Methods and frameworks of annotation cost optimization for deep learning algorithms applied to medical imaging

Camille Ruppli

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2023IPPAT039

Thèse de doctorat

Methods and frameworks of annotation cost optimization for deep learning algorithms applied to medical imaging

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Telecom Paris

École doctorale n°626 École doctorale de l'Institut Polytechnique de Paris (ED IP Paris)
Spécialité de doctorat : Informatique

Thèse présentée et soutenue à Paris, le 13 décembre 2023, par

CAMILLE RUPPLI

Composition du Jury :

Diana Mateus Professeure, Ecole Centrale Nantes	Présidente / Examinatrice
Ender Konukoglu Professeur, ETH Zurich	Rapporteur
Jean-Philippe Thiran Professeur, EPFL (Signal Processing Laboratory)	Rapporteur
Ninon Burgos Chargée de recherche, Paris Brain Institute (ARAMIS Lab)	Examinatrice
Isabelle Bloch Professeure, Sorbonne Université CNRS (LIP6), Télécom Paris (LTCI)	Directrice de thèse
Roberto Ardon Docteur, Incepto Medical	Co-directeur de thèse

Acknowledgements

I would like to sincerely thank the members of the jury: Ender Konukoglu and Jean-Philippe Thiran for reviewing my PhD thesis, Diana Mateus and Ninon Burgos for examining my work.

To my two academic supervisors, Isabelle and Pietro, thanks for having been such an inspiration. We met six years ago when I was an engineering student at Telecom and your teaching and guidance played a major part in my decision to engage in the PhD adventure. To Roberto, my Incepto supervisor, considering the available time you have, I have been beyond lucky for the writing, coding and so many brainstorming sessions you devoted to me, although it sometimes challenged my patience!

The advice, trust, and support of the three of you were incredible allies along the way and many of our weekly meetings were morale boosters when experiments, coding, or publications were not going exactly as expected.

I remember Isabelle telling me that the PhD path was not an easy one and that a healthy work environment was much needed to get through these years. I would like to thank the whole Incepto team, and especially the data science one, for building such a nice, benevolent, and ambitious workplace where I hope to keep evolving for a while! To Florence Moreau, thank you for your trust from the beginning when we met at Station F for my end-of-study internship. A special thanks to Benoit Bayol for his patience and guidance in navigating the cluster environment full of error code 137. To Martin, a huge thanks for paving the PhD way, from administrative issues to experimental and manuscript advice.

To the members of the LTCI lab, although I have not been the most present you have always welcomed me and I enjoyed our few Palaiseau running sessions and shared lunches.

This thesis marks the end of my higher education years and I would like to thank my family and friends for their support along the way and for managing, somehow, to understand what I was working on for three years. To my parents, without whom so much of this journey would not have been possible, thank you for passing on to me the value of hard work, and the love of running and music which have been such strong pillars during these years. To my sister, Sarah, my best friend, the most special thanks for having been my biggest and strongest ally for more than 20 years. It will soon be my turn to support you and I hope that I will be able to fulfill this role with the same strength and consistency as you have. Finally, to Robin, my life partner, I would like to express the deepest gratitude as I would have probably not managed these years without you and our fury four-legged friend.

A chapter is over, a new adventure begins!

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

ADC	Apparent Diffusion Coefficient
AI	Artificial Intelligence
AUC	Area Under Curve
DL	Deep Learning
DWI	Diffusion Weighted Imaging
mAP	Mean Average Precision
MRI	Magnetic Resonance Image
mp-MRI	Multi-parametric Magnetic Resonance Images
PI-RADS	Prostate Imaging Reporting and Data System
T2w	T2-weighted

Résumé

Ces dernières années, la quantité de données d'imagerie médicale n'a cessé de croître. En 1980, 30 minutes d'acquisition étaient nécessaires pour obtenir 40 images médicales. Aujourd'hui, 1000 images peuvent être acquises en 4 secondes.

Cette croissance de la quantité de données est allée de pair avec le développement de techniques d'apprentissage profond qui ont besoin d'annotations de qualité pour être entraînées. En imagerie médicale, les annotations sont beaucoup plus coûteuses à obtenir car elles nécessitent l'expertise d'un radiologue dont le temps est limité.

L'objectif de cette thèse est de proposer et de développer des méthodes permettant de limiter la charge d'annotation en imagerie médicale tout en maintenant une performance élevée des algorithmes d'apprentissage profond.

Dans la première partie de cette thèse, nous étudions les méthodes d'apprentissage auto-supervisé. Ces méthodes introduisent des sous-tâches de différents types : approches génératives, contextuelle et basée sur l'auto-distillation. Ces tâches sont utilisées pour pré-entraîner un réseau de neurones sans annotations supplémentaires afin de tirer profit des données non annotées disponibles.

La plupart de ces tâches utilisent des perturbations assez génériques, sans rapport avec la tâche supervisée sous-jacente et échantillonnées au hasard dans une liste avec des paramètres fixés. La meilleure façon de combiner et de choisir ces perturbations et leurs paramètres n'est pas encore claire. En outre, certaines perturbations peuvent être préjudiciables à la tâche supervisée objectif. Certains travaux atténuent ce problème en concevant des sous-tâches pour une tâche supervisée spécifique, en particulier dans le domaine de l'imagerie médicale. Mais ces tâches ne se généralisent pas bien à d'autres problèmes.

Un équilibre doit donc être trouvé entre l'optimisation de la perturbation ou de la sous-tâche pour un problème supervisé donné et la capacité de généralisation de la méthode.

Parmi les méthodes basées sur le contexte, les approches d'apprentissage contrastif proposent une tâche de discrimination par instance : l'espace latent est structuré suivant la similarité entre différentes instances. La définition de la similarité des instances est le principal défi de ces approches et a été largement explorée. Lorsque des perturbations sont utilisées pour définir la similarité entre les images, les mêmes questions d'optimisation des perturbations se posent.

Dans la première partie de cette thèse nous proposons d'optimiser les perturbations utilisées dans l'apprentissage contrastif. Nous introduisons un réseau de perturbation qui minimise l'information mutuelle entre les versions perturbées d'une même image sans réduire les performances des tâches supervisées. Notre réseau de perturbation produit un ensemble de paramètres définissant la composition des perturbations à appliquer. Pour construire ce réseau, chaque perturbation doit être différentiable par rapport à ses paramètres. Nous proposons donc une formulation différentiable des perturbations utilisées. Le générateur de perturbations est alors entraîné à minimiser l'information mutuelle entre les images perturbées, tandis que l'encodeur apprend à la maximiser. Une contrainte de supervision est utilisée pour s'assurer que les perturbations générées ne nuisent pas à la tâche cible supervisée. Après le pré-entraînement nous obtenons de bons résultats d'évaluation linéaire sur deux bases de données (images de cerveau et de thorax) et des perturbations pertinentes.

Les annotations de classes et certaines métadonnées ont été utilisées pour conditionner la similarité des instances, mais ces données peuvent être sujettes à la variabilité des annotateurs, en particulier dans le domaine médical. Certaines méthodes ont été proposées pour utiliser la confiance dans l'apprentissage supervisé et auto-supervisé, mais elles sont principalement basées sur les valeurs de la fonction de perte. Cependant, la confiance dans les annotations et les métadonnées est souvent liée à des connaissances a priori du domaine, telles que l'acquisition des données, l'expérience et l'accord entre les annotateurs. Ceci est encore plus pertinent pour les données médicales.

Dans la deuxième partie de cette thèse, nous proposons une fonction de perte contrastive prenant en compte la confiance des annotations pour le problème spécifique de la détection des lésions du cancer de la prostate. Nous introduisons un noyau basé sur la confiance dans la classification des lésions prostatique. Nous mesurons la confiance comme l'accord entre les annotateurs pour chaque examen. Nous définissons ensuite différents noyaux avec des niveaux de complexité croissants en ce qui concerne l'inclusion de la confiance. Après le pré-entraînement, nous avons spécialisé le réseau sur la tâche de détection des lésions avec 1% et 10% de données annotées. Avec 1% de données annotées, nous montrons des améliorations substantielles avec le noyau le plus complexe utilisant la confiance, par rapport aux autres approches d'apprentissage contrastif de la littérature.

Enfin, nous explorons quelques approches pour appliquer l'apprentissage auto-supervisé et contrastif à la segmentation des lésions du cancer de la prostate.

Abstract

In recent years, the amount of medical imaging data has kept on growing. In 1980, 30 minutes of acquisition were necessary to obtain 40 medical images. Today, 1000 images can be acquired in 4 seconds.

This growth in the amount of data has gone hand in hand with the development of deep learning techniques which need quality labels to be trained. In medical imaging, labels are much more expensive to obtain as they require the expertise of a radiologist whose time is limited.

The goal of this thesis is to propose and develop methods to limit the annotation load in medical imaging while maintaining a high performance of deep learning algorithms.

In the first part of this thesis, we focus on self-supervised learning methods which introduce pretext tasks of various types: generation-based, context-based, and self-distillation approaches. These tasks are used to pre-train a neural network with no additional annotations to take advantage of the available unannotated data.

Most of these tasks use perturbations often quite generic, unrelated to the objective task and sampled at random in a fixed list with fixed parameters. How to best combine and choose these perturbations and their parameters remains unclear. Furthermore, some perturbations can be detrimental to the target supervised task. Some works mitigate this issue by designing pretext tasks for a specific supervised task, especially in medical imaging. However, these tasks do not generalize well to other problems. A balance must be found between perturbation or pretext task optimization for a given supervised problem and the method's generalization ability.

Among context-based methods, contrastive learning approaches propose an instance-level discrimination task: the latent space is structured with instance similarity. Defining instance similarity is the main challenge of these approaches and has been widely explored. When defining similarity through perturbed versions of the same image, the same questions of perturbation optimization arise.

In the first part of this thesis, we propose an optimization of perturbations used in contrastive learning. We introduce a perturbation network minimizing the mutual information between views without reducing supervised task performances. Our perturbation network outputs a set of parameters defining the composition of perturbations to apply. To build this network, each perturbation must be differentiable with respect to its parameters. We thus propose a differentiable formulation of the used perturbations. The perturbation generator is then trained to minimize the mutual information between views while the encoder learns to maximize it. A supervision constraint is used

to ensure that the generated perturbations are not detrimental to the supervised target task. After pre-training, we obtained good linear evaluation results on two datasets (brain and chest images) and relevant perturbation outputs.

Class labels and metadata have been used to condition instance similarity, but these data can be subject to annotator variability, especially in the medical domain. Some methods have been proposed to use confidence in fully supervised and self-supervised training, but it is mostly based on loss function values. However, confidence on labels and metadata is often linked to a priori domain knowledge such as data acquisition, annotators' experience, and agreement. This is even more relevant for medical data.

In the second part of this thesis, we design an adapted contrastive loss introducing annotation confidence for the specific problem of prostate cancer lesion detection. We introduce a kernel based on classification (PI-RADS or biopsy) confidence. We measure confidence as the agreement among annotators for each exam. We then define different kernels with increasing levels of complexity with respect to confidence inclusion. After pre-training, we fine-tuned the whole encoder-decoder architecture on the lesion detection task with 1% and 10% annotated data. With 1% annotated data, we show substantial improvements with the most complex kernel using confidence compared to other contrastive learning approaches from the literature.

Finally, we explore some approaches to apply self-supervised and contrastive learning to prostate cancer lesion segmentation.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Contents

1.1 Context

In recent years, the amount of medical imaging data has kept on growing. In 1980, 30 minutes of acquisition were necessary to obtain 40 medical images. Today, 1000 images can be acquired in 4 seconds.

This growth in the amount of data has gone hand in hand with the development of deep learning techniques. These techniques allow one to automate, within a neural network, the extraction of particular characteristics, properties of textures, regularities, or irregularities of the image. Neural networks learn these features by analyzing this large volume of images to automatically classify patients in diagnostic categories, to automatically measure pathological or non-pathological structures, or to give elements to indicate the next steps of patient care.

Neural networks have proved effective with large databases of annotated photographic images, including ImageNet [Deng et al., 2009]. Annotated databases allow one, during training, to associate an image with a class (car, cat, dog... for ImageNet for example) and thus to guide the network in its training for the recognition of this class.

In the case of medical data, we are confronted with a dimensionality problem. Today, we have access to databases containing tens of thousands of images. However, the image labels are much more expensive to obtain, requiring the expertise of a radiologist whose time is limited. If the annotation can take only a few minutes when giving a simple label in simple cases (pathological or not, presence of a tumor or not, etc.), it can take several hours in complex cases or when it is necessary to segment precisely 3D normal or pathological structures. We therefore have a large database of three-dimensional images and a small database of annotations with a heterogeneous level of difficulty.

In medical imaging, the deep learning techniques and the types of annotations vary [Chartrand et al., 2017, Litjens et al., 2017, Yu et al., 2018]. Annotations can be separated into three categories depending on the learning task at hand: classification, localization, or segmentation. Obtaining ground truth for this type of annotation usually involves consensus building. Depending on the task, the annotation cost varies. Empirically, we observe that the cost of annotation is correlated with the difficulty of the task. Internal reports show that cartilage anomaly detection and localization take 90 seconds on average while bone segmentation takes 5 minutes.

