consequently, many approaches rely on models pretrained on natural images or limited medical datasets. This raises concerns about domain shift: to what extent do representations learned outside the medical domain transfer to clinical imaging tasks, and how sensitive are these models to changes in modality, acquisition protocol, or patient population?

Recent self-supervised vision models, such as DINOv2 and DINOv3, have shown strong performance across a variety of downstream tasks, suggesting that rich visual representations can emerge without explicit supervision [7–10]. Preliminary studies indicate that these models can be competitive with, or even outperform, supervised and medical-domain-specific pretraining strategies in certain settings. Nevertheless, reported gains are not always consistent across imaging modalities and clinical tasks, highlighting the need for a more systematic understanding of their assumptions, limitations, and failure modes.

Building on these observations, this thesis investigates the use of self-supervised foundation models for medical imaging, with a particular focus on vision transformers pretrained using DINOv2-style objectives. The central goal is to understand how such models can be adapted more effectively to medical domains characterized by strong domain shift and limited annotations. To ground this investigation, the thesis focuses on digital breast tomosynthesis (DBT), an imaging modality widely used in breast cancer screening and assessment of breast density. The remainder of this introduction first provides background on vision transformers and self-supervised pretraining, followed by an overview of DBT and its clinical relevance. Next, related work on AI for mammography and DBT is reviewed, including recent attempts to adapt DINOv2 to this domain. Motivated by the limitations observed in prior research done here at QTIM, this thesis then explores strategies to improve self-supervised pretraining for medical images, emphasizing domain-specific augmentations and the exploitation of three-dimensional image context.

1.1 From Transformers to Self-Supervised Vision Foundation Models

In 2017, Transformers were introduced with the attention and self-attention mechanisms, enabling the modeling of relationships between all pairs of tokens in a sequence [11]. This architecture showed strong scalability properties compared to CNNs and RNNs, particularly as the amount of training data increased.

Three years later, Dosovitskiy et al. proposed applying the Transformer encoder to images in *"An Image is Worth 16×16 Words"* [12], introducing the Vision Transformer (ViT). In this approach, images are divided into fixed-size patches that are treated as tokens. While ViTs demonstrated competitive performance, they required significantly more data than CNNs to outperform them. This observation led the authors to suggest that future research should focus on combining ViTs with self-supervised learning methods.

In parallel, research in Transformer-based NLP introduced Masked Language Modeling (MLM) with BERT in 2018 [13]. By masking up to 15% of the input tokens and predicting them using the remaining context, BERT learned rich bidirectional representations capturing deep semantic relationships.

Following a similar idea, MLM was adapted to the vision domain and led to Masked Image Modeling (MIM) in 2021 with BEiT [14]. In this framework, image patches are first converted into discrete visual tokens using a pre-trained tokenizer, and the model predicts these tokens rather than raw pixels. BEiT demonstrated that masked prediction could be effectively applied to images, but also revealed a strong dependence on the quality of the tokenizer.

Less than six months later, Masked Autoencoders (MAE) were proposed as a more scalable alternative [1]. Unlike BEiT, MAE directly predicts raw pixel values, removing the need for a tokenizer. This design choice is motivated by the high level of spatial redundancy present in natural images, where neighboring patches often contain strongly correlated information. As a result, a large fraction of the image can be masked while still allowing accurate reconstruction. In contrast, natural language exhibits lower redundancy, as masking too many tokens quickly removes essential semantic and syntactic cues, which explains why Masked Language Modeling typically relies on lower masking ratios (around 15%). Leveraging this redundancy, MAE encodes only the visible patches using a lightweight encoder and employs a shallow decoder to reconstruct the full image. This approach enables efficient training and demonstrates successful reconstruction even when masking up to 75% of the image patches.

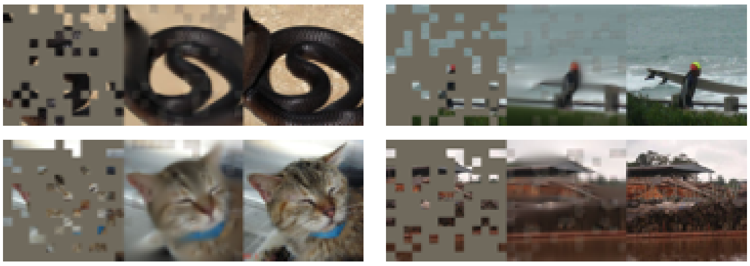

Figure 1.1: Reconstruction of an image with 75% of its patches masked using Masked Image Modeling [1]

Although MAE outperformed BEiT on several benchmarks, the two methods are generally considered to represent distinct branches within the MIM landscape. Token-based MIM approaches, such as iBOT, mc-BEiT, and BEiTv2, address the limitations of BEiT by improving or fixing the tokenizer, thereby reducing the performance gap [15]. These methods benefit from explicitly encoding semantic information, which tends to be advantageous for high-level tasks such as classification and object detection. In contrast, pixel-based MIM approaches, including SimMIM, MaskFeat, and PeCo, adopt a simpler design and focus more on low-level visual details.

Meanwhile, Facebook AI Research explored self-supervised learning with Vision Transformers and introduced DINO [2]. DINO relies on a student–teacher architecture in which both networks receive different augmented views of the same image, as illustrated in Fig. 1.2. The teacher network is updated using an exponential moving average of the student parameters after centering. While teacher–student frameworks had been previously explored in vision [16, 17], DINO was the first to apply self-distillation to Vision Transformers. The method demonstrated that ViTs could compete with convolutional networks in this setting and suggested that self-supervised learning could be a key component for developing BERT-like models for vision.

Building on these ideas, iBOT combined the self-distillation mechanism of DINO with masked patch prediction inspired by BERT and BEiT [15]. By performing predictions at the patch level, iBOT was able to learn both local and global representations, demonstrating that self-distillation and masked modeling are complementary approaches.

In 2023, Facebook AI Research released DINOv2, an improved version of DINO that integrates masked modeling techniques from iBOT [18]. A large-scale data

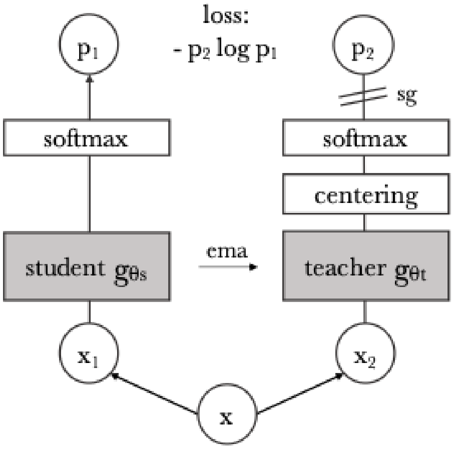

Figure 1.2: Self-distillation without labels in DINO [2]

curation pipeline enabled a substantial increase in the scale of pre-training. DINOv2 achieved performance comparable to supervised models using only weakly trained task-specific heads, highlighting its strong generalization capabilities. At the time of writing, DINOv2 can be considered a foundation model for self-supervised visual representation learning.

More recently, DINOv3 was introduced with significantly larger models and training datasets [3]. The largest DINOv3 model contains 6.7 billion parameters, compared to 1.1 billion for DINOv2, and is trained on approximately 1.1 billion images instead of 142 million. The introduction of Gram Anchoring improves training stability by preventing representation collapse, while support for higher input resolutions enhances cross-domain applicability. Unlike DINOv2, which is limited to 512 × 512 pixel inputs, DINOv3 demonstrates promising results with resolutions up to 4096 × 4096, which is particularly relevant for domains such as medical imaging.

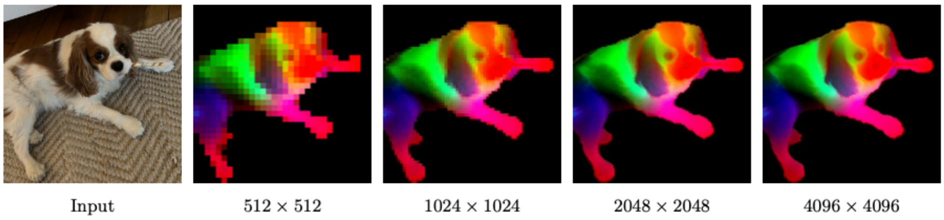

Figure 1.3: Supported input resolutions in DINOv3 [3]

1.2 Screening Method: Digital Breast Tomosynthesis

Breast cancer is often asymptomatic in its early stages, which makes systematic screening essential. Mammography is an X-ray–based imaging technique designed to highlight variations in breast tissue, based on differences in radiation absorption. Current screening recommendations advise individuals with breasts to undergo annual screening starting in their 40s, or earlier, typically in their 30s, in the presence of clinically relevant risk factors such as a family history of breast cancer.

Conventional two-dimensional mammography has been used since the 1960s and represented a significant improvement over earlier chest X-rays performed without breast compression in the 1950s, which were able to detect breast abnormalities only with limited accuracy. By compressing the breast and focusing on dedicated imaging views, 2D mammography enabled more precise and reliable screening.

Digital Breast Tomosynthesis (DBT) is a more recent imaging modality that was approved by the U.S. Food and Drug Administration in 2011. DBT provides a near three-dimensional representation of the breast using both cranio-caudal and mediolateral oblique views. As illustrated in Fig. 1.4, multiple two-dimensional mammographic images are acquired along an arc defined by the X-ray source [19]. These projections are subsequently post-processed to reconstruct layered representations of breast tissue. The scan angle, defined as the angle between the first and last acquisition, varies depending on the imaging system.

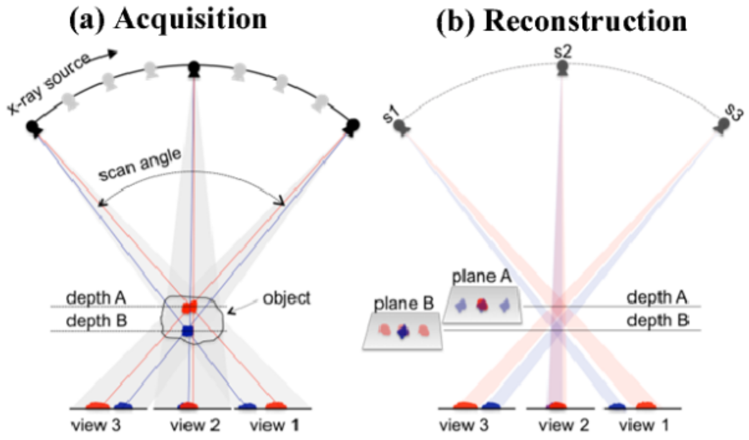

Figure 1.4: Digital breast tomosynthesis acquistion and image reconstruction, taken from [4]

By producing a near-3D representation, DBT improves the detection of smaller tumors and reduces tissue overlap, which in turn lowers the rate of false-positive findings. This leads to fewer unnecessary biopsies and supports more accurate diagnosis [20].