For a classification task, the annotation consists of assigning a label to an image: sick or healthy, cracked or not cracked. This label can also be multi-class: one among several types of blood cells or chest pathologies for example. This type of annotation, for classification, is generally simpler to obtain and can sometimes be obtained by non-expert individuals trained in the pathology being studied.

For localization tasks, the annotation consists of placing a box on a structure of interest: around a meniscus or an intestinal obstruction for example. Depending on the pathologies and structures, annotations are more expensive and cannot be performed by non-experts.

Finally, the annotation for a segmentation task is the most expensive of the three because it consists of assigning a class to each pixel of the input image.

To address the dimensionality problem posed by medical imaging, various methods have been proposed: active learning, transfer learning, and self-supervised learning. Active learning [Ren et al., 2020] methods can be used to evaluate the prediction uncertainty of an unannotated image using, for instance, dropout methods and submit it to an oracle for annotation (see Figure 1.1).

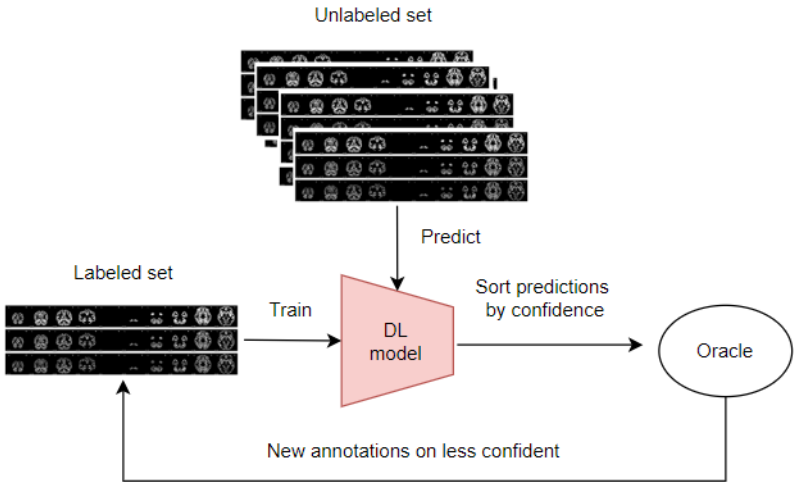


Figure 1.1: Active learning method.

In medical imaging, these methods are widely applied to segmentation tasks. In addition to uncertainty, Yang et al. [2017] have evaluated the image representativeness and Sadafi et al. [2019] the presence of a class rarity in the image. These additional evaluations showed a performance improvement over random annotation.

Transfer learning methods propose to pre-train a model with an available large annotated database (such as ImageNet [Deng et al., 2009]). The learned weights are then used as initialization for fine-tuning on another smaller database (see Figure 1.2).

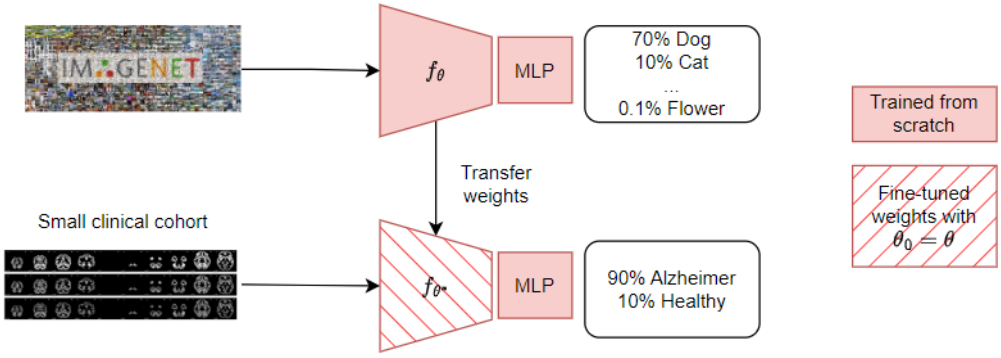


Figure 1.2: Transfer learning method.

This method has been widely used in medical imaging to overcome data scarcity [Ding et al., 2019, Esteva et al., 2017, Wang et al., 2017]. Many works have investigated the conditions of transfer learning efficacy (i.e feature reuse) from a pre-trained model on ImageNet to a medical imaging database. Mustafa et al. [2021] show that transfer learning performances are increased with large architecture and large pre-training dataset (compared to the size of the supervised target task). Matsoukas et al. [2022] add that feature reuse is fostered when models have small inductive biases. Both these works emphasize that benefits from transfer learning increase when there is a small domain distance between source and target datasets. However, as shown by Raghu et al. [2019] small models trained from scratch on medical images can perform better than transfer learning from ImageNet. Transferring from ImageNet might not be relevant for medical imaging especially because the domain gap [Matsoukas et al., 2022] between medical and natural images is large as shown in Figure 1.3.

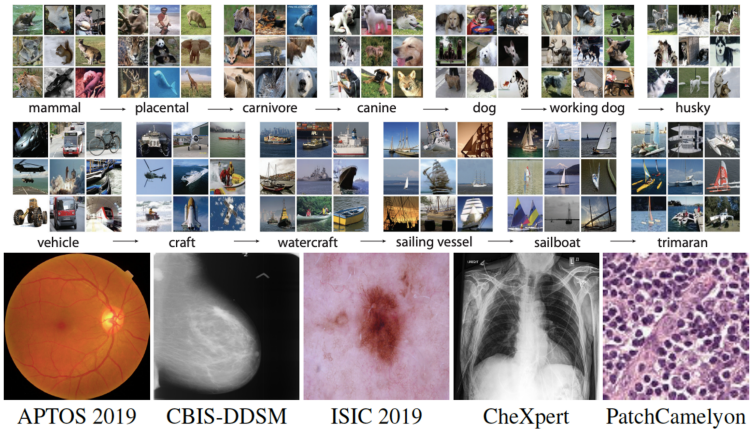


Figure 1.3: Domain gap between ImageNet (top image from Deng et al. [2009]) and medical images (bottom image from Matsoukas et al. [2022]).

To overcome the data domain differences between natural and medical images, adaptation methods have been developed [Guan and Liu, 2021]. Their goal is to transfer knowledge learned from data in a source domain to data in another target domain.

Differently from transfer learning, the learning tasks are the same. Some methods used in medical imaging propose to learn domain invariant features than can be transferred to the target domain. Dou et al. [2019] train a model to learn clustering and class alignment in the latent space using data from different domains. Liu et al. [2020] use meta learning to simulate domain shift and learn the shape of interest. These methods still need an annotated dataset to be trained on the target domain which sometimes lead to domain generalization problems when labels are scarce.

Recently, large datasets of healthy images have emerged such as UK Biobank [Littlejohns et al., 2020] and OpenBHB [Dufumier et al., 2022b]. These datasets are fuel for self-supervised learning methods which propose to build annotation-free pretext tasks used to pre-train deep learning models (see Figure 1.4). We focus particularly on this family of methods which are presented in greater length in Chapter 2.

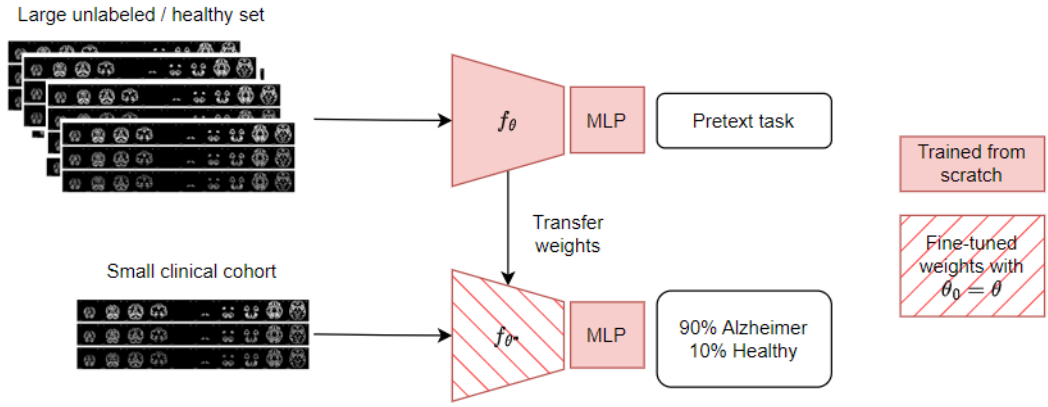


Figure 1.4: Self-supervised learning method.

In a concrete industrial and clinical scenario where one wants to build an algorithm to detect some pathologies on medical images to be used by radiologists in their daily workflow, the available data (annotated or unannotated) can be very heterogeneous compared to research dataset [Loizillon et al., 2023]. We investigate the specific problem of automatic prostate cancer lesion detection on MRI with deep learning which is an application with high clinical implications and where available datasets can be quite heterogeneous.

Prostate cancer is the second most common cancer in men worldwide [Bray et al., 2018]. Its early diagnosis is crucial for efficient treatment. Multi-parametric MRI (mp-MRI) is widely used for lesion detection and has been shown efficient for diagnosis [Rouvière et al., 2019]. The Prostate Imaging Reporting and Data System (PI-RADS) [Turkbey et al., 2019] has standardized interpretation of prostate MRI by defining a score for lesion malignancy. However, this score is subject to high inter and intra-reader variability which can highly affect deep learning algorithm performances [Güneş et al., 2023]. Leveraging annotation variability and confidence in training fully or self-supervised approaches can thus help improve performances and allow to use of imprecise data during training thus reducing the need for more annotations.

1.2 Goal

The goal of this thesis is to propose and develop methods to limit the annotation load in medical imaging while maintaining a high performance of deep learning algorithms. The research questions we are answering throughout this work are:

1. Many self-supervised learning methods have been designed using various perturbations often quite generic and unrelated to the objective task. How can perturbations used in these methods be optimized to benefit the most the supervised target task? Can some amount of supervision benefit self-supervised pre-training (*i.e.,* semi-supervised)? In Chapter 3, we introduce a perturbation generator pre-trained in a self-supervised framework with some amount of supervision. We show that the proposed pre-training method leads to latent representations of better quality.

2. A large amount of metadata, clinical information, and labels of various sources and confidence are available along with unlabeled images. These available data should be included in pre-training as they can be quite informative of pathology characteristics. How can labels of varying confidence be taken into account in semi-supervised pre-training? In Chapter 4, we propose a specific loss taking global labels provided by multiple annotators into account. We show substantial improvements in fine-tuning performances on two datasets with few annotations.

3. Self-supervised learning methods have widely been applied for classification tasks: how can the methods proposed in this thesis be applied to segmentation tasks? We present preliminary results in Section 4.6.

1.3 Publications

International workshops

(1) Camille Ruppli, Pietro Gori, Roberto Ardon, Isabelle Bloch. *Optimizing Transformations for Contrastive Learning in a Differentiable Framework.* Accepted for a poster presentation at Medical Image Learning with Limited and Noisy Data. MILLanD workshop 2022.

(2) Camille Ruppli, Pietro Gori, Roberto Ardon, Isabelle Bloch. *Decoupled conditional contrastive learning with variable metadata for prostate lesion detection.* Accepted for an oral presentation at Medical Image Learning with Limited and Noisy Data. MILLanD 2023.

National conferences

(1) Camille Ruppli, Pietro Gori, Roberto Ardon, Isabelle Bloch. *Optimisation des perturbations pour l'apprentissage contrastif.* Accepted for an oral presentation at 28^{eme} Colloque sur le Traitement du Signal et des Images (GRETSI), 2022

Journal article

Article on prostate cancer lesion detection with contrastive learning (Chapter 4) in preparation

Chapter 2

Related works

Contents

As introduced in Section 1.2, our goal is to reduce annotation costs to train neural networks for medical imaging problems. We focus on self-supervised learning methods that have been introduced to take advantage of the large amount of unannotated data available. In these methods, a neural network is pre-trained on pretext tasks that require no supplementary annotations. These methods can be split into three categories: generation-based (Section 2.1), context-based (Section 2.2), and self-distillation approaches (Section 2.3). While developing methods to reduce annotation costs, some amount of supervision can be introduced to guide the model. This supervision can be subject to annotator variability, especially when dealing with medical data. In Section 2.4, we present fully and self-supervised methods dealing with annotation variability and confidence. In the following sections, we give a general presentation of existing related works. Some of these methods, closer to our work, will be presented in larger details in the following chapters.

2.1 Generation based pretext tasks

Most pretext tasks entail applying a transformation sampled within a list of set transformations with fixed parameters. A perturbed (or transformed) image is then generated. In generation-based tasks, the perturbed image is fed to the network which has to generate or reconstruct the original unperturbed input image. The network is pre-trained on this reconstruction task and will then be fine-tuned on the supervised target task

with the available annotated data. An example of this approach for fine-tuning on a classification task on prostate MRI is given in Figure 2.1.