1.2.1 BI-RADS Scoring

The Breast Imaging Reporting and Data System (BI-RADS) is a standardized framework used to describe and classify findings in breast imaging, as well as to estimate the likelihood of breast cancer. Its primary objective is to ensure clear and consistent communication between radiologists and clinicians, while providing well-defined recommendations to guide patient management. BI-RADS defines assessment categories ranging from 0 to 6, each corresponding to a different level of suspicion for malignancy and an associated clinical action.

An essential component of the BI-RADS framework is the characterization of breast composition, which reflects the relative proportions of fatty and fibroglandular tissue. Breast composition is classified into four categories, as illustrated in Fig. 1.5. These categories are directly related to the sensitivity of mammography, as denser breast tissue can obscure abnormalities.

In particular, fibroglandular tissue appears radiopaque on mammograms, similarly to many breast lesions, which can reduce contrast and make small masses or calcifications more difficult to detect. As breast density increases, the likelihood that lesions are masked by surrounding tissue also increases, potentially lowering the diagnostic performance of mammography and increasing the risk of false-negative findings.

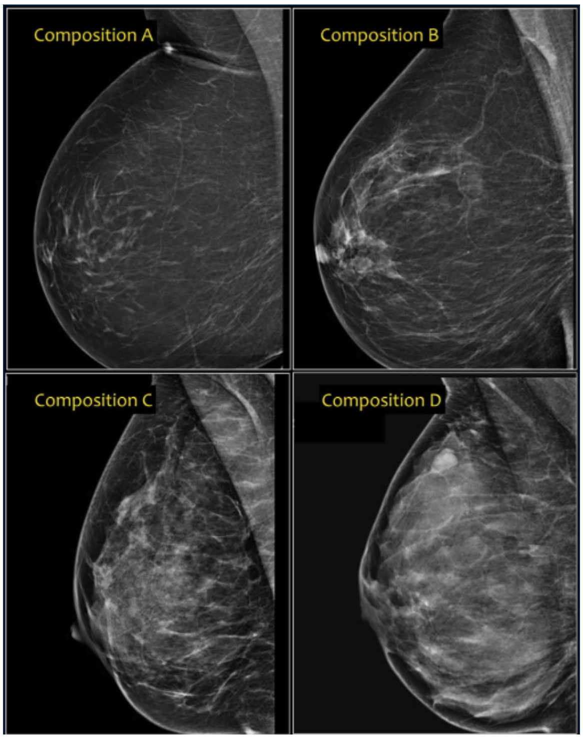

Figure 1.5: BI-RADS breast composition categories: A) almost entirely fatty, B) scattered areas of fibroglandular density, C) heterogeneously dense, D) extremely dense [5]

1.2.2 State of the Art of AI on DBT and Mammography

Self-supervised learning (SSL) methods have gained increasing attention in breast imaging, particularly for two-dimensional mammography. Previous studies have shown that SSL can leverage large amounts of unlabeled data to learn meaningful representations that are effective for downstream tasks such as breast density classification, risk prediction, and cancer detection [21, 22].

More recently, foundation models trained using SSL have demonstrated promising results in medical imaging. For example, RAD-DINO introduced a framework based on DINOv2 that was adapted to radiology images, showing that such models can learn transferable representations across a variety of radiological tasks [23]. Similarly, MedDINOv3 extended the DINOv3 architecture to large-scale medical datasets, improving generalization across imaging modalities by adapting vision foundation models to the specific characteristics of medical images [24]. Together, these works highlight the potential of SSL-based foundation models to serve as strong pretrained backbones for a wide range of medical imaging applications.

Despite these advances, most existing SSL research in breast imaging remains focused on 2D mammography, with comparatively limited exploration of Digital Breast Tomosynthesis (DBT). While DBT differs from mammography by providing a quasi three-dimensional representation of the breast, this increased dimensionality also introduces additional computational complexity. As a result, the application of DINO-style SSL models to DBT remains largely underexplored.

1.3 DINO for DBT

Previous experiments conducted by colleagues at QTIM have shown that the foundation model DINOv2 is a promising encoder for DBT images [25]. However, DINOv2_b14 was originally trained on ImageNet-1k and is therefore not specifically adapted to medical images, including DBTs. To further investigate its performance in this domain, Felix Dorfner developed DBT DINO, a version of DINOv2 pretrained on approximately 487k DBT scans over 475k epochs [6].

Transformer-based models such as DINO rely on self-attention mechanisms to aggregate information across image patches. The way attention is spatially distributed—here referred to as *attention distribution*—reflects how localized or diffuse the learned representations are. Early in training, attention maps tend to be spatially structured and focused on coherent regions, while prolonged pre-training often leads to more uniform or noisy attention patterns, indicating a loss of locality in the learned features.

As observed in the DINOv3 study [3], attention disparity increases over the course of pre-training, as illustrated in Fig. 1.7. In practical terms, this means that feature similarity becomes less dependent on spatial proximity, suggesting that the model progressively shifts from learning localized visual structures to more globally distributed representations. When transferring such models to medical imaging—where fine-grained, localized patterns are often critical—this evolution may be detrimental.

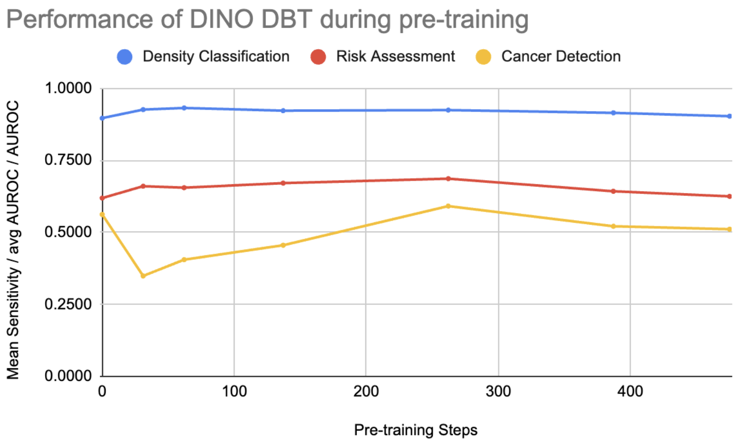

Figure 1.6: Downstream task performance as a function of pre-training on DBT scans

This observation suggests that when continuing the pre-training of DINOv2 on DBT data, the final checkpoint is not necessarily the most suitable for encoding task-relevant information. By evaluating performance on downstream tasks (Fig. 1.6), it was possible to (A) select the most appropriate backbone checkpoint for each task and (B) observe that continued pre-training on DBT data does not consistently improve downstream performance. In the case of lesion detection, performance was even observed to decrease.

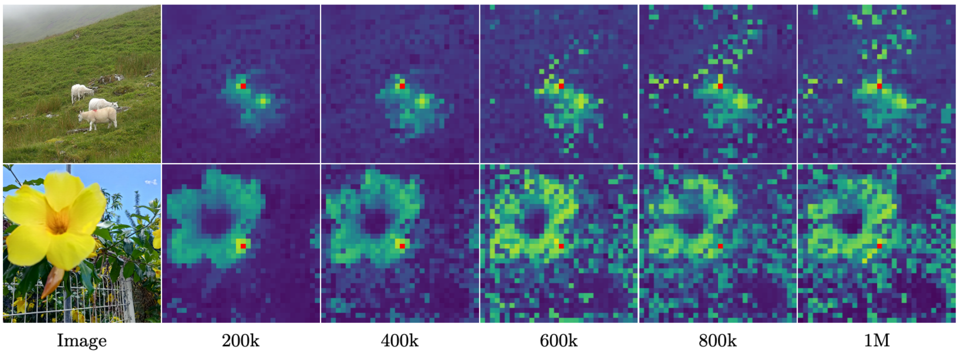

Figure 1.7: Evolution of the cosine similarity between the patch highlighted in red and all other patches at different stages of pre-training. Each panel corresponds to a specific training epoch (indicated by the number below the image). As training progresses, the features produced by the model become less localized and the similarity maps become noisier [3].

These observations motivated the present thesis. They highlight that pre-training strategies and inductive biases inherited from natural image datasets do not necessarily transfer optimally to medical imaging. In this work, I systematically study how domain shift affects DINOv2 pre-training for DBT, with a particular focus on the role of data augmentations. I show that augmentations effective for natural images are not universally beneficial and must be adapted to the medical domain.

Furthermore, I propose and evaluate methods that leverage the intrinsic 3D context of DBT scans—a structural property absent from most natural image datasets, to improve representation learning. Together, these contributions provide practical guidelines for adapting self-supervised foundation models to medical imaging and demonstrate that domain-aware design choices such as the augmentations and crop sizes are essential for effective pre-training.

Chapter 2

Methods

2.1 Pre-training

2.1.1 Pre-training Dataset

This work relies on large-scale Digital Breast Tomosynthesis (DBT) data collected within the Mass General Brigham hospital network in Boston, MA, between 2011 and 2024. The DBT datasets were acquired, filtered, and pre-processed prior to this work. Consequently, we closely follow the data curation procedures described in the *DINO-DBT* and *Towards Early Detection* studies [6, 25].

Across all internal datasets, a total of **130,621 DBT studies** from **34,422 patients** were available and split for training, validation, and testing purposes.

Pre-training dataset The dataset used for continuous pre-training (hereafter referred to as the *pre-training dataset*) consists exclusively of DBT studies from patients who were diagnosed with breast cancer at some point in their medical history. This design choice was made to ensure that the model is exposed to clinically relevant imaging patterns during representation learning.

To prevent data leakage across tasks, a portion of this dataset was held out and explicitly excluded from the datasets used for downstream risk assessment and breast density classification.

Figure 2.1 provides a global overview of all datasets used in this study, including those employed for downstream tasks, while Table 2.1 summarizes their key statistics.

2.1.2 Continuous Pre-Training of DINOv2

To pre-train the model on DBTs, the data is directly streamed from the hospital's picture archive and communication system (PACS) database. The DBT volumes are divided into 2D slices, which are used for training and validation. In addition, several adaptations were applied to the data augmentation pipeline.