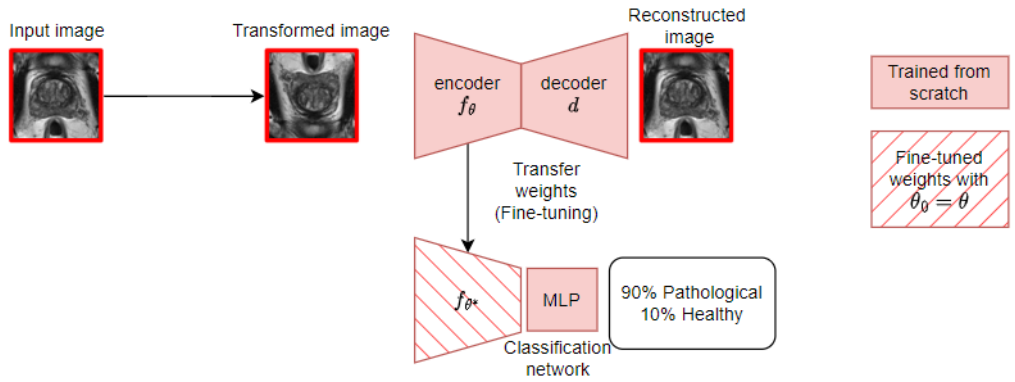


Figure 2.1: Example of a generation-based pretext task.

Exploring existing methods thus induces looking at (i) transformation types and (ii) the link between these transformations and the supervised target task.

2.1.1 Transformation and reconstruction types

Some methods train a Generative Adversarial Network [Goodfellow et al., 2014] (GAN) to perform inpainting: recovering the missing part of an image. The following cited works are typical examples, yet nonexhaustive. Pathak et al. [2016] train a GAN to recover the input image missing block while preserving the whole image context with an adversarial loss function. Tao et al. [2020] use the same GAN framework to reconstruct 3D medical volumes corrupted with a "Rubik's Cube" perturbation: shuffling and rotating sub-volumes. The goal of these approaches is to give the network context information on the input image. However, GAN approaches are costly to optimize, subject to mode collapse, and need a lot of training data to converge [Saxena and Cao, 2020].

Some works thus propose to solve the reconstruction task without a GAN architecture. For natural images, colorization has widely been applied [Zhang et al., 2016, Larsson et al., 2017]. The input image is converted to grayscale, only keeping the intensity information, and the network is trained to recover color channels. These methods do not apply to medical images as they differ, by nature, from natural images and solve different problems.

Many papers propose to build auto-encoders [Hinton and Salakhutdinov, 2006] to reconstruct perturbed images especially in medical imaging [Zhou et al., 2019b, Tardy and Mateus, 2021, Feng et al., 2020, Chen et al., 2019]. In the work by Chen et al. [2019], the input image is perturbed by swapping randomly selected patches, and the network is trained to reconstruct the input image.

Zhou et al. [2019b] introduce a novel set of perturbations better suited for medical images: Models Genesis. A nonlinear pixel intensity transformation is applied to learn organ appearance. Pixels are shuffled locally and the reconstruction to such perturbation leads the model to learn local boundaries and texture. Out-painting is performed by selecting an image patch at random and assigning a random value out to pixels

outside this patch. By reconstructing the original image after out painting, the model must learn global geometry and organ layout. In-painting is, reversely, performed by filling a randomly selected patch and filling it with random values. Reconstructing the original image after in-painting makes the model focus on local continuities. Figure 2.2 shows examples of the aforementioned perturbations.

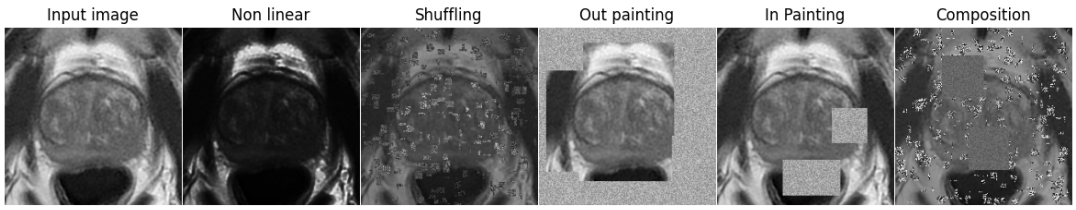


Figure 2.2: Models Genesis perturbations[Zhou et al., 2019b].

These perturbations are composed and applied to input images with some probability. A U-Net [Ronneberger et al., 2015] is then pre-trained to reconstruct the original image. This pre-training has led to great performances on a variety of tasks: lung nodule segmentation, liver segmentation, or pulmonary diseases classification, among others.

Reconstruction methods are a powerful pre-training tool especially when using appropriate perturbations but finding how to best compose and configure them for optimal performances in the subsequent task remains unclear.

2.1.2 Link with the supervised target task

To overcome this limitation some works introduce perturbations better linked to the target supervised task while adding some amount of supervision in pre-training. Tardy and Mateus [2021] propose the generation of synthetic artifacts to perturb the input mammogram image which has to be reconstructed by the network while separating normal from abnormal content. These synthetic artifacts are masses, distortions, and clusters specific to breast cancer. A weakly supervised loss function is also introduced to classify input images as normal or abnormal when labels are available. Pre-training is then better linked to the target supervised task.

Generation-based pretext tasks such as GANs or colorization have proven successful with natural images. Given the different nature of medical images, some works have introduced specific pretext tasks such as those proposed by Zhou et al. [2019b]. These tasks, although well adapted to the image nature, are unrelated to the supervised target tasks. Their parameters and application probability are arbitrarily fixed, which is not satisfactory as some transformations could hurt performances. A link to the supervised task is necessary to choose relevant transformations. This link has been introduced by Tardy and Mateus [2021] but the transformations applied are specific to one pathology.

2.2 Context based pretext tasks

The other family of pretext tasks taken from Jing and Tian [2019] is context-based. Images contain context information that can be useful for the supervised target task. Learning context has been done by training a model to predict a perturbation parameter (Section 2.2.1) or to learn similarity in the latent space (Section 2.2.2).

2.2.1 Relative position and parameter prediction tasks

Prediction tasks

The most common context pretext task consists of predicting the relative position of two randomly selected patches in the image [Doersch et al., 2015]. Many works build on this idea by introducing more complex tasks such as Jigsaw puzzle [Noroozi and Favaro, 2016]: the image is divided into patches which are permuted and the model is trained to predict the applied permutation. The rationale is that the model will learn spatial context while trying to predict patch positions which can be useful for object classification or detection.

Parameter prediction has been applied to other perturbations than patch swap. Gidaris et al. [2018] propose to rotate images and train the model to predict the applied rotation angle. Figure 2.3 shows a schematic example of this approach. The model is then supposed to learn the object's location and pose in the image.

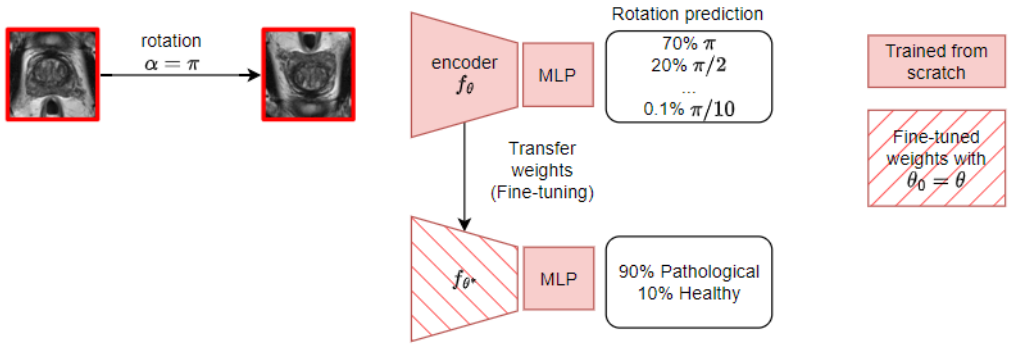


Figure 2.3: Prediction pretext task.

For both context and parameters prediction methods, perturbation parameters such as patch size, overlap or rotation angle must be chosen among a large number of possible permutations. No optimization is done on this set of parameters and their impact on the objective supervised task is not explored.

Application to medical images

The aforementioned prediction tasks have widely been applied to medical images where spatial information is of utmost importance to detect organs or pathologies. Riva et al. [2022] especially show that U-Net models implicitly use spatial relationships between objects to obtain high segmentation performances without the use of a separated task which underlines the importance of context. Blendowski et al. [2019] build on the method proposed by Doersch et al. [2015] to learn context in 3D Magnetic Resonance Images (MRI). They extract two 2.5D sub-volumes in the coronal plane which are chosen without overlap. The second volume position is further shifted along the normal direction which gives two offsets δ_1 and δ_2 that are to be predicted by a decoder. Bai et al. [2019] define nine anatomical positions on 2D Cardiac MRI. A U-Net is trained to segment these nine anatomical boxes. During fine-tuning, to not forget about information learned during pre-training, the pretext task loss function is computed along with the supervised task one.

These pretext tasks solve the issue of not being related enough to the objective task but lack generality and cannot be applied to other supervised tasks.

Limitations

Given the variety of proposed approaches to pre-train a model, it is hard to find which pretext task is the most suitable for pre-training. To mitigate this issue, some methods propose to combine different tasks during pre-training to take the most advantage of each approach. Kim et al. [2018], Doersch and Zisserman [2017] combine different self-supervised tasks. After a shared network backbone, a smaller network is added for each pretext task. As different tasks can require different types of information, Doersch and Zisserman [2017] introduce a Lasso penalty to select which features from the backbone network get into each specific layer task. The goal is also to select features that will be more useful for the subsequent supervised task.

Although self-supervised learning methods have proven successful [Goyal et al., 2019, Kolesnikov et al., 2019], their performances struggle to reach those of fully supervised learning. One possible explanation is that they are agnostic to the supervised target task and could benefit from some amount of labeled data. This has been done in some works combining Jigsaw puzzle [Carlucci et al., 2019] or rotation prediction [Zhai et al., 2019] with a supervised branch predicting image class with a small amount of labeled data.

Adding some supervision to increase the link with the supervised target task is a first approach to solve the aforementioned limitations. However, it remains unclear why a particular task among relative position, parameter prediction, and generation-based (see Section 2.1) would be most beneficial to the target supervised task.

Some authors, such as Wu et al. [2018], have thus proposed instance-level discrimination tasks: learning similarity between instances in the latent space. These approaches are mostly based on the noise contrastive estimation loss function [Gutmann and Hyvärinen, 2010] which leads to grouping them as contrastive learning methods.

2.2.2 Contrastive learning

To solve the relevance issue presented by generation and prediction-based pretext tasks, contrastive learning approaches have been introduced to solve a new pretext task. Each image is considered as an instance of its own and the model is trained to distinguish between the different classes i.e. images.

The exemplar task, introduced by Dosovitskiy et al. [2014], is the pretext task closest to contrastive learning. Input images are divided into patches, these patches are perturbed by sampling randomly in a set of transformations with a set of parameters. A surrogate class, the input image generating the perturbed patches set, is defined. The model is then trained to predict this surrogate class. By predicting the surrogate class, the model is expected to become invariant to the applied transformations and to learn semantic similarity between images.

As mentioned by Wu et al. [2018] the exemplar task is parametric whereas contrastive learning approaches are mostly nonparametric.

The following paragraphs present: (i) base contrastive learning methods, (ii) pair sampling, (iii) use of auxiliary information, (iv) use of noisy labels in contrastive learning, and (v) contrastive learning in medical imaging.

Base contrastive learning methods

He et al. [2019], Wu et al. [2018] train a neural network to solve the instance discrimination task: the network is trained to map semantically similar examples close in the latent space. Each image instance is treated separately. A memory bank is built [Wu et al., 2018] or updated during training [He et al., 2019] and stores image latent representations with which an input image will be compared using the noise contrastive estimation loss function [Gutmann and Hyvärinen, 2010]. Figure 2.4 shows a schematic view of these approaches.

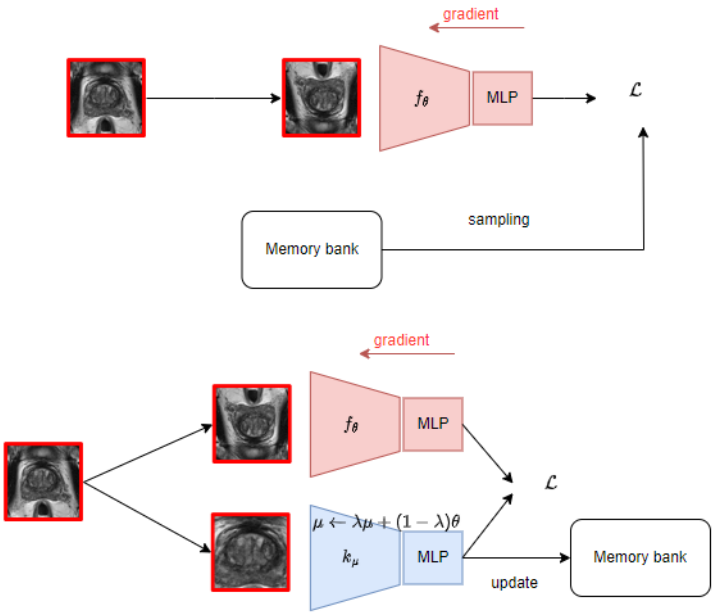


Figure 2.4: Schematic views of memory bank approaches. Wu et al. [2018] (top): the memory bank contains the latent representations of all the dataset, at each training iteration instances are sampled in the memory bank to compute the loss and the memory bank is updated with the latent representations of the current batch. He et al. [2019] (bottom): the memory bank is updated at each iteration with the latent representations generated by the momentum encoder k_μ which parameters are updated through momentum update.