As explained in Section 1.1, local and global crops are fed to the student and teacher models. Since DBT slices contain a substantial amount of background, the local

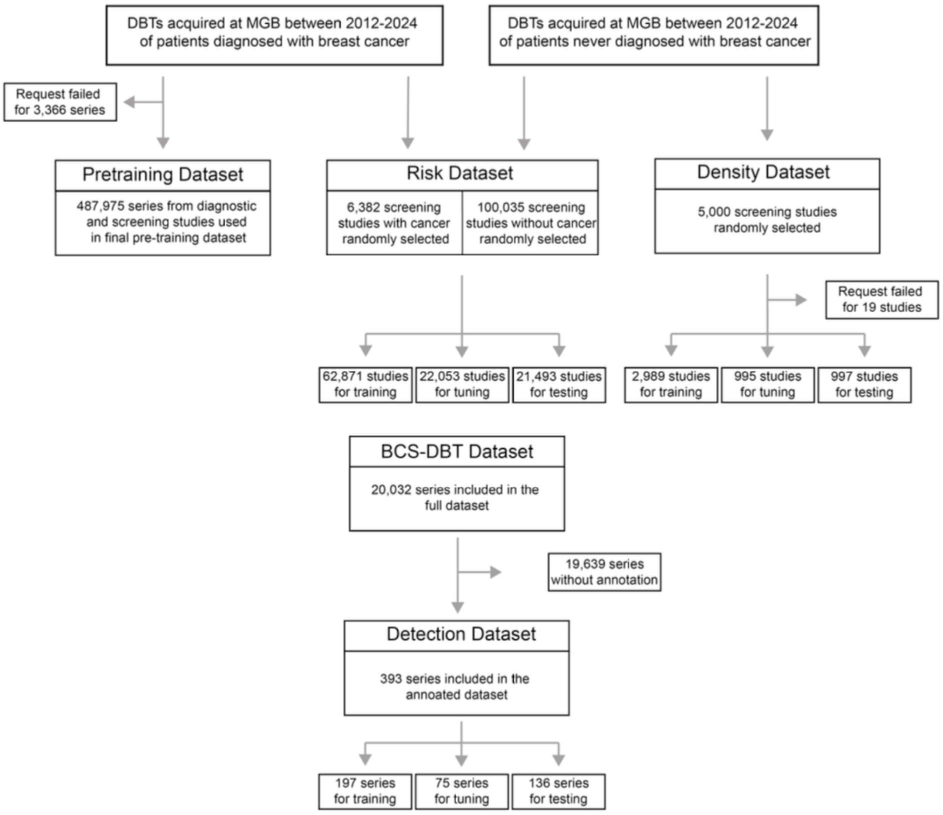
DBTs acquired at MGB between 2012-2024 of patients diagnosed with breast cancer
DBTs acquired at MGB between 2012-2024 of patients never diagnosed with breast cancer
Request failed for 3,366 series
Pretraining Dataset
487,975 series from diagnostic and screening studies used in final pre-training dataset
Risk Dataset
6,382 screening studies with cancer randomly selected
100,035 screening studies without cancer randomly selected
Density Dataset
5,000 screening studies randomly selected
Request failed for 19 studies
62,871 studies for training
22,053 studies for tuning
21,493 studies for testing
2,989 studies for training
995 studies for tuning
997 studies for testing
BCS-DBT Dataset
20,032 series included in the full dataset
19,639 series without annotation
Detection Dataset
393 series included in the annoated dataset
197 series for training
75 series for tuning
136 series for testing

Figure 2.1: Patient flow chart for the datasets used in this study. The *Risk Dataset* and *Density Dataset* refers to the downstream tasks used to evaluate the performance of the pre-training. BCS-DBT: Breast Cancer Screening—Digital Breast Tomosynthesis dataset is used for the cancer detection task. [6].

crop selection was modified to discard crops containing more than 75% background. Further augmentation changes include reducing the maximum radius of GaussianBlur from 2 to 0.5, and removing the GrayScale and Solarize augmentations, as they are not relevant for DBT images.

DINO DBT was trained on four A100-40GB graphics cards (NVIDIA, Santa Clara, USA) for 60,000 optimizer steps. DINOSCAR, our own pre-trained version of DINOv2, was trained on two A100-40GB graphics cards for approximately 30,000 optimizer steps. This setup enabled the parallelization of multiple pre-training configurations and allowed for earlier stopping, as DINO DBT showed its best performance at checkpoints around 17,400 optimizer steps.

2.1.3 Embedding Aggregation for Density and Risk Tasks

Each DBT study consists of four views, referred to as series, with each series containing multiple frames. For each 2D slice processed by DINO (and DINO DBT by extension), the model outputs one CLS token and 1369 patch tokens, each of dimension 768. To obtain predictions at the volume level, these tokens are aggregated across all slices.

To reduce dimensionality while preserving information, summary statistics (mean,

Statistic	**Pre-Training**	**Density**	**Risk**	**Detection**
Volumes	487,975	19,924	425,668	393
Exams	129,883	4,981	106,417	199
Patients	27,990	4,981	31,561	199
Age at study (mean ± SD) [years]	63.54 ± 11.64	57.76 ± 11.40	60.09 ± 10.47	-
Sex				
Female	26,735 (95.52%)	4855 (97.47%)	30,797 (97.57%)	199 (100%)
Male	83 (0.30%)	1 (0.02%)	0	0
Unknown/Other	1,172 (4.19%)	125 (2.51%)	764 (2.42%)	0
Race				
White	22,954 (82.01%)	3772 (75.73%)	25,616 (81.17%)	-
Asian	1,176 (4.20%)	290 (5.82%)	1,245 (3.94%)	-
Black	1,106 (3.95%)	248 (4.98%)	1,934 (6.13%)	-
Other/Unknown	2,754 (9.84%)	671 (13.47%)	2,766 (8.76%)	-

Table 2.1: Dataset summary statistics for the four datasets used in the study. Information is provided as available for the particular dataset [**?**]

standard deviation, minimum, and maximum) are computed either on the CLS tokens across all slices or on the patch tokens over the full volume. Previous experiments have shown that the Risk Assessment task performs best when using all four statistics computed on the patch tokens [25]. In contrast, the Density Prediction task achieves the highest accuracy when using only the mean and standard deviation [6].

The resulting embedding has 768 dimensions per view and per summary statistic. For the Density Prediction task, embeddings are concatenated across both views of each breast, resulting in a 3072-dimensional representation for a full study. For the Risk Assessment task, due to computational limitations, only a single view per study is used, leading to a 768-dimensional embedding. This choice was made to increase dataset variety while limiting its overall size (see Section 2.2.4).

2.2 Benchmark

To evaluate the quality and generalization ability of the learned representations, the model is assessed on three downstream tasks:

- Breast density classification
- Five-year breast cancer risk assessment
- Breast cancer detection

Together, these tasks probe both global and fine-grained semantic understanding of DBT images. Specifically, they evaluate how effectively information encoded in the CLS token and in the patch-level tokens produced by the attention mechanism transfers to clinically relevant objectives.

2.2.1 Downstream Task Datasets

Density dataset The breast density dataset is derived from the hospital archive database (PACS) described in Section 2.1.1 and consists of DBT studies from patients who were never diagnosed with breast cancer. For the original study, 5,000 studies containing all four standard views (bilateral CC and MLO) were randomly selected and balanced across the four breast density categories. The data were split into 3,000 training, 1,000 validation, and 1,000 test studies.

For benchmarking purposes, this dataset was further down-sampled to 1,250 studies (approximately 6,000 volumes) in order to significantly reduce computational cost. As shown in Table 2.2, this reduction resulted in only a limited decrease in performance while enabling substantially more efficient experimentation.

Same experiment	**Density Dataset**	**Density Subset**
# Studies	20,000	6,000
AUROC	0.9257	0.8966
Computation Time [h]	∼24:00	∼5:45

Table 2.2: Performance comparison between the full breast density dataset and the down-sampled subset using the same backbone architecture.

Breast cancer risk dataset The breast cancer risk assessment dataset is also sourced from the hospital PACS archive described in Section 2.1.1 and is composed of two cohorts: a *pre-cancer* cohort and a *healthy* cohort. The pre-cancer cohort includes DBT studies from patients who were diagnosed with breast cancer within five years following the referenced imaging study, whereas the healthy cohort includes only patients who were never diagnosed with breast cancer.

For the pre-cancer cohort, a `years_to_cancer` label was computed to represent the time interval between the DBT study and the subsequent cancer diagnosis. This label is used for classification in the risk prediction task, as described in detail in Section 2.2.4.

Cancer detection dataset The breast cancer detection task relies on the publicly available Duke University BCS-DBT dataset, which provides bounding box annotations for suspicious lesions required for training and evaluating the detection head.

The dataset contains 20,032 DBT series acquired between 2014 and 2018 [26, 27]. Among these, 393 series include bounding box annotations as part of the DBTex challenge. These annotated series are split into 197 training, 75 validation, and 136 test series, following the protocol described in [6].

2.2.2 Density classification

Breast composition assessment is a key component of the BI-RADS protocol, as discussed in Section 1.2.1. As illustrated in Figure 1.5, breast density is categorized into four distinct classes. Breast density is strongly associated with the risk of developing breast cancer and also affects the sensitivity of screening, as small masses may be less visible in dense breast tissue.

The level of semantic understanding required for breast density classification makes representation learning approaches such as DINOv2 particularly well suited for this task. In particular, the limited input resolution primarily affects patch-level tokens, while the CLS token retains a high-level and semantically rich representation of the breast.

For evaluation, a linear probing head is trained using the embeddings produced by the pre-trained backbone, and performance is reported using the AUROC metric on the test set.

2.2.3 Cancer detection

In this work, *detection* is formulated as the identification of bounding boxes that contain cancerous findings in screening digital breast tomosynthesis (DBT) exams. The objective is therefore to localize regions of interest that are likely to contain cancer, rather than performing pixel-level segmentation or volume-level classification. This formulation reflects the primary task of radiologists during screening, which consists of identifying suspicious regions that require further assessment.

Cancer detection is a well-known but challenging task, as it requires a large amount of labeled data. As mentioned in the introduction, the goal is to leverage the large collection of unlabeled DBT volumes available at MGH and use the smaller labeled Duke dataset to train the detection head. Given the limited size of the labeled dataset, we employ the DeiT model (Data-efficient Image Transformers) [28], which is designed to perform well in low-data regimes.

The detection pipeline consists of three main steps: (A) training the detection head on 3D volumes, (B) evaluating the model on 2D slices, and (C) post-processing the 2D bounding box predictions within a 3D context.