He et al. [2019], Misra and van der Maaten [2019] build instances by applying two randomly sampled perturbations to the input image.

In [Chen et al., 2020a], the memory bank is removed and stochastic data perturbation is introduced to create instances. The context is learned in the latent space through similarity between positive pairs and dissimilarity between negative pairs. Positive pairs are defined as two perturbed versions of the input image while negative pairs are defined as perturbed versions of all other images. The instance-level discrimination task can then be seen as an invariance task: the neural network is trained to consider similar two perturbed versions of the same image, thus becoming invariant to the

applied perturbation.

The best results are obtained when augmenting the input image batch twice (rather than keeping one unperturbed version) and when composing multiple augmentations. This composition, however, is done at random with sampling probabilities and perturbation parameters that are arbitrarily set. A large batch size is also needed to obtain better results, which has been criticized by later works and becomes problematic when working with 3D images and network architectures for medical images.

To remove the batch size constraint, Zbontar et al. [2021] replace the cross entropy involved in the contrastive loss function by variance reduction. Some other approaches remove negative pairs altogether, as will be presented in Section 2.3.

From these methods some questions arise: which perturbations composition to apply? is invariance to all the applied perturbations relevant for the target task? how to sample positive and negative pairs? Most of the aforementioned papers investigate the impact of removing some perturbations or changing perturbation parameters on the supervised target task performances. However, only few papers optimize these perturbations during pre-training [Tian et al., 2020] or question the invariance approach [Xiao et al., 2021, Purushwalkam and Gupta, 2020]. Tian et al. [2020] show that perturbations can be optimized to find a balance between the amount of information shared between perturbed versions of images and the relevant information needed for the supervised target task. We present this approach in more detail in Chapter 3.3. The work of Xiao et al. [2021] and Purushwalkam and Gupta [2020] show that invariances to some perturbations could be detrimental to the supervised target task and propose building different latent spaces and transformations.

Exploring pair sampling

For the instance discrimination task to be a useful pre-training tool, considering one instance per image might be sub-optimal. In a dataset, different images can share semantic information and class labels. Structuring the latent space by pushing these similar instances apart might be detrimental to the target supervised task. To counter this limitation many works have investigated how to better sample positive and negative pairs.

In a fully supervised setting, Khosla et al. [2020a] condition positive pairs selection with class labels: latent representations of samples with the same class labels will be attracted together. Cho et al. [2021] follow the same line of work by assigning a higher dissimilarity weight to negative pairs with a different label from the input image than negative pairs with the same label. These methods are not self-supervised anymore and introduce a new fully-supervised loss function to structure the latent space. However, they propose an interesting approach to condition positive and negative pairs sampling with the available information.

Negative pairs sampling has also been investigated in many works either through designing a specific sampling distribution [Robinson et al., 2020] or by computing similarity between alleged negatives and a support set of multiple perturbations of the input image [Huynh et al., 2022]. This support set contains latent representations of the input image perturbations and is used to increase comparison diversity. Alleged negatives are compared with the two sampled perturbed versions of the input image but also with

other perturbed versions in the support set.

Building a support set of latent representations has widely been done in contrastive learning to increase generalization and diversity in pairs sampling. This support set can be fixed or updated during training as done with memory bank approaches. Wei et al. [2020] build a support set with every image but the two perturbed input images. Similarity with the support set is then used as a consistency regularization term in the loss function. In [Dwibedi et al., 2021], the support set is implemented as a queue (first-in first-out) containing latent representations of each batch of images from each training step. It is built to be large enough to approximate the dataset distribution. It is used to find the nearest neighbor of one perturbed version of the input image. This nearest neighbor will then be used in the contrastive loss, replacing one input image perturbation, in positive and negative pairs. Figure 2.5 shows a schematic view of this approach.

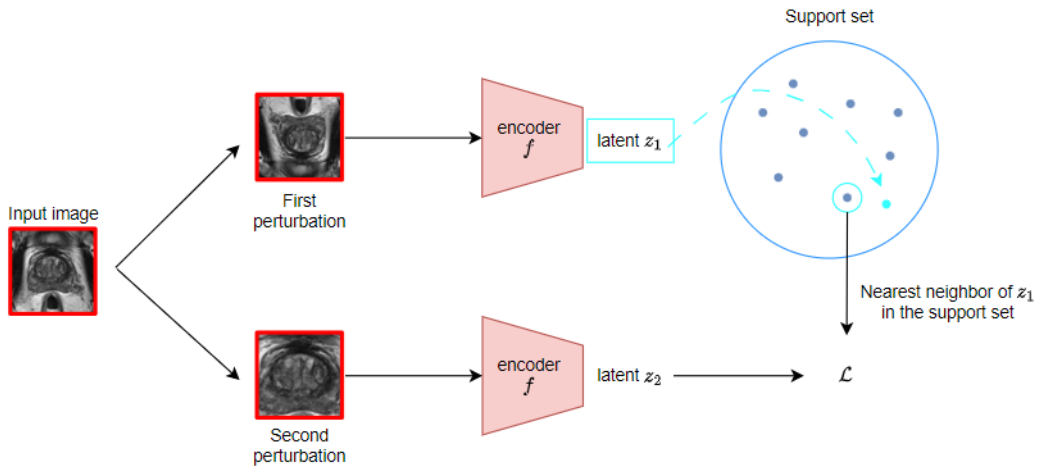


Figure 2.5: Nearest neighbor Contrastive Learning (nnCLR) approach [Dwibedi et al., 2021].

Building on this work, Ge et al. [2023] introduce soft neighbors: k nearest neighbors are selected from support set similarities, and a softmax function is applied to latent representations of the neighbors and the input image. The function output, named positiveness, is used to weigh each neighbor's positive contribution in the contrastive loss. These nearest neighbors approaches yield satisfying results compared to other state-of-the-art contrastive learning methods.

However, some questions remain on support set design: which data to include in the support set? How to make sure that the latent space is sufficiently well structured for nearest neighbors to be meaningful?

Contrastive learning with auxiliary information

Another approach to better structure the latent space has been to include auxiliary information in pair sampling. Rather than uniformly attract or repel positive and negative pairs, a kernel is introduced to condition similarity. Image attributes, such as bird color [Tsai et al., 2022], or medical metadata, such as patient age [Dufumier et al., 2021b], are fed to a kernel computing similarity between attributes of different

Figure 2.6: Three different dog breeds (top row) and three cases of cardiomegaly (left), hernia (middle) and emphysema (right) from Chest Xray dataset [Wang et al., 2017] (bottom row). The three chest images look much more similar than the three dog images.

images. Positive pairs sampling is conditioned on the computed similarity: positive pairs are attracted proportionally to the kernel output value. To avoid repulsing data with similar attributes but not sampled as positive pairs, negative pairs repulsion can also be conditioned with attributes dissimilarity [Dufumier et al., 2021a]. Kernel values are also applied to negative pairs and samples are repulsed proportionally to attributes dissimilarity. Going one step further, positive and negative pairs sampling can also be conditioned with attributes generated from a previously learned VAE or GAN distribution [Dufumier et al., 2022a].

Contrastive learning in medical imaging

Most of the introduced contrastive learning methods have been applied to medical images because labels are scarce and costly to obtain, as annotating a medical image often requires expert knowledge. However, they cannot always be applied straightforwardly as medical images have different characteristics from natural images.

They encode strong context information: to detect, segment, or localize a lesion on a medical image, knowing the relative position of different organs is essential as lesions can change their structure.

Some papers have thus developed pretext tasks designed to learn these strong spatial relationships. Li et al. [2021a], Yan et al. [2020] combine base contrastive learning objective with a loss enforcing similarity between neighboring or overlapping image patches. Zeng et al. [2021] define image similarity through slice position in 3D volumes: given two 3D volumes of two different patients, 2D slices are extracted from both volumes, and their relative position along the z axis is computed. If the position difference is small, then slices are considered similar.

In medical imaging, differences between healthy and pathological images can be much scarcer than differences between natural images of different classes. When trying to detect different pathologies on Chest Xray, differences between pathologies can be much subtler than between different dog breeds as shown in Figure 2.6.

Some papers have thus proposed to model the problem as anomaly detection: Bozorgtabar et al. [2020] structure the latent space, at first, with healthy images only, then small pathological examples are introduced and pushed away from healthy representations. Dufumier et al. [2021b] propose to learn healthy brain representations by conditioning on the a priori knowledge that anatomical structures of healthy brains are similar in patients of close age. Fine-tuning on a dataset with a specific pathology will benefit from this anatomical knowledge of healthy patients.

Using patient metadata, such as age, in contrastive pre-training, sometimes implies dealing with noisy and diverse data. Some works have proposed to include this variability in pre-training: using radiological labels to improve prediction on histological ones [Sarfati et al., 2023], combining contrastive loss with weak supervision [Lubrano et al., 2022] or dynamically adapting samples importance in contrastive loss to account for label noise [Peng et al., 2021]. These metadata can be provided by multiple doctors and can thus present inter and intra-annotator variability which is not taken into account in the existing works.

Most of the previously mentioned approaches are well adapted to image classification. But contrastive pre-training has been much less applied to segmentation problems which are quite common in medical imaging. Some works have proposed methods to perform efficient pre-training for medical image segmentation. Most of the proposed approaches introduce ways to pre-train the decoder along with the encoder. Zheng et al. [2021] propose to combine contrastive learning at different scales in the encoder with image reconstruction done by a decoder.

Some works introduce a local contrastive loss in the decoder to contrast latent representation at the pixel level, which is deemed more accurate for the segmentation task. One of the state-of-the-art approaches is proposed by Chaitanya et al. [2020] which combines the base global contrastive loss function with a local loss term on patches from decoder feature maps. The idea behind the local loss term is to bring closer, in the latent space, the representations of patches from the same local region of the feature map. Hu et al. [2021] add a supervised constraint to this local loss: pixel patches are considered similar if they share the same segmentation label, when available. A supervised segmentation branch is added to this framework by Chaitanya et al. [2023]. Supervised segmentation training is performed for a few epochs with available annotated data. The trained segmentation network is then used to produce pseudo labels on unannotated data. Supervised local contrastive loss is applied at the pixel level with the available segmentation masks and pseudo labels.

2.3 Self-distillation methods

Self-distillation methods use two different neural networks to perform pre-training: online (student) and target (teacher) networks. They only use positive pairs and avoid collapse (i.e the two networks outputting the same constant representation) using asymmetric architectures and different optimization procedures (such as Exponential Moving Average or stop gradients) for both networks [Zhang et al., 2022, Tian et al., 2021, Garrido et al., 2022].

In the approach of Grill et al. [2020], the online network is an encoder followed by a projector and a predictor models (usually built as two or three layers multi-layer perceptron) and the target network is another encoder followed by another projector. The parameters of the online (p_o) and target (p_t) networks are different and the target network parameters are updated using an exponential moving average (EMA): $p_t = \tau p_t + (1 - \tau)p_o$. Gradients are not back-propagated through the target network (stop-gradient). Both the online and target networks are then optimized to maximize agreement (measured as the cosine similarity) between the latent representations of two perturbed versions of the same image. Figure 2.7 shows a schematic view of this approach.

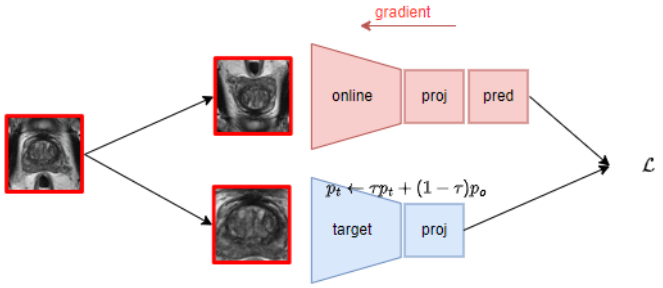


Figure 2.7: BYOL approach [Grill et al., 2020].

Chen and He [2020] use a similar architecture but remove the EMA. To avoid collapse, only stop-gradient is applied to the target network.

Caron et al. [2021] introduce DINO which has a similar architecture to the two aforementioned approaches but no predictor is used for the online network and the encoder is a vision transformer [Vaswani et al., 2017]. They transform the projector outputs into "soft labels" by feeding them to a softmax function. A cross-entropy loss function is then computed between the soft labels of online and target networks from two perturbed versions of the input image. EMA is performed to update the parameters of the teacher model.