Model performance is evaluated using sensitivity at different false positives per volume (FPs/volume). Because radiologists operate at varying false-positive rates, averaging sensitivity across multiple FP rates provides a more realistic evaluation metric than using a single fixed threshold [29]. This evaluation protocol also ensures consistency with the Duke DBT Challenge [26], from which the labeled dataset was obtained.

During training, hyperparameters are tuned using Optuna by maximizing the mean sensitivity. The best trial is selected based on sensitivity at 1 FP per volume, as we observed that the mean sensitivity is strongly influenced by performance at this operating point.

Best optuna trial, same exp	**Paper setup**	**Our setup**
# Trials	225	100
Best trial	146	72
Best trial mean sensitivity	0.9000	0.8693
Mean sensitivity on test set	0.6765	0.5861
Compute time	∼48:00	∼24:00

Table 2.3: Difference in results between our setup and the paper setup. All settings are identical except for the maximum number of Optuna trials.

Initially, Optuna-based hyperparameter optimization was intended to be part of the final pipeline. However, we were unable to reproduce the results reported in the original paper, as shown in Table 2.3. We also trained the detection model using the hyperparameters reported in the paper (Table 2.4). Two main observations emerged: (A) the highest mean sensitivity achieved on the test set was `mean_sensitivity_1_5_fps` = **0.6054**, compared to `mean_sensitivity_1_5_fps` = **0.6765** reported in the paper, suggesting limited reproducibility; and (B) there was substantial variability across runs, likely due to the non-deterministic nature of the training process.

We concluded that running Optuna for 225 trials was comparable to sampling multiple random initializations and selecting the best-performing run. This observation

Paper results	Run 1	Run 2	Run 3	Run 4	Run 5
0.6765	0.5625	0.6054	0.5750	0.5920	0.4760

Table 2.4: Mean sensitivity on the test set for the paper results and five independent runs using the same hyperparameters.

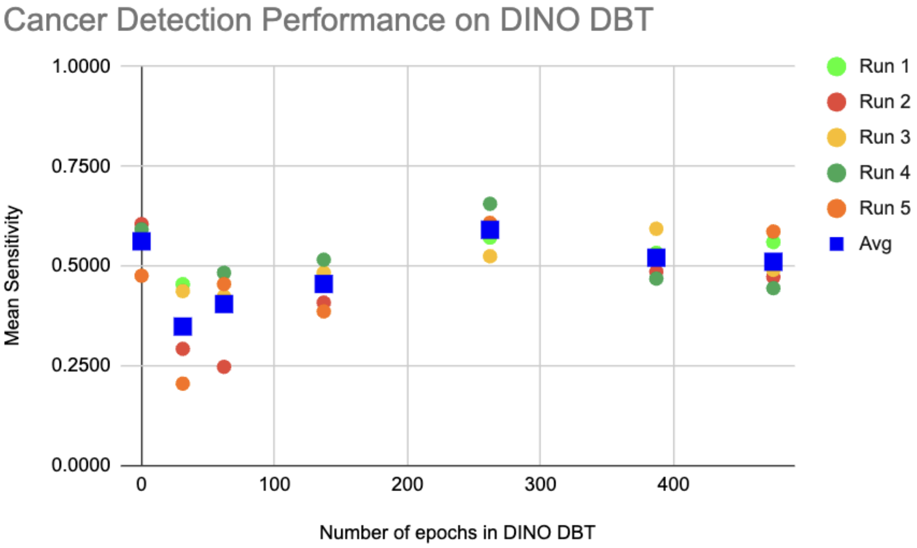

Figure 2.2: Averaging results over five independent runs reduces performance variability caused by the non-deterministic training process.

was corroborated by a colleague who was also unable to reproduce the original results. Furthermore, we were unable to reduce the observed variability, likely due to the small size of the Duke dataset. Consequently, we decided to fix the hyperparameters to those reported in the paper, reduce computational cost, and report results averaged over five independent runs to mitigate variability (Figure 2.2).

2.2.4 Risk assessment

This task is less common in the clinical field compared to breast composition reporting or cancer lesion detection. Implemented at QTIM by Manon Dorster, its objective is to estimate the likelihood of developing breast cancer within five years. This is achieved by comparing pre-cancer studies to healthy scans. We define *pre-cancer* studies as examinations from patients who were diagnosed with breast cancer within five years following the scan. In contrast, *healthy* studies correspond to patients who were not diagnosed with breast cancer within five years after the examination. As this task originates from recent research and is not yet clinically deployed, we consider it particularly relevant for evaluating the impact of pre-training.

The experiments were conducted on an RTX 8000 GPU server with 128 GB of memory and three CPUs. The most time-consuming steps are embedding generation (inference on the backbone) and hyperparameter optimization using Optuna. To limit computational cost, we used a subset of 20k series, resulting in an average runtime of 8h40 for embedding generation and 2h30 for Optuna optimization.

As for the previous tasks, we evaluated the risk assessment task across multiple checkpoints from the DBT DINO and DINOSCAR models to identify which backbone is most suitable. During this initial evaluation, we observed a large variability in performance for repeated runs using the same checkpoint (DINOv2_b14 backbone).

Our first hypothesis was that the linear head was overly sensitive to Optuna initialization. This was ruled out by performing a grid search, which still resulted in substantial variability, as illustrated in Figure 2.3.

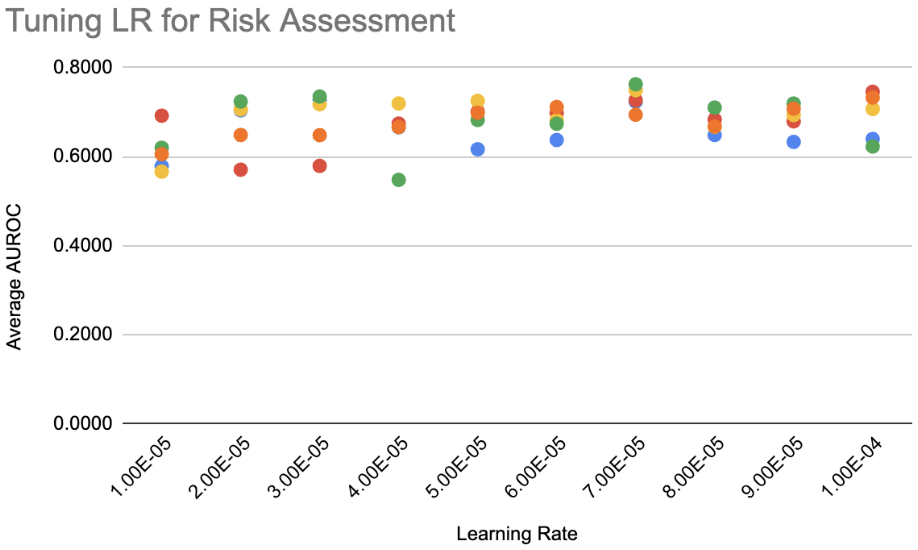

Figure 2.3: Learning rate tuning for the risk assessment task.

Our second hypothesis concerned the subset used for training. Due to computational constraints and the requirement to avoid overlap with the pre-training dataset, the resulting data splits became unbalanced. We ran the task ten times on both the dataset used in the original paper and on the constructed subset. The stability observed on the dataset used in the paper confirmed that the subset (lated called *Unbalanced Subset*) was the source of variability. We therefore created a second subset with a rearranged split (*Balanced Subset*), leading to more stable results, as shown in Figure 2.4. Although the overall performance of this second subset is lower, potentially due to differences in the distribution of `years-to-cancer` stability makes it more appropriate for comparing pre-training experiments.

This task requires a more localized understanding than density classification, as the indicators of elevated cancer risk are expected to lie in subtle temporal or spatial changes within the scans. However, when comparing performance using CLS embeddings versus local patch embeddings, the CLS token consistently performs better. This may be explained by the averaging strategy applied to patch tokens.

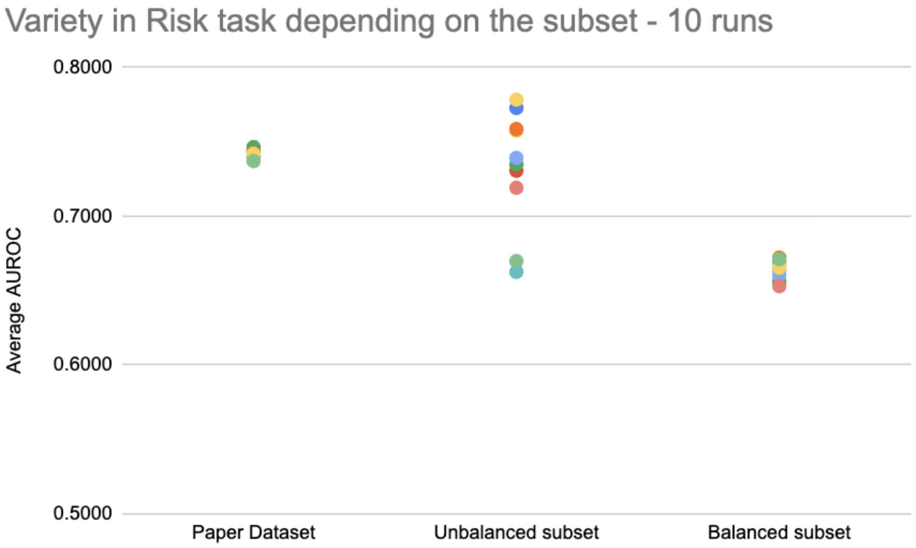

Figure 2.4: Comparison of result variability across datasets and subsets. The *Balanced Subset* shows lower performance than the *Unbalanced Subset*, but yields more consistent results across runs.

Chapter 3

Experiments

As shown in Section 1.3, continuing the pre-training of DINOv2 does not lead to a significant improvement in downstream task performance. (#add plot of task performance along fine-tuning). The first part of this chapter focuses on analyzing downstream task performance, attention distribution, and cosine similarity at various pre-training checkpoints. The objective is to identify potential trends in the most suitable checkpoint for each downstream task, assess the consistency of pre-training across different runs, and observe any abnormal behavior in the attention distribution.

The second set of experiments aims to better understand why self-supervised learning methods, such as DINOv2, do not transfer well to medical imaging data, and more specifically to digital breast tomosynthesis (DBT) scans. We first analyze the two main losses used in DINOv2, namely the iBOT loss and the DINO loss. We then explore the limitations of these techniques, focusing on the masking ratio used in iBOT and the cropping size used in DINO. More generally, we test several assumptions related to the specific properties of our data, such as input dimensionality and adapted augmentations. Finally, we evaluate the importance of input resolution using DINOv3.