For the three aforementioned methods, in contrast to contrastive learning approaches, there is no agreement minimization between latent representations of different images. Although these methods do not need negatives and avoid collapse using different solutions, it is still not clear why they work, as shown by Zhang et al. [2022].

2.4 Annotation variability and confidence

When building natural image datasets for training a deep learning neural network, to retrieve as many annotations as possible, some annotations can be provided by large-scale mining of search engines [Li et al., 2017] or downloading social media images with labels [Mahajan et al., 2018], which leads to label noise. In medical imaging, building large datasets with high annotation quality is even more difficult as annotations require expert knowledge. Furthermore, some tasks are subject to inter and intra-observer variability which necessitate obtaining a consensus label [Bridge et al., 2016].

For both natural and medical imaging tasks, many methods have been proposed to tackle annotation variability and label noise during training [Karimi et al., 2019]. In this section, we present some methods addressing this issue, again without being exhaustive.

Annotation variability in fully-supervised learning

When training a fully-supervised model for a classification task, noisy data have been taken into account in Learning with Noisy Labels (LNL) methods. The most common LNL approach, DivideMix [Li et al., 2020a], uses a Gaussian mixture model (GMM) to divide training data into clean (labeled) and noisy (unlabeled) samples which are fed to two parallel networks (see Figure 2.8). New labels are guessed for data with noisy labels by ensembling predictions from both networks. MixMatch approach [Berthelot et al., 2019] is then used to train the model: both labeled and unlabeled examples are linearly mixed and added to the set of training data as an augmentation.

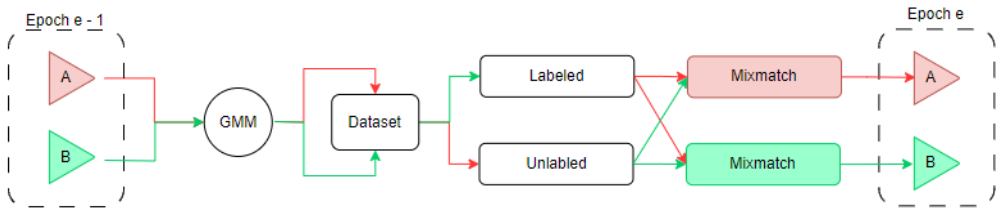


Figure 2.8: DivideMix approach [Li et al., 2020a].

To deal with annotator variability for a medical image classification task, Jiménez-Sánchez et al. [2019] uses curriculum learning [Bengio et al., 2009]: starting to train the model with "easy" samples, i.e with high kappa score (synonym of high inter-rater agreement), and adding harder samples gradually.

For fully-supervised segmentation, different strategies have been proposed in the literature. Many of the existing methods [Cardoso et al., 2013, Yang et al., 2023] are based on the STAPLE algorithm [Warfield et al., 2004] which assigns an accuracy weight through majority voting to each segmentation.

Mirikharaji et al. [2021] propose to build disjoint subsets of annotations containing one unique annotation per image. Each subset is associated with a model trained on a union of all annotations available. During training, the annotations subsets are used to build weight maps to cancel the contributions of inconsistent annotations from the union. At inference, a fusion of the different models' predictions is performed to generate the final prediction.

Another line of research is proposed by Zhang et al. [2023] through jointly learning the expert consensus label and the characteristics of each annotator using two parallel neural networks.

Contrastive learning with noisy labels

Contrastive learning has proven a successful tool to increase robustness against label noise [Xue et al., 2022]. Zheltonozhskii et al. [2022] build on DivideMix [Li et al., 2020a] by adding a contrastive learning step before LNL. Yi et al. [2022] use the contrastive loss function as a regularization term along the supervised binary cross-entropy loss

function. Regularization aims to learn similar representations for data with similar linear classifier probability.

A similar approach is proposed by Wang et al. [2022a] for partial label learning. MoCo [He et al., 2019] architecture is used along a classifier head. The output of the classifier head is used to sample positive pairs: samples with similar classifier predictions are defined as positive. Class prototypes are built, updated during training, and used to update pseudo-labels with which a classification loss is computed. More elaborate approaches use a two-step training by starting with contrastive learning followed by one or more fine-tuning steps.

Ciortan et al. [2021] use unsupervised contrastive learning followed by fine-tuning with noise robust classification loss as a first training step. Pseudo labels are generated from the fine-tuned model, a Gaussian mixture model is applied after fine-tuning to evaluate pseudo labels correctness. The correctness output is then used to weight positive pairs in supervised contrastive learning. A final classification fine-tuning is then performed.

A notion of confidence is introduced by Li et al. [2022b]. For each sample, pseudo labels are computed for each possible class by averaging labels of the nearest neighbors in the latent space (with respect to cosine distance). An unsupervised contrastive learning step is performed during the first epochs to structure the latent space.

Pseudo labels are used to approximate class posterior probability. A set of confident samples is built by computing the cross-entropy loss using original labels and posterior probability. Supervised contrastive learning is then performed by sampling positive pairs in the confident set. A classification loss is also added to confident examples. Pseudo labels and nearest neighbors definition are here strongly linked to the unsupervised contrastive training phase which does not guarantee a sufficient latent space structure for subsequent pre-training.

2.5 Conclusion

To take advantage of the large amount of unannotated data available, self-supervised learning methods have been proposed to pre-train neural networks without the need for manual annotations. Generation-based (Section 2.1), context-based (Section 2.2), and self-distillation (Section 2.3) approaches often use perturbations sampled at random in a fixed list with fixed parameters. How to best combine and choose these perturbations and their parameters remains unclear. Furthermore, some perturbations can be detrimental to the target supervised task. Some works mitigate this issue by designing pretext tasks for a specific supervised task, especially in medical imaging. However, these tasks do not generalize well to other problems. A balance has to be found between perturbation or pretext task optimization for a given supervised problem and method generalization ability.

Among context-based methods, contrastive learning approaches (see Section 2.2.2) propose an instance-level discrimination task: the latent space is structured with instance similarity. Defining instance similarity is the main challenge of these approaches and has been widely explored. When defining similarity through perturbed versions of the same image, the same questions of perturbation optimization arise. In Chapter 3, we introduce a perturbation generator optimized for contrastive pre-training guided by a

small amount of supervision.

Class labels and metadata have been used to condition instance similarity, but these data can be subject to annotator variability, especially in the medical domain.

In the methods summarized in Section 2.4, the confidence is mostly based on loss function values. However, confidence in labels and metadata is often linked to a priori domain knowledge such as data acquisition, annotators' experience, and agreement. This is even more relevant for medical data.

In Chapter 4, we design an adapted contrastive loss introducing annotation confidence for the specific problem of prostate cancer lesion detection.

Chapter 3

Perturbations Optimization for Contrastive Learning

Contents

As introduced in Chapter 2, self-supervised learning methods have been developed to take advantage of unannotated data during pre-training to increase performances on the objective supervised task when only a few annotated data are available. Many of these methods use perturbations that are often sampled at random among a fixed perturbation list with fixed parameters. Few works, such as the ones by Tian et al. [2020], Seyfiouglu et al. [2022], have investigated perturbation optimization in contrastive pre-training. These perturbations, which can be seen as the counterpart of data augmentation in fully supervised learning, should be optimized to be most beneficial to the supervised target task.

Looking for the best pretext tasks or perturbations implies finding an order between them. To that end, we first tried to create pretext task clusters. This approach has not been successful but is presented in Appendix A.

In this chapter, we first present the existing methods to optimize data augmentation in fully and self-supervised learning (Section 3.1), we then introduce in greater detail the contrastive learning framework in Section 3.2 and we present our proposed perturbation optimization method in Section 3.3.

3.1 Automatic augmentation methods

Data augmentation has widely been applied in fully supervised learning to increase the amount of available training samples [Krizhevsky et al., 2012, Devries and Taylor, 2017]. Finding the most appropriate augmentation method is, however, challenging and time-consuming if done manually. Some works have thus proposed ways to automate this search. One of the most common approach is through reinforcement learning: AutoAugment, proposed by Cubuk et al. [2019], combines two different augmentations which are defined by their parameter magnitude and application probability to define sub-policies. The sub-policies search space is then defined by all the combinations of perturbation parameters and probability values. A reinforcement learning network is trained to find five optimal sub-policies performing best on an external validation set. This approach has been simplified in RandAugment by Cubuk et al. [2020] where the search space is reduced to only two parameters: the number of perturbations to apply and perturbation magnitude, which are set equal for all selected perturbations. Optimal parameters are then found with a simple grid search.

Many works have later built on these approaches, introducing new network layers or optimization schemes. Rommel et al. [2022] introduce an augmentation layer learning perturbation parameters and weights balancing each perturbation contribution.

Li et al. [2020c] (DADA method) use a differentiable optimization model rather than reinforcement learning. They expand the search space introduced in AutoAugment to a joint distribution rather than a discrete parameter space. Augmentation policies are treated as sampling from Categorical and Bernoulli distributions. Data augmentation optimization is then modeled as a Monte Carlo problem. Liu et al. [2021a] explicitly model the augmentation probability and shows performance improvement compared to DADA.

These methods are well suited for fully supervised learning as the optimization framework is based on the supervised validation loss. However, these approaches would not be suited to contrastive or self-supervised learning where annotations are scarce.

Some papers have thus investigated how to optimize perturbations applied in contrastive learning. To optimize contrastive learning perturbations, their characteristics have been investigated. Patrick et al. [2021] propose a theoretical framework to best compose transformations in self-supervised learning by defining a contrast function enforcing invariance or distinctiveness to a set of transformations. Koyama et al. [2021] show that an augmentation random variable can be introduced and optimized along with the network to find the perturbation to apply. They add an entropy constraint to enforce generalization on the supervised target task. They validate this framework by optimizing crop positions to apply with a softmax function. Both the aforementioned approaches introduce a theoretical framework rather than practical generic methods to optimize perturbations in contrastive learning.

Bridging the gap between theoretical analysis and practical solutions, Tian et al. [2020] define optimal contrastive perturbations as nonredundant and useful for the subsequent supervised task and introduce a perturbation generator. They train a color space perturbation generator in an adversarial framework.

Similarly, Tamkin et al. [2020] generate perturbed images from a noise distribution and define a distortion budget to constrain the strength of applied perturbations. Both these approaches are based on generative networks which have small interpretability.

Without the use of a generative network, Shi et al. [2022] introduce a mask generator trained to occlude meaningful parts of the image. Mask sparsity is enforced to avoid trivial solutions. Both Tian et al. [2020] and Shi et al. [2022] introduce interesting and promising frameworks for perturbations optimization for contrastive learning. Our method, presented in Section 3.3, builds on these works.

Some works propose to apply perturbations in the feature space rather than directly to input images. Li et al. [2022a] add augmentation networks in the feature space optimized to find the most discriminative representation. Optimization is performed with a base contrastive loss function and a loss function including augmented features. Yang et al. [2022b] add adversarial attacks to a generative adversarial network latent space. The generated image, along with the input one, is used to train the contrastive learning framework. The gradient of the contrastive loss function is then used to update the strength of the adversarial attack.

These approaches also allow for little control of the perturbations. The link between latent and image space perturbations remains unclear: perturbing the input image does not translate to a similar perturbation function in the latent space and vice versa. These approaches thus lack explainability.

3.2 Introduction to contrastive learning

Among self-supervised learning methods, contrastive learning approaches propose to pre-train a model to learn invariance and similarity in the latent space. Different approaches have been proposed to achieve this goal but one of the most successful is simCLR [Chen et al., 2020a]. In simCLR, input images are transformed with perturbations sampled at random in a list with fixed parameters to produce two different *views*. These views are then fed to the neural network, and projected into its latent space. The noise contrastive estimation loss [van den Oord et al., 2018] is then used to bring closer, in the latent space, views of the same image while pushing apart views of different images: the instance discrimination task introduced in Section 2.2.2 A schematic view of this approach is shown in Figure 3.1.