3.1 DINOSCAR

3.1.1 Checkpoint selection in DINOSCAR

As mentioned in the introduction, the attention distribution evolves throughout pre-training and progressively decays, as illustrated in Figure 3.1. This behavior, previously reported in the DINOv3 literature, suggests that later checkpoints may exhibit increasingly localized or less diverse attention patterns. Consequently, the final pre-training checkpoint is not necessarily the most suitable choice for all downstream tasks.

To better characterize this phenomenon, we analyze the cosine similarity of attention representations across pre-training checkpoints. Cosine similarity measures the angular similarity between two vectors, independently of their magnitude, and is defined as:

$$\text{cosine\_sim}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|_2, \|\mathbf{b}\|_2}, \tag{3.1}$$

where values close to 1 indicate highly aligned representations, while lower values indicate more diverse or dissimilar representations. Including this formulation is useful for clarity and reproducibility, as it explicitly defines the metric used to quantify representational similarity across checkpoints.

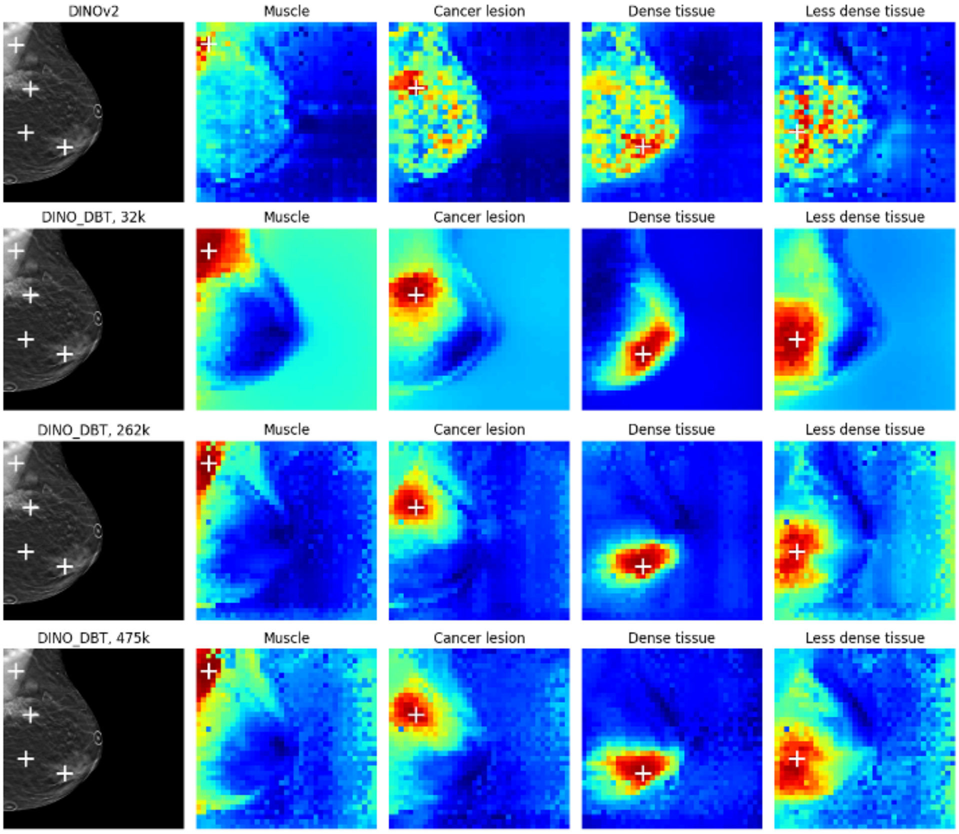

Figure 3.1: Cosine similarity of attention representations across DINOSCAR pre-training checkpoints. Red is 1, showing high similarity and blue represents very low similarity

Figure 3.1 provides additional insight into the dynamics of pre-training. In the early stages of training, the model exhibits higher confusion, reflected by lower and less stable cosine similarity values, and in fact performs worse than the DINOv2_b14. As training progresses, representation quality improves, with the checkpoint at approximately 262k epochs yielding the best overall behavior in terms of cosine similarity and downstream performance. At later stages (e.g., the 475k checkpoint), we observe a clear decay in attention similarity, consistent with the attention collapse trend described in the DINOv3 paper.

Motivated by these observations, we evaluate each downstream task using multiple checkpoints sampled along the DINOSCAR pre-training trajectory. The pre-training process was repeated twice to assess whether the observed trends are consistent across runs.

Based on the results shown in Figure 3.2, the relationship between checkpoint selection and downstream task performance is not straightforward and varies across tasks:

- **Breast density prediction.** Performance remains relatively stable across checkpoints, with limited variation. As a result, checkpoint 50 is selected as a representative reference.

- **Cancer detection.** This task exhibits substantial variability across checkpoints. Nevertheless, performance generally improves with training and peaks at the final checkpoint, which is therefore selected.

- **Risk assessment.** No clearly optimal checkpoint emerges, and the large variability raises concerns regarding result reliability. In cases where the second pre-training run appears to have converged to a local minimum that adversely affects risk assessment performance, checkpoint 100 is selected, as it achieved stronger results in the first run.

Pre-training Steps	**Density Classification** (AUROC)	**Cancer Detection** (mean sensitivity)	**Risk Assessment** (avg AUROC)
0	0.8934 | 0.8972	0.5622 | 0.5882	0.6377 | 0.6408
50	**0.9277 | 0.9285**	0.3580 | 0.2968	0.6344 | 0.6261
100	0.9198 | 0.9176	0.5134 | 0.4009	**0.6326 | 0.6678**
150	0.9100 | 0.9144	0.4338 | 0.4891	0.6251 | 0.6416
200	0.8967 | 0.9151	0.5321 | 0.4111	0.6191 | 0.6531
243	0.9088 | 0.9113	**0.5249 | 0.5175**	0.6220 | 0.6348

Table 3.1: Performance metrics across training steps, two separated pre-training runs

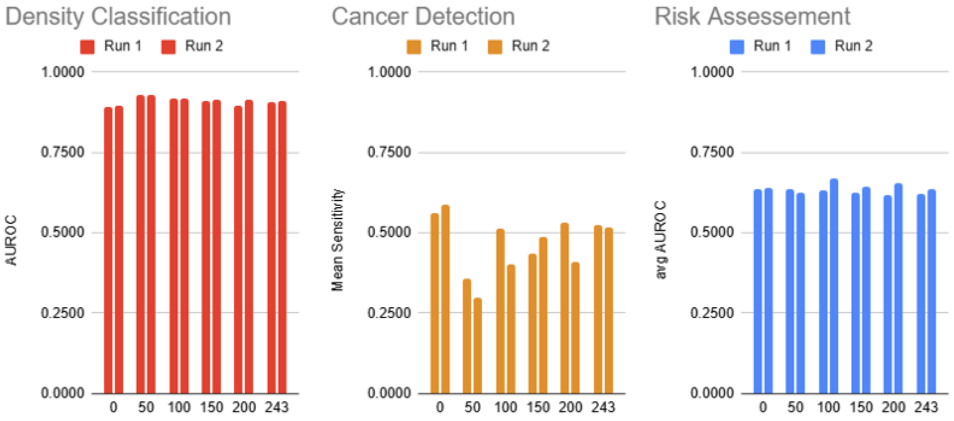

Figure 3.2: Downstream task performance across DINOSCAR pre-training checkpoints, used to determine the most suitable checkpoint for each task.

3.2 Masking Ratio

As recalled in Section 1.1, Masked Image Modeling (MIM) consists in masking a proportion of image patches before feeding them to a Vision Transformer (ViT), and training the model using an auxiliary reconstruction loss. This idea originates from natural language processing, where it was introduced as Masked Language Modeling (MLM) in BERT. BEiT later extended this paradigm to images by encoding image patches into discrete tokens and masking a subset of them during training. iBOT further adapted this approach by operating directly on pixel-level tokens.

MIM has gained attention not only for its strong empirical performance, but also because it reduces the reliance on handcrafted data augmentation strategies, which may be highly domain-dependent.

The optimal masking ratio is modality-dependent. In the MAE paper, the authors note that while a typical masking ratio of 15% is used in NLP, substantially higher ratios are more effective for images. In particular, they report that masking up to 75% of image patches yields good performance, which they attribute to the higher redundancy present in visual data compared to language [1].

One possible explanation for the limited performance gains observed when fine-tuning DINOv2 is that, although the backbone learns a strong global representation of the image, fine-tuning may lead to a degradation of patch-level information. In practice, different self-supervised frameworks adopt different masking strategies: iBOT samples its masking ratio uniformly in the range $(0, 0.3)$, whereas DINOv2 uses a broader range of $(0.1, 0.5)$.

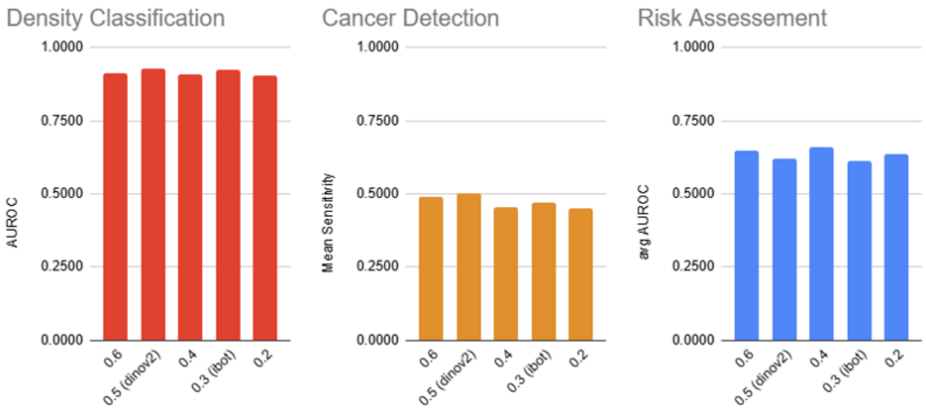

Figure 3.3: Impact of the masking ratio on downstream task performance.

The experimental results do not reveal a single clearly optimal masking ratio across tasks. For both the density estimation and cancer detection tasks, the best performance is obtained with a masking ratio of 0.5, which corresponds to the default value used in DINOv2. A second local maximum can be observed around a masking ratio of 0.3, which aligns with the upper bound of the masking range used in iBOT. For the risk prediction task, the best performance is achieved for masking ratios around 0.4. While no definitive explanation emerges from these results, this behavior may be related to task-specific sensitivity to local versus global information and to differences in effective redundancy within the images.

3.3 Crop size

As explained in the introduction, DINOv2 is trained using a contrastive self-supervised learning framework. The student network receives local crops of the input image, while the teacher network processes global crops. The DINO loss encourages consistency between the representations extracted from local and global views, thereby enforcing both local and global semantic understanding.

The motivation for using smaller local crops is to force the model to focus on fine-grained, localized information rather than relying solely on global context. In our setting, where images are strongly compressed and spatially structured, local crops can contain highly informative regions, sometimes even more discriminative than global crops that include large background areas.

However, our scans contain a substantial amount of background, and reducing the local crop size excessively can be detrimental if the resulting crops do not contain sufficient relevant signal. To mitigate this issue, the DINO-DBT framework introduces an additional constraint during cropping: local crops containing more than 80% background (defined as pixel intensity below 10/255) are discarded.

We evaluate three different configurations that modify the size of global and local crops: 1) reducing the minimum size of local crops; 2) increasing from 50% to 60% the minimum size of global crops and the maximum size of local crops, under the assumption that larger crops may capture more contextual and structural information; 3) reducing the maximum size of local crops, thereby encouraging a stronger focus on fine-grained local patterns.

Setup	Global Crop	Local Crop	Density	Detection	Risk
DINOv2	0.5, 1.0	0.2, 0.5	0.9145	0.4899	0.6489
Setup 1	0.5, 1.0	0.3, 0.5	0.9154	0.4571	0.6402
Setup 2	0.6, 1.0	0.2, 0.6	0.9084	0.4257	0.6280
Setup 3	0.5, 1.0	0.2, 0.4	**0.9233**	**0.4922**	**0.6731**

Table 3.2: Crop size changes the amount and quality of information a crop can contain and affect its learning.

The first setup yields results comparable to the default DINOv2 configuration for the density and risk tasks, but slightly degrades performance on the detection task. The second setup leads to a clear decrease in performance across all three tasks, suggesting that overly large crops may dilute discriminative information by reintroducing excessive background. In contrast, the third setup improves performance consistently across all tasks, indicating that constraining the size of local crops helps the model focus on informative regions while still preserving sufficient contextual information (Table 3.2.

3.4 Loss ratio

The previous experiments on masking ratio and crop size suggest that these hyperparameters may affect the balance between the two learning objectives used in DINOv2. This motivates a closer examination of the relative contribution of the DINO and iBOT losses in our specific setup.

DINOv2 learns visual representations using a discriminative self-supervised approach that combines the DINO and iBOT objectives [18]. Let p_s denote the probability distribution obtained from the student network after softmax, and p_t the probability distribution obtained from the teacher network after centering and softmax. The DINO loss is defined as:

$$\mathcal{L}_{\text{DINO}} = -\sum p_t \log p_s \tag{3.2}$$

The iBOT loss operates at the patch level, where p_{ti} and p_{si} denote the teacher and student probability distributions for patch i:

$$\mathcal{L}_{\text{iBOT}} = -\sum_i p_{ti} \log p_{si} \tag{3.3}$$

In the standard DINOv2 formulation, the two losses are weighted equally:

$$\mathcal{L} = \mathcal{L}_{\text{DINO}} + \mathcal{L}_{\text{iBOT}} \tag{3.4}$$

Given that changes in masking ratio and crop size repeatedly affect the local (iBOT) and global (DINO) objectives, we investigate how sensitive downstream performance is to their relative weighting. To this end, we introduce a weighted loss:

$$\mathcal{L} = \alpha\mathcal{L}_{\text{DINO}} + \beta\mathcal{L}_{\text{iBOT}} \tag{3.5}$$

$$\alpha + \beta = 1 \tag{3.6}$$

We consider two configurations:

$$\texttt{up\_DINO} : \alpha = \frac{2}{3}, \beta = \frac{1}{3}, \qquad \texttt{up\_iBOT} : \alpha = \frac{1}{3}, \beta = \frac{2}{3} \tag{3.7}$$

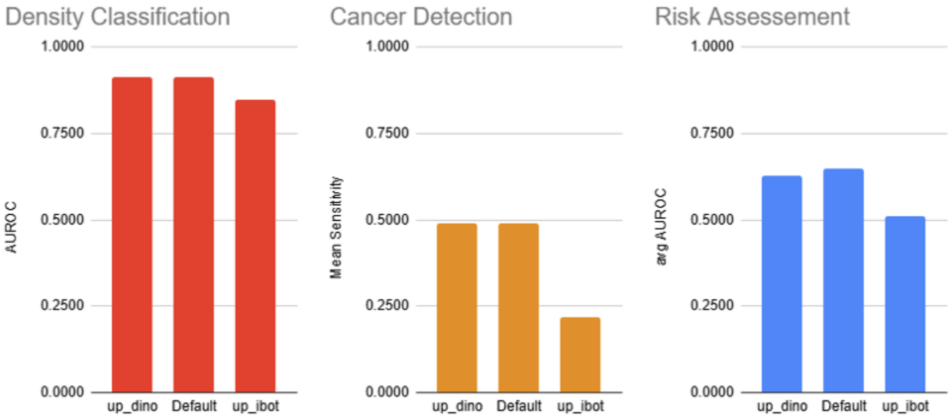

Figure 3.4: Effect of varying the relative weighting of the DINO and iBOT losses on downstream task performance.

The results show that increasing the weight of the DINO loss while decreasing the contribution of the iBOT loss leads to performance comparable to the baseline, albeit slightly lower. In contrast, increasing the weight of the iBOT loss results in a significant degradation of downstream performance. This suggests that, in our data regime, global image-level representations learned through the DINO objective capture most of the information relevant for downstream tasks. The limited benefit of the iBOT objective may be linked to the relatively low redundancy present in the images: while masking-based reconstruction provides some regularization, it appears to contribute less than global semantic alignment to the final performance (Figure 3.4).

3.5 Augmentation techniques

Several image transformations that are essential for effective contrastive learning in natural images are not directly applicable to medical imaging [?]. In particular, DBT images exhibit distinct intensity distributions and structural characteristics that require task-aware augmentation design. For this reason, we revisit the augmentation strategy adopted in DINOv2 and adapt it to better reflect the properties of DBT data, following observations from recent studies highlighting the importance of modality-specific augmentations [30, 31].

Rather than seeking an ideal or universally optimal augmentation scheme, we design a set of DBT-tailored augmentations guided by visual inspection and empirical constraints. For each transformation, suitable parameter ranges were determined through qualitative assessment on DBT images and compared against the parameters used during the original DINOv2 pretraining. These choices are supported by the visual examples presented in Figure 3.5, which illustrate the effect of each transformation on relevant anatomical structures.

Given the large number of augmentation parameters, it was not feasible to evaluate each transformation independently. Moreover, since these experiments are specific to DBT imaging, the objective is not to isolate the individual contribution of each augmentation. Instead, the goal is to demonstrate that adapting augmentations to the target medical domain is necessary, rather than relying on configurations optimized for natural images. The selected DBT-tailored parameters and their comparison with the original DINOv2 configuration are summarized in Table 3.3.

The considered augmentations include contrast, brightness, hue, posterization, saturation, Gaussian blur, sharpening, and solarization. Some transformations, such as hue and saturation, have limited relevance for black-and-white medical images. As shown in Figure 3.5, contrast and brightness adjustments, as well as posterization, preserve the visibility of cancerous structures while modifying global image statistics. Gaussian blur was reduced and sharpening was introduced, as DBT imaging benefits from higher sensitivity to fine spatial details compared to natural images. Solarization was entirely removed, as it suppressed high-intensity pixels corresponding to cancerous regions, effectively cancelling diagnostically relevant signals. Additional visual references are provided in the annex, Figure 3.5.

	DINOv2 Augmentations	**DBT-tailored Augmentations**
Contrast	0.4	0.6
Brightness	0.4	0.6
Hue	0.1	–
Posterize	–	4
Saturation	0.2	–
Gaussian Blur	2	0.5
Random Sharpen	–	2
Solarize	128	–

Table 3.3: Augmentation difference between DINOv2 and DINOSCAR

The impact of these DBT-tailored augmentations is reported in Table 3.4. Performance improves consistently across all three downstream tasks. These results indicate that adapting augmentation strategies to the characteristics of DBT images enables the model to extract more relevant features, resulting in higher-quality representations and improved downstream performance.

	DINOv2 Augmentations	**DBT-tailored Augmentations**
Density Classification (AUROC)	0.8966	**0.9361**
Cancer Detection (Mean Sensitivity)	0.5622	**0.5354**
Risk Assessment (avg AUROC)	0.6190	**0.6620**