More formally, given an encoder f, followed by a projection head g (in practice implemented as a nonlinear multi-layer perceptron), and sim the cosine similarity measure defined as $sim(u, v) = \frac{u^T v}{||u||||v||}$, the contrastive loss function writes :

$$\mathcal{L}_{NCE} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(X_i^1, X_i^2) = \frac{1}{N} \sum_{i=1}^{N} -\log \left(\frac{e^{sim(x_i^1, x_i^2)}}{\sum_{j, j \neq i} e^{sim(x_i^1, x_j^2)}} \right) \tag{3.1}$$

where $x_i^1 = g \circ f(X_i^1)$, $x_i^2 = g \circ f(X_i^2)$ are latent representations of the two views of an input image X_i and N the number of images.

van den Oord et al. [2018] showed that minimizing this loss function maximizes a lower bound of the mutual information shared between latent representations of different views of the same image. Minimizing this loss function thus enforces similarity between these different views and perturbation invariance is learned. Learning invariance

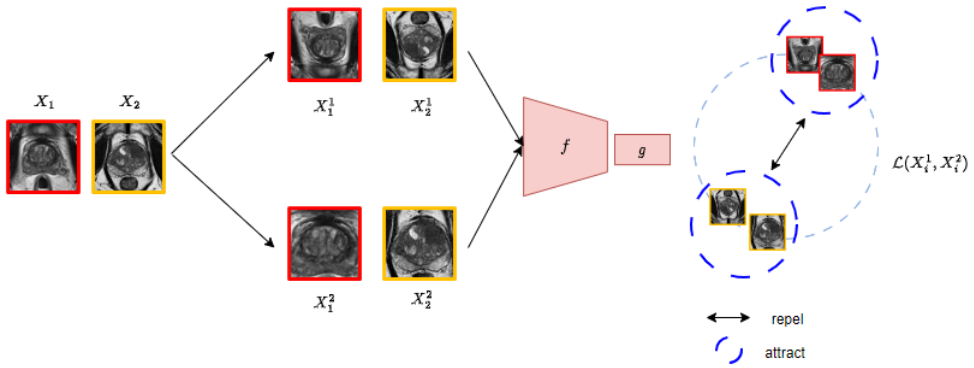


Figure 3.1: Schematic view of simCLR approach: a batch of images (X_1, X_2) is given as input, two perturbations are sampled at random to create two different views of the input batch (X_1^1, X_2^1), (X_1^2, X_2^2), these two views are fed to an encoder f followed by a projection head g (a nonlinear multi-layer perceptron), the model is trained to optimize the contrastive loss function $\mathcal{L}$ which goal is to attract latent representation of views of the same image while repelling latent representations of views of different images.

to perturbations can be detrimental to the supervised target task: Xiao et al. [2021] build perturbation-specific latent spaces because learning invariances to every perturbation might be adverse to classification performances. They show that, for a flower classification problem, learning invariance to color can hurt classification as it is a distinctive attribute of flower classes. This is even more relevant for medical imaging where pathologies are localized. Therefore, learning invariance to a crop perturbation that might remove the region of interest is counter-intuitive.

3.3 Perturbation generator optimization

Tian et al. [2020] introduce a "sweet spot" characterizing views for contrastive learning:

i mutual information between views should be small enough to avoid redundancy,

ii mutual information between views should be high enough to avoid discarding useful information for the supervised target task.

An example of such views is shown in Figure 3.2. On the top left, the two views are too redundant. On the bottom left, the two views are not redundant anymore but the tumor (shown in red overlay) is cropped from one view, and learning invariance to this perturbation will be detrimental to the tumor presence classification task. On the right a sweet spot is shown: the two views share a small enough amount of information not to be too redundant but tumor-relevant pixels are kept.

To find optimal views for pre-training, Tian et al. [2020] introduce a flow-based perturbation generator based on the work of Kingma and Dhariwal [2018] to generate images in different color spaces. This perturbation generator is not well-suited for medical images. To overcome this limitation, we introduce a differentiable perturbation generator not based on GAN (see Section 3.3.1), as Cohen et al. [2018] showed that synthesizing anatomically relevant images with generative models can be difficult. We have also tried a GAN approach which has not been successful but led to the idea presented in Section 3.3.1 and is presented in Appendix B.

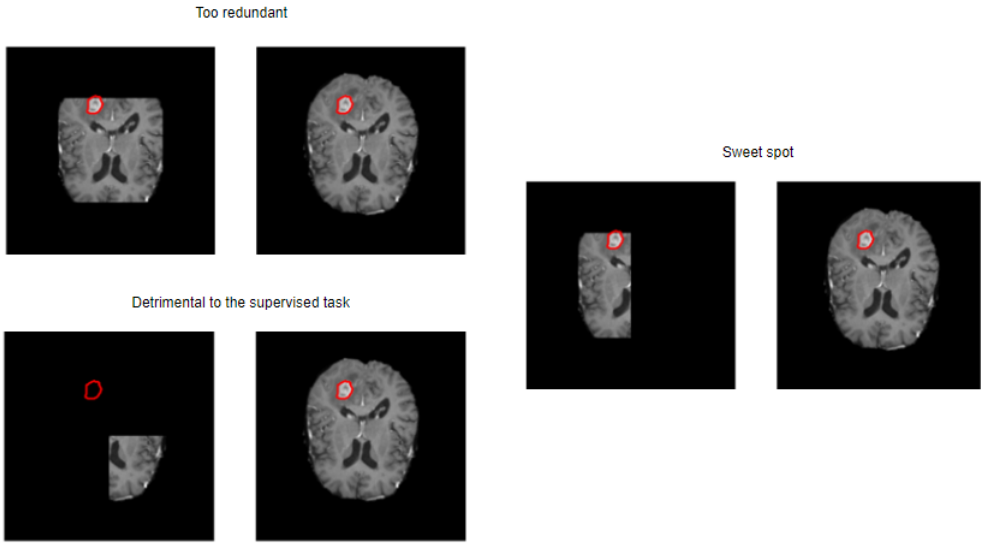


Figure 3.2: Example of views reaching the sweet spot (on the right) in contrast to views detrimental to the supervised task (on the bottom left) or too redundant (on the top left).

3.3.1 Perturbation network

We introduce a perturbation network (M) that minimizes the mutual information between images of a positive pair without compromising the supervised task performance. For each image of the training set, M, implemented as a neural network, outputs a set of parameters (Λ) defining the perturbations to apply (T_{Λ_M}). As in [Chen et al., 2020a, Xiao et al., 2021], the latent space of the encoder (f) is optimized using a projection head (g) into a lower dimension space where a contrastive loss function (I_{NCE}) is minimized. Supervision is added in the latent space using a linear classifier (p) that minimizes a classification loss function ($\mathcal{L}$). Figure 3.3 shows a schematic view of the architecture used (X denotes an image from the training set and X_M its transformed version).

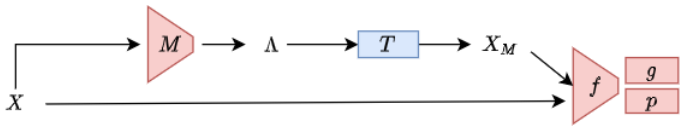


Figure 3.3: Proposed architecture (red color indicates a trainable element, blue color indicates a non-trainable element).

Optimizing perturbations

We consider a finite set of intensity and geometric perturbations acting on images. Each perturbation is parameterized by a vector of parameters (for example, the parameter vector of a 2D rotation around a fixed point only contains its angle). The perturbation function (T_{Λ_M}) is the composition of perturbations applied in a fixed order. The perturbation network (M) outputs the perturbation function parameters. We propose to train M to find the optimal perturbations for the semi-supervised contrastive problem. The network M maps an image to the space of parameter vectors normalized to $[0, 1]$.

Let λ_k be the vector of parameters for a given perturbation, then the perturbation function, noted as T_{Λ_M}, is parameterized by $\Lambda = [\lambda_1, \cdots, \lambda_K]$ (where K is the number of perturbations considered).

Optimal perturbations for the semi-supervised contrastive problem are then obtained via M, which is thus responsible for finding the optimal Λ_M^*. As already done by Chaitanya et al. [2020] and in contrast to Chen et al. [2020a], we keep one version of the image batch unperturbed.

Formally, we are solving the following coupled optimization problem:

$$\begin{cases} \min_M & \alpha_0 I_{NCE}\Big(g \circ f(X_M), g \circ f(X)\Big) + \alpha_1 \mathcal{L}\Big(p \circ f(X_M), y\Big) \\ \min_{f,p,g} & -\alpha_2 I_{NCE}\Big(g \circ f(X_M), g \circ f(X)\Big) + \alpha_3 \mathcal{L}\Big(p \circ f(X_M), y\Big) \\ & +\alpha_4 \mathcal{L}\Big(p \circ f(X), y\Big) \end{cases} \tag{3.2}$$

where α_i are weights balancing each loss term, y is the classification label when available, and $\mathcal{L}$ is the binary cross entropy loss function for the supervised constraint. I_{NCE} is the mutual information lower bound introduced by van den Oord et al. [2018] which is the opposite of the contrastive loss function [Chen et al., 2020a]:

$$I_{NCE}\Big(g \circ f(X_M), g \circ f(X)\Big) = \sum_{i=1}^{N} \log\left(\frac{e^{sim(g \circ f(X_{Mi}), g \circ f(X_i))}}{\sum_{j, j \neq i} e^{sim(g \circ f(X_{Mi}), g \circ f(X_j)))}}\right) \tag{3.3}$$

where the index i defines positive pairs (X_{Mi}, X_i), j negative ones $(X_i, X_j), (X_{Mi}, X_j)$, and sim is a similarity measure defined as $sim(x, x') = \frac{x^T x'}{\tau}$ where τ is a fixed scalar, here equal to 1. In comparison to Equation (3.1), the perturbed image version X_i^1 now depends on the perturbation generator M (X_{Mi}), and the second view X_i^2 is kept unperturbed (X_i in Equation (3.3) above).

The perturbation network, M, and the encoder, f, are optimized alternatively.

Perturbation network optimization steps:

i M generates a batch of Λ_M vectors defining a perturbation T_{Λ_M}. For every image X in a batch, a transformed version is generated: $X_M = T_{\Lambda_M}(X)$.

ii The transformed and untransformed data batches are passed through the encoder f, the projection head g, and the linear classifier p.

iii The gradients of both the contrastive $-I_{NCE}$ and classification $\mathcal{L}$ loss terms (see Equation (3.3)) are computed to update the weights of the network M, aiming to minimize the mutual information and the classification loss function.

Encoder optimization steps:

i From the previous optimization steps of M, one transformed version of the data is generated. Latent projections of the transformed and untransformed data are generated using encoder f and projection head g.

ii The contrastive and classification loss gradients are computed, and parameters of f, g and p are updated. This brings closer positive pairs and further away negative ones and ensures that transformed images are properly classified.

Differentiable formulation of perturbations

A fundamental difference of the proposed perturbation optimization, compared to Li et al. [2020b], Liu et al. [2021b], Tian et al. [2020], is the use of explicit perturbation differentiation.

During training, gradient computations of Equation 3.2 involve the derivative of T_{Λ_M} with respect to the weights (w) of M: $d_w(T_{\Lambda_M}) = dT_{\Lambda_M} \circ d_w M$. This requires the explicit computation of the derivatives of T with respect to its parameters Λ and the differential calculus for each perturbation composing T. Thus, we introduce specific formulations and normalized parameterization for the perturbations used in our experiments.

We use the following perturbations: crop ($Crop$), Gaussian blur (G), additive Gaussian noise (N), rotation (R) around the center of the image, horizontal ($Flip_0$) and vertical ($Flip_1$) flips. Table 3.1 lists the differentiable expressions of these perturbations where S is the sigmoid function, s the size of our images, erfinv the inverse of the error function $2\pi^{-\frac{1}{2}} \int_x^\infty e^{-u^2} du$, $\mathcal{U}$ the uniform distribution and x is a point of the image grid. We fix the maximum Gaussian blur standard deviation to $\sigma_{max} = 2.0$ and the maximum additive noise standard deviation to $\tilde{\sigma}_{max} = 0.1$. The final perturbation function is defined as:

$$T_\Lambda = (R \circ Flip_1 \circ Flip_0 \circ Crop \circ N \circ G)(X, \Lambda) \tag{3.4}$$

and T_Λ thus depends on 7 parameters (the crop has 2 parameters: the position of its center in each axis) which are generated by M.

Table 3.1: Differentiable expressions of the perturbations used, parameterized by $\lambda \in [0, 1]$.

Flip around axis e	$Flip(X, \lambda, e)(x) = (1-\lambda)X(x) + \lambda X(x - 2\langle x, e\rangle e)$
Crop centered at $c_\lambda = [\lambda_1 s, \lambda_2 s]$	$Crop(X, \lambda)(x) = X(x) \times S(\frac{s}{8} - \|\|x - c_\lambda\|\|_\infty)$
Gaussian blur with kernel $g_{\lambda\sigma_{max}}$	$G(X, \lambda) = g_{\lambda\sigma_{max}} * X$
Rotation	$R(X, \lambda)(x) = X\left(\begin{pmatrix} \cos(\lambda 2\pi) & -\sin(\lambda 2\pi) \\ \sin(\lambda 2\pi) & \cos(\lambda 2\pi) \end{pmatrix} x\right)$
Additive Gaussian noise	$N(X, \lambda) = X + \lambda\tilde{\sigma}_{max} \times \sqrt{2}\,\text{erfinv}(\mathcal{U}[-1, 1])$

3.3.2 Experiments

Datasets

Experiments are performed on BraTs MRI [Menze et al., 2015] and Chest X-ray [Tang et al., 2020] datasets. The Chest X-ray dataset is composed of 10000 images. BraTs volumes are split along the axial axis to get 2D slices. Only slices with less than 80% of black pixels are kept. This results in 34000 slices. For both datasets, we study the supervised task of pathology presence classification (binary classification, present/not present). In medical imaging problems, it is common to have labels only for a small part of the dataset. We thus chose 10% of supervision in all of our experiments. We randomly select three hold-out test sets of 1000 slices for BraTs experiments. With the Chest dataset, we use the provided test set of 1300 images, from Tang et al. [2020], evenly split in three to evaluate variability.