Table 3.4: Augmentation difference between DINOv2 and DINOSCAR

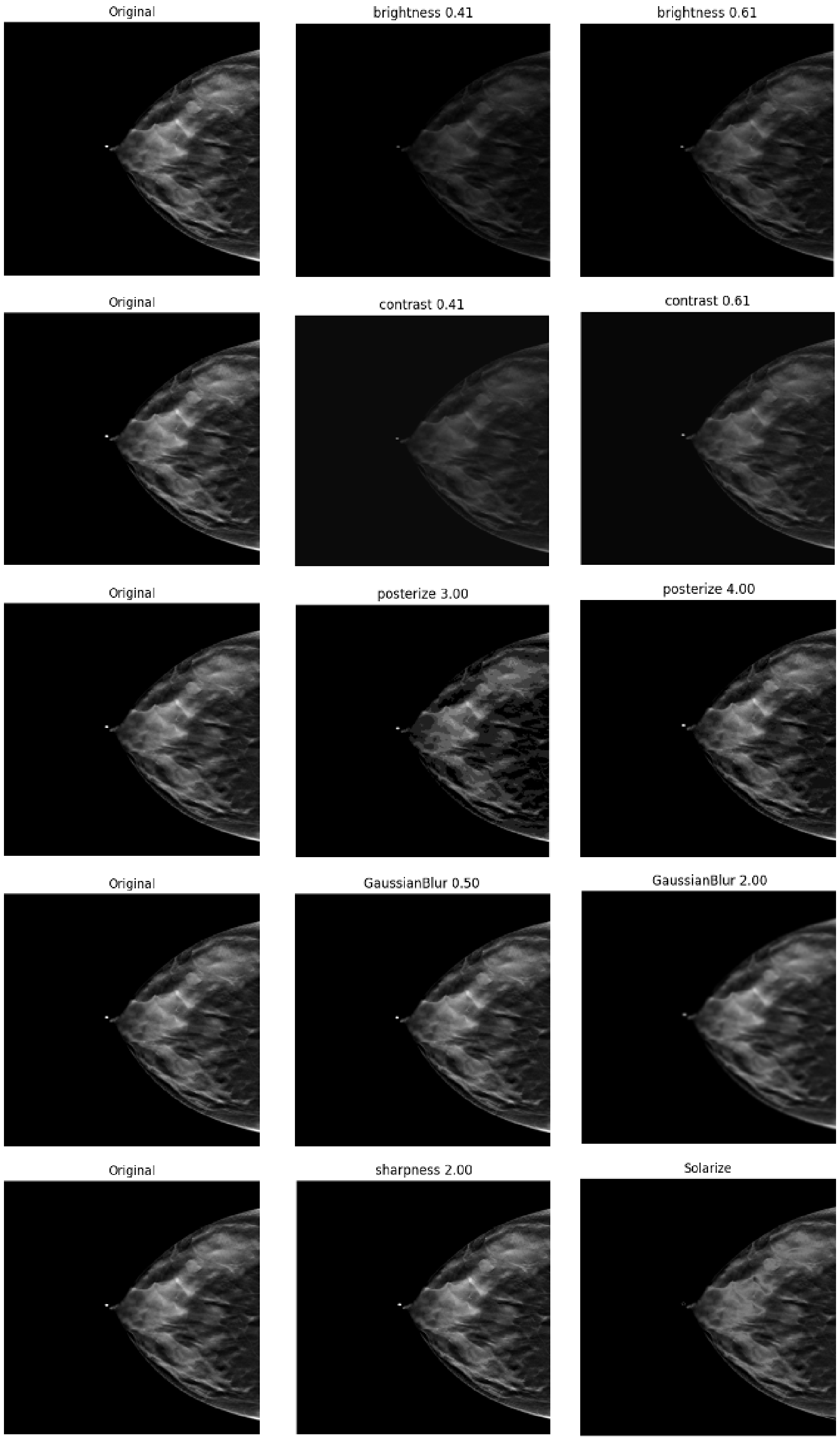

Figure 3.5: Illustration about the maximum range of the used augmentations. Center column display what is used in our 'DBT-tailored Augmentations' and in the right columns what was used in DINOv2

3.6 Multi-slice

One important difference between natural and medical images lies in their dimensionality. Medical imaging modalities such as DBT inherently capture volumetric information, and restricting learning to individual 2D slices may overlook valuable contextual cues. This experiment explores a self-supervised learning strategy designed to learn representations of 2D images while explicitly leveraging their position within a volume.

This approach is inspired by the work of [32], which proposes swapping patches across different satellite sensors to incorporate complementary views during contrastive learning. Motivated by this idea, we hypothesize that using slices from the same volume as alternative views in contrastive learning could help the model better understand each slice in its anatomical context.

As illustrated in Figure 3.6, standard DINO training samples two global crops and eight local crops from the same image, which are fed to the teacher and student networks, respectively. In our multi-slice variant, each input is loaded as a small stack of neighboring slices. The global crops are extracted from the central slice i, while the local crops are sampled from adjacent slices $i-1$ and $i+1$. This design encourages the student network to learn features that are consistent across neighboring slices, effectively capturing how structures evolve across the volume.

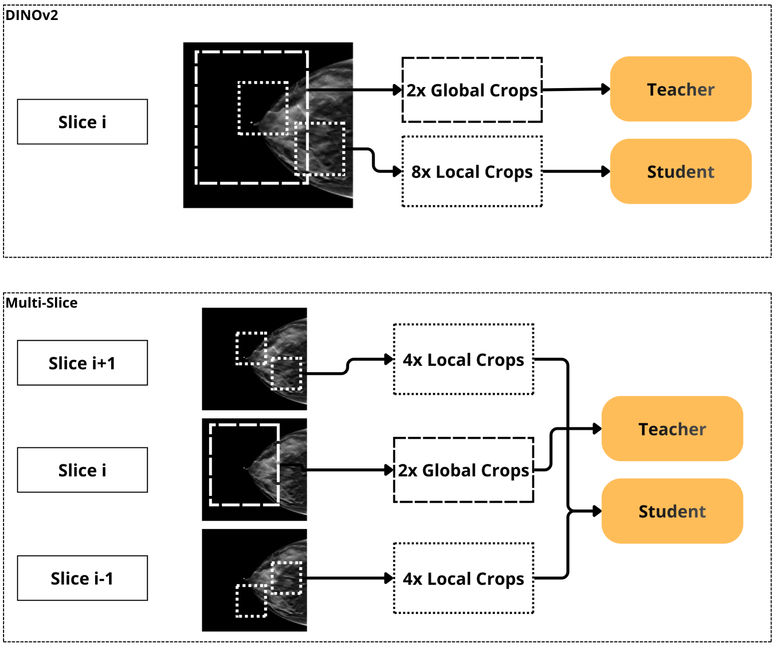

Figure 3.6: Overview of the proposed multi-slice contrastive learning strategy, where local crops are sampled from neighboring slices within the same DBT volume.

The quantitative results are presented in Table 3.5. Performance improves across all three downstream tasks. Notably, the detection task benefits the most, suggesting that this approach is particularly effective for capturing the volumetric extent of localized 2D features in DBT scans.

We further analyze the learned representations by examining the evolution of cosine similarity between embeddings produced by DINOSCAR and DINOSCAR_multi_slice,

	Single slice learning	**Multi-slice learning**
Density Classification (AUROC)	0.8966	**0.9215**
Cancer Detection (Mean Sensitivity)	0.5622	**0.5354**
Risk Assessment (avg AUROC)	0.6190	**0.6502**

Table 3.5: Multi-slice learning improves the pre-training

shown in Figure 3.7. This visualization provides additional insight into how multi-slice training alters the embedding space.

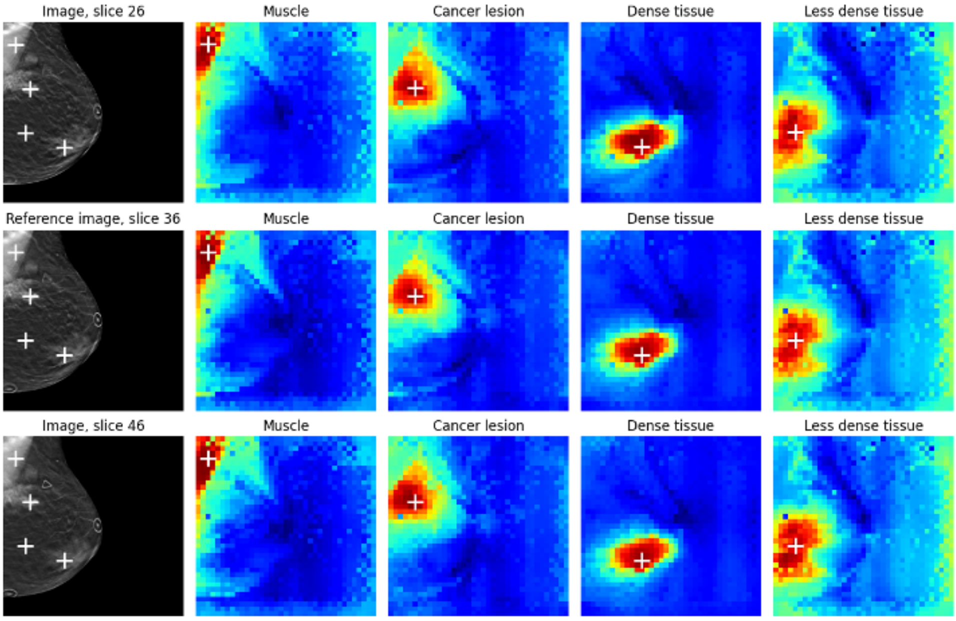

Figure 3.7: Learning from adjacent slices enables the model to learn the context of the slice

3.7 Registers

In the previous experiment, we analyzed the cosine similarity attention distribution at the last layer of both DINODBT and DINOv2 models. We noticed that nearly all attention was concentrated in the top-left four patches. This behavior is consistent with the findings of [33], which report that when the model is presented with more information than can be stored by the available patches, a small subset of patches may arbitrarily accumulate most of the information.

During the course of this thesis, Meta AI released a DINOv2 ViT-B/14 model pre-trained with four register tokens. We continued pretraining using this architecture to evaluate whether registers could improve fine-tuning performance. However, this experiment was conducted using a backbone that differed from the DINOv2 ViT-B/14 model used in all other experiments, making direct comparison inappropriate. To address this, we additionally compared DINOv2 backbones with and without registers to assess the isolated effect of registers on performance.

The results of this comparison are reported in Table 3.6.

When using the DINOv2 backbone pretrained with registers, we observe a slight decrease in performance for the density estimation task and a more pronounced degradation for the detection task, while performance on the risk prediction task improves. In contrast, when the backbone is pretrained on DBTs with registers, the density estimation performance remains comparable to the version without registers. However, registers have a strong negative impact on the detection task. For the risk prediction task, using registers during DBT pretraining does not lead to a clear performance improvement.

A possible interpretation of these results is that register tokens tend to aggregate information in a way that is not well suited for tasks requiring fine-grained, local understanding, such as detection. While registers may help store global or abstract information, this aggregation can be detrimental when precise spatial localization is required. For density estimation, the effect of registers appears limited, and for risk prediction, registers do not seem to provide additional benefits during pretraining on DBTs.

In conclusion, although the attention analysis suggests that information is concentrated in a small number of tokens, introducing additional register tokens does not consistently improve learning performance during pretraining. Moreover, registers appear to degrade performance for tasks that rely on strong local representations, such as detection.

Model	**Density Classification (AUROC)**	**Risk Assessment (avg AUROC)**
DINOv2_vitb14	0.8966	0.6190
DINOSCAR **backbone w/o registers** **& pre-train w/o registers**	0.9277	0.6326
DINOv2_vitb14_reg	0.8890	0.6354
DINOSCAR **backbone w/ registers** **& pre-train w/ registers**	0.9261	0.6365

Table 3.6: Performance comparison of DINOv2 backbones with and without registers, including additional pre-training on DBTs. Cancer detection task was not implemented as it required modifying the model to include the extra register tokens

Chapter 4

Conclusion

This thesis investigated the adaptation of self-supervised foundation models to medical imaging, with a particular focus on vision transformers pretrained using DINOv2-style objectives and their application to digital breast tomosynthesis (DBT). Motivated by prior observations that continued pre-training on DBT data does not necessarily translate into improved downstream performance, this work aimed to better understand how domain shift, training dynamics, and design choices influence representation quality in a low-label medical setting.