Experimental settings

For every experiment with the BraTs dataset, the encoder f is a fully convolutional network composed of four convolution blocks with two convolutional layers in each block. Following the architecture proposed by Tang et al. [2020], the encoder f for experiments on the Chest dataset is a Densenet121. The network M is a fully convolutional network composed of two convolutional blocks with one convolutional layer. The projection head g is a two-layer perceptron as in [Chen et al., 2020a].

On BraTs dataset (resp. Chest dataset), we train with a batch size of 32 (resp. 16) for 100 epochs. In each experiment, the learning rate of f is set to 10^{-4}. When optimizing M with (resp. without) supervision, M learning rate is set to 10^{-3} (resp. 10^{-4}).

When using 10% of labeled data for the supervision task, on relatively small databases (10^5 images), there is a risk of overfitting on the classification layer (p in Equation 3.2). Contrastive and supervision loss terms need to be carefully balanced while optimizing both the encoder and the perturbation generator. To evaluate the impact of hyper-parameters, we carry out experiments with $(\alpha_0, \alpha_2) \in \{1, 0.1\}$ and $(\alpha_1, \alpha_3, \alpha_4) \in \{1, 10\}$. Linear evaluation results (see following subsection) on BraTs dataset after convergence are summarized in Table 3.2. Results in Section 3.3.3 are shown with the best values found for each method.

Table 3.2: 3-fold cross-validation mean linear evaluation AUC (computed on thresholds applied to network prediction probability output) after convergence with different α_i values (standard deviation in parentheses).

	α_i values	AUC
Optimizing M	$\alpha_{0,2} = 1, \alpha_{3,4} = 1, \alpha_1 = 10$	**0.884** (0.042)
	$\alpha_0 = 0.1, \alpha_{1,3,4} = 10, \alpha_2 = 0.1$	0.868 (0.030)
	$\alpha_0 = 0.1, \alpha_1 = 10, \alpha_2 = 1, \alpha_{3,4} = 1$	**0.887** (0.013)
Random M	$\alpha_2 = 1, \alpha_{3,4} = 1$	0.874 (0.000)
	$\alpha_2 = 0.1, \alpha_{3,4} = 10$	0.820 (0.037)
	$\alpha_2 = 1, \alpha_{3,4} = 10$	**0.883** (0.003)
base simCLR Chen et al. [2020a]		0.730 (0.020)

The fully supervised experiments described in Section 3.3.3 are optimized with the same encoder architecture and one dense layer followed by a sigmoid activation function for the classification task. For the fully supervised experiments, we use a learning rate of 10^{-4}.

Linear evaluation

To evaluate the representation quality learned by the encoder, we follow the linear evaluation protocol used in the literature [Chen et al., 2020a, Perakis et al., 2021, Tian et al., 2020]. The encoder is frozen with the weights learned with our framework. One linear layer is added, after removing the projection head (g) and trained with a test set of labeled data, not used in the previous training phase. This means that we first project the test samples in the latent space of the frozen model and then estimate the most discriminative linear model. The rationale here is that a good representation should make the classes of test data linearly separable. This hypothesis implies that the instance discrimination task performed by contrastive pre-training can be a surrogate

of the classification task. Wang et al. [2022b] showed that proper perturbations would ensure such a separation between latent representations of different classes.

3.3.3 Results

To assess the impact of each term in Equation 3.2 we perform optimization using the following strategies:

- **Random** (without M, without supervision): each image is transformed with parameters generated by a uniform distribution: $\Lambda = \mathcal{U}\left([0,1]^7\right)$, and $\alpha_{1,3,4} = 0$.
- **Random with supervision** (without M, with supervision): we add the supervision constraint to the random strategy. We set $\alpha_2 = 1$ and $\alpha_{3,4} = 10$.
- **Self-supervised** (with M, without supervision): while setting α_1, α_3 and α_4 to 0, we optimize Equation 3.2.
- **Self-supervised with supervision constraint** (with M and supervision): setting $\alpha_1 = 10$ and $\alpha_{0,2,3,4} = 1$, we optimize Equation 3.2.

We split the data into pre-training and test sets. Data from the pre-training set are further split into training and validation sets for the perturbator/encoder optimization. For optimizations with supervision constraint (self-supervised and random), all pre-training data is used for self-supervision, and a small set of labeled data is used for the supervision constraint. For variability analysis, three optimizations are performed by changing the supervision set. A schematic view of our splitting strategy is shown in Figure 3.4.

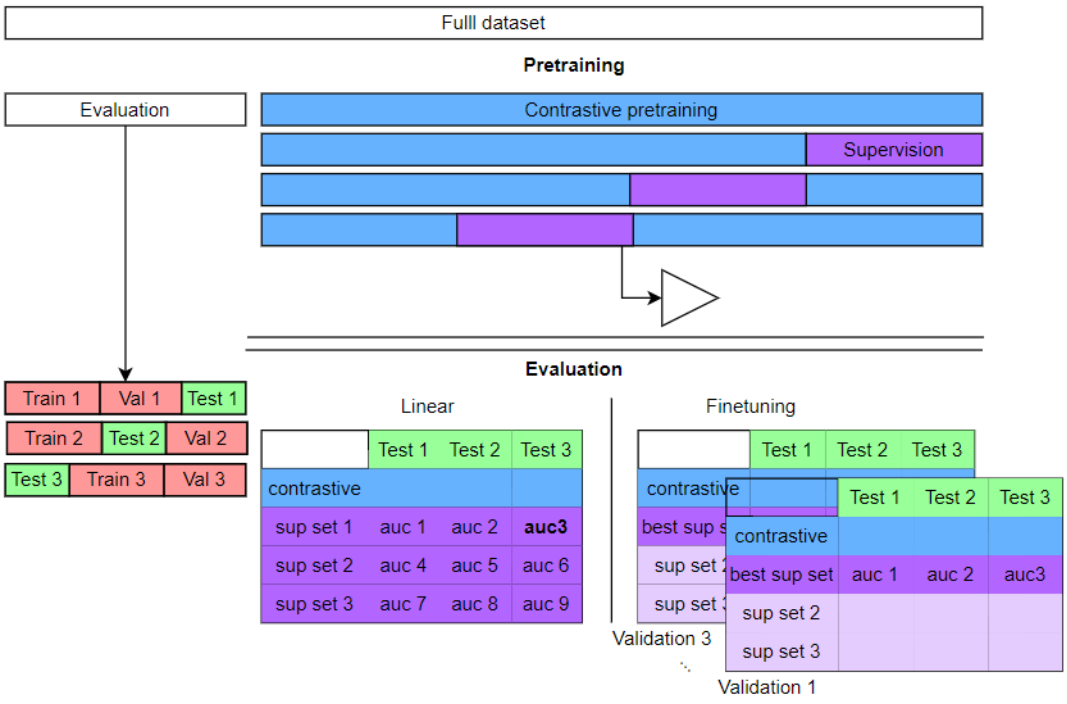


Figure 3.4: Splitting strategy for M/f optimization.

With the BraTs dataset, as slices come from 3D volumes, we split the data ensuring that all slices of the same patient are in the same set.

Linear evaluation is performed on the four optimization strategies with the hold-out test set. Performances are evaluated with the weights obtained at different epochs. We aim to evaluate if our method outputs better representations during training.

In Figure 3.5, we show performances (mean and standard deviation) on three different test sets for both datasets. We also train the encoder on the classification task in a fully supervised setting with 10% and 100% labeled data. For the fully supervised training, we use data augmentation composing the tested perturbations randomly. Each perturbation has a 0.5 probability of being sampled.

We perform linear evaluation on the frozen encoder with the hold-out test set and report the obtained AUC as horizontal lines in Figure 3.5. Figure 3.5 also reports linear evaluation results of the base simCLR optimization by Chen et al. [2020a] where only one image is transformed by a random composition of the tested perturbations. As with the fully supervised experiments, each perturbation has a 0.5 probability of being sampled.

Figure 3.5 shows that optimizing M with supervision helps to have better representations for both datasets. It also shows that optimizing with only 10% of labeled data allows to reach the same quality of representation as the fully supervised training with 100% of labels.

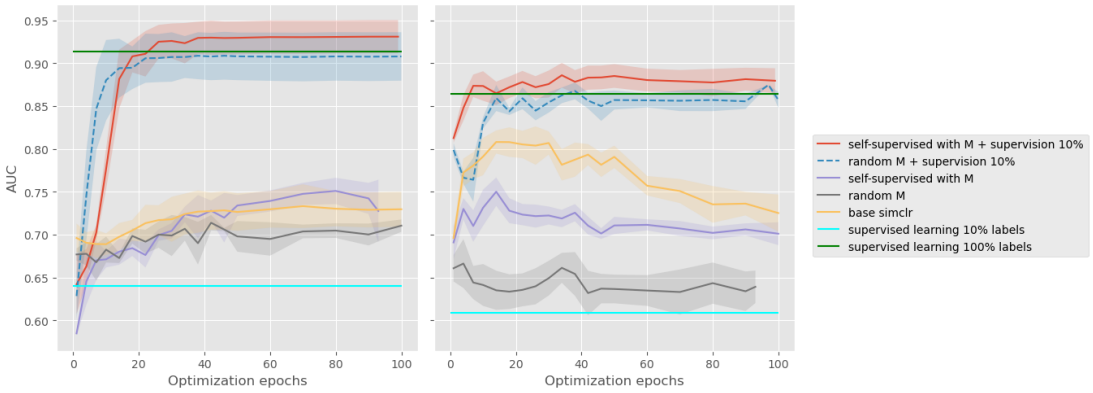


Figure 3.5: Linear evaluation results comparing with other methods (left BraTs dataset, right Chest dataset, the y-axis is shared by both plots).

To investigate the impact of the supervised loss function, we launched an experiment with the supervised contrastive loss function introduced by Khosla et al. [2020b] using only 10% of labeled data. After convergence, we obtain a mean AUC of 0.52 ± 0.12 compared to 0.93 ± 0.01 with our method.

On the Chest X-ray database, strong results were obtained by Tang et al. [2020] using a network pre-trained on ImageNet. Optimizing M with 10% supervision on this ImageNet pre-trained network has a smaller impact compared to random perturbations (0.96 ± 0.001 for both approaches). However, ImageNet pre-trained networks can only be used with 2D slices or 2.5D volumes (composed of 3 consecutive slices, one in each color channel) whereas our strategy could be easily extended to 3D volumes.

Quality of learned representation

To evaluate the quality of the learned representation we perform t-distributed Stochastic Neighbor Embedding (tSNE [van der Maaten and Hinton, 2008]) dimensionality reduction of the latent representations of the test set data. The test set data for both datasets are fed to the trained encoders, and the latent representations before the projection layer are used to fit a 2D tSNE projection. Each point in Figures 3.6 and 3.7 represent the latent projections on a 2D plane of a test set data.

For the BraTs dataset, Figure 3.6 shows that normal and abnormal cases are better clustered when optimizing the perturbation generator with 10% supervision than when performing base contrastive learning with simCLR [Chen et al., 2020a]. A good separation is obtained with fully supervised learning with 100% annotated data.

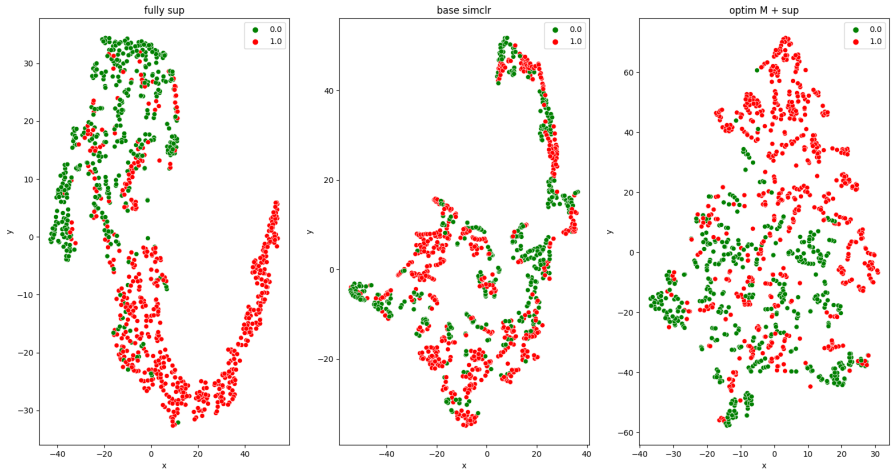


Figure 3.6: tSNE of learned representation for different experiments on BraTs dataset: fully supervised with 100% annotated data (left), base simCLR (middle), optimizing M with 10% supervision (right)

On the Chest dataset, Figure 3.7 shows that optimizing the perturbation generator with 10% supervision gives a better class separation than training with fully supervised learning with 100% annotated data or base simCLR.