A first key finding of this work is that prolonged self-supervised pre-training does not monotonically improve downstream performance in DBT. Consistent with observations reported in DINOv3, extended pre-training leads to a progressive loss of spatial locality in attention patterns, resulting in more diffuse and noisy representations. Empirically, intermediate checkpoints along the pre-training trajectory often outperformed the final model, and in some cases continued training even degraded performance, particularly for lesion detection. These results highlight the importance of checkpoint selection and suggest that representation collapse or over-globalization can be detrimental for medical tasks that rely on fine-grained, localized information.

The analysis of masking strategies further supports the idea that optimal self-supervised configurations are task- and modality-dependent. While no single masking ratio emerged as universally optimal, consistent trends were observed across tasks. Density estimation and cancer detection benefited most from higher masking ratios around 0.5, whereas risk prediction peaked at slightly lower values. These findings suggest a trade-off between preserving local structural details and encouraging global semantic abstraction, with different clinical tasks exhibiting varying sensitivity to this balance.

The comparison between DINO and iBOT objectives further revealed that, in the DBT setting, global image-level semantic alignment plays a more critical role than patch-level reconstruction. Increasing the weight of the iBOT loss led to degraded performance, while prioritizing the DINO loss yielded results comparable to the baseline. This suggests that the limited redundancy present in DBT images reduces the effectiveness of aggressive masking-based reconstruction and that learning robust global representations is more beneficial for the studied tasks.

A central contribution of this thesis is the evaluation of domain-specific augmentations. Augmentation strategies originally designed for natural images were shown to be suboptimal for DBT, and adapting them to reflect medical imaging properties, such as limited texture variability, strong anatomical structure, and acquisition. Specific noise resulted in consistent performance gains across all downstream tasks. These results provide concrete evidence that augmentation design is inherently domain-dependent and help explain why foundation models pretrained on natural images often fail to transfer optimally to medical imaging without careful adaptation.

Beyond two-dimensional representations, this work explored the integration of three-dimensional context through a multi-slice DINO framework. By sampling local crops from neighboring slices while anchoring global crops to a central slice, the model was encouraged to learn slice-consistent semantic representations. This approach effectively teaches the network to assign similar semantic meaning to corresponding structures across adjacent slices, while becoming more invariant to slice-specific noise and quality variations. Leveraging this intrinsic 3D structure, which is absent from most natural image datasets, represents a promising direction for improving self-supervised learning in volumetric medical imaging.

Despite these contributions, several limitations must be acknowledged. The lesion detection task suffers from high variability due to the limited amount of labeled data, which makes performance estimates sensitive to dataset composition and sampling effects. Conversely, the risk prediction task requires very large cohorts, as pre-cancer signals manifest as subtle, diffuse changes that are difficult to capture without extensive longitudinal data. These constraints limit the statistical power of some conclusions and highlight the broader challenge of evaluation in low-prevalence medical tasks.

Looking forward, recent advances such as DINOv3 offer promising avenues to address several of the limitations identified in this work. In particular, the ability to process larger input images, as demonstrated in MED-DINOv3, may help preserve spatial detail while maintaining global coherence, potentially mitigating the attention collapse observed in earlier models. Combined with domain-aware augmentations and volumetric training strategies, such architectures could further improve the transferability of foundation models to medical imaging.

In conclusion, this thesis demonstrates that the effective application of self-supervised foundation models to medical imaging requires more than large-scale pre-training. Careful consideration of domain-specific properties, including augmentations, spatial sampling, masking strategies, and volumetric context—is essential to preserve clinically relevant information. By systematically analyzing and adapting these components, this work provides practical guidelines for improving representation learning in DBT and contributes to a broader understanding of how foundation models can be meaningfully deployed in medical imaging.

Bibliography

[1] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021.

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021.

[3] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025.

[4] Ingrid S. Reiser and Ioannis Sechopoulos. A review of digital breast tomosynthesis. 2014.

[5] S. V. Destounis, S. M. Friedewald, L. J. Grimm, S. P. Poplack, and J. S. Sung. Mammography. In *ACR BI-RADS v2025 Manual*. American College of Radiology, Reston, VA, 2025.

[6] Felix J. Dorfner, Manon A. Dorster, Ryan Connolly, Oscar Gentilhomme, Edward Gibbs, Steven Graham, Seth Wander, Thomas Schultz, Manisha Bahl, Dania Daye, Albert E. Kim, and Christopher P. Bridge. Dbt-dino: Towards foundation model based analysis of digital breast tomosynthesis, 2025.

[7] Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, March 2024. Publisher: Nature Publishing Group.

[8] Matthew Christensen, Milos Vukadinovic, Neal Yuan, and David Ouyang. Vision–language foundation model for echocardiogram interpretation. *Nature Medicine*, 30(5):1481–1488, May 2024. Publisher: Nature Publishing Group.

[9] Suraj Pai, Ibrahim Hadzic, Dennis Bontempi, Keno Bressem, Benjamin H. Kann, Andriy Fedorov, Raymond H. Mak, and Hugo J. W. L. Aerts.

Vision Foundation Models for Computed Tomography, January 2025. arXiv:2501.09001 [eess].

[10] Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Matthew P. Lungren, Maria Teodora Wetscherek, Noel Codella, Stephanie L. Hyland, Javier Alvarez-Valle, and Ozan Oktay. Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence*, 7(1):119–130, January 2025. Publisher: Nature Publishing Group.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[12] 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.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[14] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers, 2022.

[15] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer, 2022.

[16] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020.

[17] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.

[18] 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.

[19] Parvinder Sujlana, Mahadevappa Mahesh, Srinivasan Vedantham, Susan C. Harvey, Lisa A. Mullen, and Ryan W. Woods. Digital breast tomosynthesis: Image acquisition principles and artifacts. *Clinical imaging*, 2019.

[20] Idan Kassis, Dror Lederman, Gal Ben-Arie, Maia Giladi Rosenthal, Ilan Shelef, and Yaniv Zigel. Detection of breast cancer in digital breast tomosynthesis with vision transformers. *Scientific Reports*, 14(1):22149, 2024.

[21] Han Chen and Anne L. Martel. Enhancing breast cancer detection on screening mammogram using self-supervised learning and a hybrid deep model of swin transformer and convolutional neural network, 2025.

[22] Mehmet Ali Karagoz and Okan Umut Nalbantoglu. A self-supervised learning model based on variational autoencoder for limited-sample mammogram classification. *Applied Intelligence*, 54:3448–3463, 2024.

[23] Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Matthew P. Lungren, Maria Teodora Wetscherek, Noel Codella, Stephanie L. Hyland, Javier Alvarez-Valle, and Ozan Oktay. Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence*, 7(1):119–130, January 2025.

[24] Yuheng Li, Yizhou Wu, Yuxiang Lai, Mingzhe Hu, and Xiaofeng Yang. Meddinov3: How to adapt vision foundation models for medical image segmentation?, 2025.

[25] Manon Dorster, Felix J. Dorfner, Mason Cleveland, Melisa Guelen, Jay Patel, Dania Daye, Jean-Philippe Thiran, Albert Kim, and Christopher Bridge. *Towards Early Detection: AI-Based Five-Year Forecasting of Breast Cancer Risk Using Digital Breast Tomosynthesis Imaging*, pages 63–71. ResearchGate, 09 2025.

[26] Mateusz Buda, Ashirbani Saha, Ruth Walsh, Sujata Ghate, Nianyi Li, Albert Swiecicki, Joseph Y. Lo, and Maciej A. Mazurowski. A data set and deep learning algorithm for the detection of masses and architectural distortions in digital breast tomosynthesis images. *JAMA Network Open*, 4(8):e2119100, August 2021.

[27] M. Buda, A. Saha, R. Walsh, S. Ghate, N. Li, A. Swiecicki, J. Y. Lo, J. Yang, and M. Mazurowski. Breast Cancer Screening – Digital Breast Tomosynthesis (BCS-DBT), 2020. Version Number: 5.

[28] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers and distillation through attention, 2021.

[29] Nicholas Konz, Mateusz Buda, Hanxue Gu, Ashirbani Saha, Jichen Yang, Jakub Chłedowski, Jungkyu Park, Jan Witowski, Krzysztof J. Geras, Yoel Shoshan, Flora Gilboa-Solomon, Daniel Khapun, Vadim Ratner, Ella Barkan, Michal Ozery-Flato, Robert Martí, Akinyinka Omigbodun, Chrysostomos Marasinou, Noor Nakhaei, William Hsu, Pranjal Sahu, Md Belayat Hossain, Juhun Lee, Carlos Santos, Artur Przelaskowski, Jayashree Kalpathy-Cramer, Benjamin Bearce, Kenny Cha, Keyvan Farahani, Nicholas Petrick, Lubomir Hadjiiski, Karen Drukker, III Armato, Samuel G., and Maciej A. Mazurowski. A competition, benchmark, code, and data for using artificial intelligence to detect lesions in digital breast tomosynthesis. *JAMA Network Open*, 6(2):e230524–e230524, 02 2023.

[30] Fabio Garcea, Alessio Serra, Fabrizio Lamberti, and Lia Morra. Data augmentation for medical imaging: A systematic literature review. *Computers in biology and medicine*, 152:106391, 2023.

[31] Xinyu Liu, Gizem Karagoz, and Nirvana Meratnia. Analyzing the impact of data augmentation on the explainability of deep learning-based medical image classification. *Machine Learning and Knowledge Extraction*, 7(1):1, 2024.

[32] Gencer Sumbul, Chang Xu, Emanuele Dalsasso, and Devis Tuia. Smarties: Spectrum-aware multi-sensor auto-encoder for remote sensing images. *arXiv preprint arXiv:2506.19585*, 2025.

[33] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers, 2024.

[34] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249, 2021.

[35] Robbie Holland, Oliver Leingang, Hrvoje Bogunović, Sophie Riedl, Lars Fritsche, Toby Prevost, Hendrik PN Scholl, Ursula Schmidt-Erfurth, Sobha Sivaprasad, Andrew J Lotery, et al. Metadata-enhanced contrastive learning from retinal optical coherence tomography images. *Medical Image Analysis*, 97:103296, 2024.