Relevance

When optimizing without supervision, the network M needs to minimize the mutual information and it can therefore generate perturbations that create images that are very different from the untransformed images but that do not contain relevant information for the downstream task, in particular for medical images. Without the supervision constraint, the optimal crop can be found, for instance, in a corner, leading to an image with a majority of zero values (i.e., entirely black), thus useless for the supervised task. The supervision constraint helps M to generate relevant images that keep pathological pixels (see some examples in Figures 3.8 and 3.9).

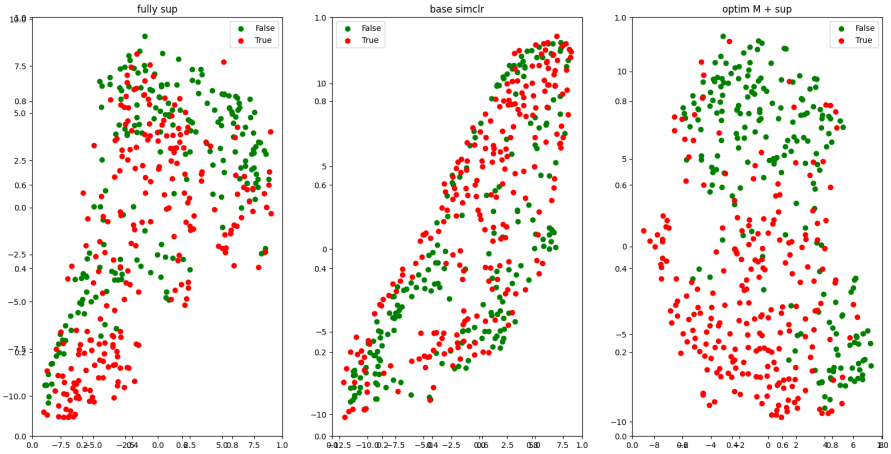


Figure 3.7: tSNE of learned representation for different experiments on Chest dataset: fully supervised with 100% annotated data (left), base simCLR (middle), optimizing M with 10% supervision (right)

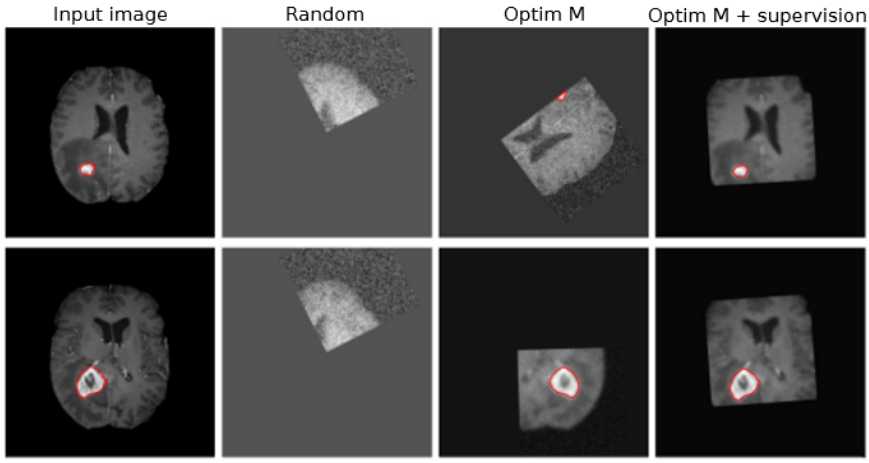


Figure 3.8: Two examples (rows 1 and 2) of generated perturbations on the BraTs dataset with different optimization strategies (the red contour highlights the tumor).

Runtime

The addition of the network M increases the training computational time by around 20-25% which is balanced by a performance gain.

Perturbation composition order

As done in Chen et al. [2020a], the perturbation order is fixed. We launched an additional experiment with a different perturbation order for both simCLR and our method. Linear evaluation results after convergence are shown in Table 3.3.

We see that the perturbation order has little impact on our results and, above all, our method substantially outperforms simCLR in both experiments.

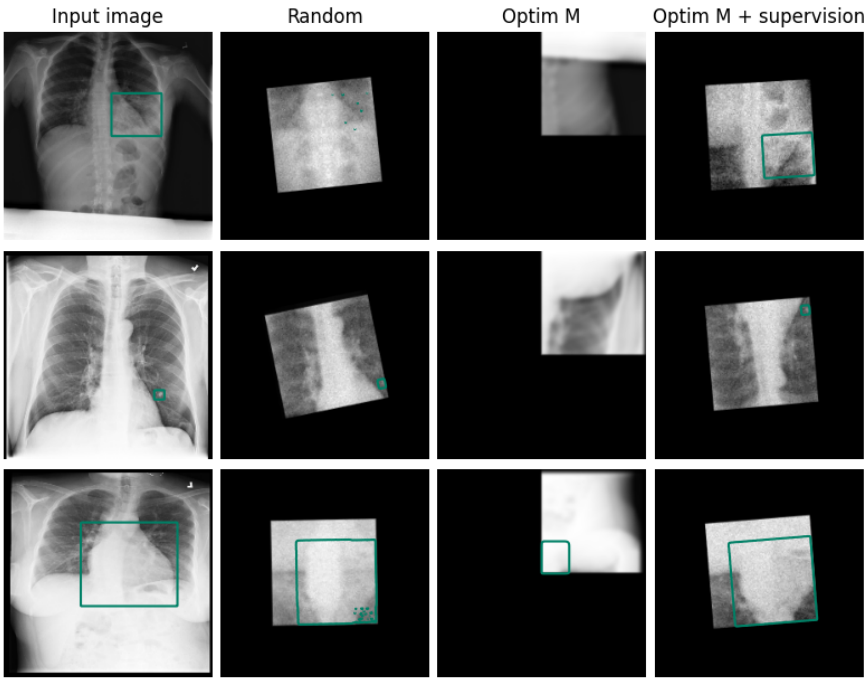


Figure 3.9: Examples (rows 1, 2, and 3) of generated perturbations on the Chest dataset with different optimization strategies (the green contour highlights the pathology localization).

Method	First composition order	Second composition order
simCLR	0.730±0.020	0.760±0.027
Ours	0.926±0.020	0.923±0.021

Table 3.3: Linear evaluation AUC after convergence when changing perturbation composition order.

3.3.4 Improving the perturbation generator

Mode collapse

Given the promising results obtained with linear evaluation using the pre-trained perturbation generator, we could take advantage of the found perturbations using them as augmentations in fine-tuning.

We thus use the perturbation generated by M to fine-tune neural networks pre-trained with our approach and simCLR method. We also use these augmentations to train a neural network from scratch on the supervised classification task. Table 3.4 shows that using the perturbations learned by M during fine-tuning is detrimental to performances.

Looking at the covariance matrix of the generated perturbation parameters (see Table 3.5), we see that M has a mode collapse. There is not much variability in the perturbations generated by M which explains why this constant augmentation is not useful for fine-tuning.

To increase the variability of the augmentation generated by M during fine-tuning, we add some noise to M outputs. We also use the pre-trained classification layer during fine-tuning. Using the pre-trained classification layer and adding noise on M we get: AUC $= 0.83 \pm 0.03$. Comparing with results in Table 3.4, we get better results than when not using the learned classifier but we still get worse results than with random augmentation. This advocates for a study of the modes of M to avoid generating only one constant augmentation.

Table 3.4: Fine-tuning results on BraTs dataset with different augmentation strategies.

	Random	None	M-produced
Ours	0.84 (0.03)	0.82 (0.04)	0.81 (0.05)
simCLR	0.68 (0.03)	0.73 (0.01)	0.67 (0.01)
Random init	0.60 (0.008)	0.55 (0.013)	0.60 (0.002)

Table 3.5: Covariance matrix of perturbation parameter values obtained on BraTs dataset with the pre-trained perturbation generator.

	λ_0	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
λ_0	0.0014	-0.0011	-0.0004	4.70e-05	-2.84e-05	2.97e-05	0.0003
λ_1	-0.0011	0.0009	0.0003	-3.75e-05	2.24e-05	-2.35e-05	-0.0002
λ_2	-0.0004	0.0003	0.0001	-1.43e-05	8.30e-06	-8.69e-06	-7.93e-05
λ_3	4.70e-05	-3.75e-05	-1.43e-05	1.63e-06	-9.42e-07	9.86e-07	8.99e-06
λ_4	-2.84e-05	2.24e-05	8.30e-06	-9.42e-07	8.35e-07	-8.59e-07	-6.90e-06
λ_5	2.97e-05	-2.35e-05	-8.69e-06	9.86e-07	-8.59e-07	8.92e-07	7.15e-06
λ_6	0.0003	-0.0002	-7.93e-05	8.99e-06	-6.90e-06	7.15e-06	6.01e-05

Generating two perturbed views with M

To increase the variability of the perturbations generated by M and get closer to the work done by Chen et al. [2020a], we propose to train M to generate two perturbed views of the input image. We change the used flip by :

$$\text{Flip}(x, \lambda) = \sigma(\lambda)\text{Flip}(x) + (1 - \sigma(\lambda))x \tag{3.5}$$

to force λ values generated by M to be either 0 or 1 and avoid 0.5 that generated irrelevant images. Figures 3.10 and 3.11 show results obtained after two optimizations with different weights given to the supervision constraint. The obtained perturbations are satisfactory but do not contain enough tumor information even when increasing the supervision constraint.

Alignment and uniformity framework

Wang and Isola [2020] introduce another contrastive loss function formulation that performs better than that proposed by Chen et al. [2020a]. They rewrite the contrastive loss function through alignment and uniformity. Positive pairs similarity in the latent space defines the alignment term. They show that negative pairs dissimilarity is equivalent to a uniform distribution of latent representations. Their loss function writes

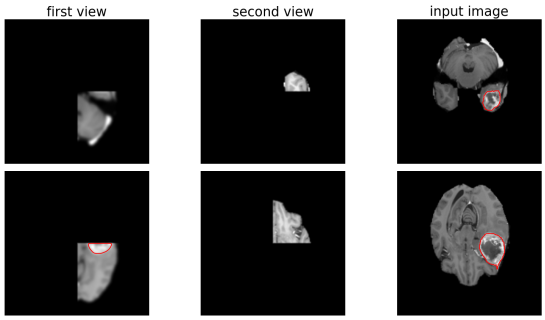


Figure 3.10: Example of generated views optimizing M with supervision giving equal importance weight to contrastive and supervision constraints.

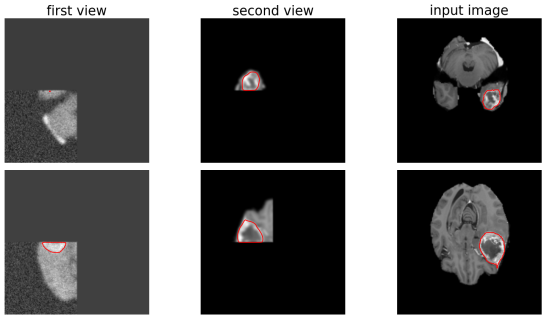


Figure 3.11: Example of generated views optimizing M with supervision giving a larger weight to supervision constraint.

:

$$\mathcal{L}_{NCE} = \underbrace{\frac{1}{N}\sum_{i=1}^{N} ||x_i^1 - x_i^2||_2}_{\text{Alignment: } \mathcal{L}_{align}} + \underbrace{\log\left(\frac{1}{N^2}\sum_{i,j=1}^{N} e^{-||x_i^1 - x_j^2||_2}\right)}_{\text{Uniformity: } \mathcal{L}_{uniform}} \tag{3.6}$$

We use this new formulation on the perturbation optimization method presented in Section 3.3.1. Maximizing both alignment and uniformity for the perturbation generator while minimizing them for the encoder shall give the same results as previous experiments with mutual information, asymptotically. The optimization problem introduced in Equation (3.2) now writes (without supervision constraint for better readability):

$$\begin{cases} \max_M & \mathcal{L}_{align}\Big(g \circ f(X_M), g \circ f(X)\Big) + \mathcal{L}_{uniform}\Big(g \circ f(X_M), g \circ f(X)\Big) \\ \min_{f,g} & \mathcal{L}_{align}\Big(g \circ f(X_M), g \circ f(X)\Big) + \mathcal{L}_{uniform}\Big(g \circ f(X_M), g \circ f(X)\Big) \end{cases} \tag{3.7}$$

However, this formulation implies that the perturbation generator will push negative samples to collapse to one point in the latent space which is counterintuitive. The uniformity term can thus be minimized for the perturbation generator:

$$\begin{cases} \max_M & \mathcal{L}_{align}\Big(g \circ f(X_M), g \circ f(X)\Big) + \min_M \mathcal{L}_{uniform}\Big(g \circ f(X_M), g \circ f(X)\Big) \\ \min_{f,g} & \mathcal{L}_{align}\Big(g \circ f(X_M), g \circ f(X)\Big) + \mathcal{L}_{uniform}\Big(g \circ f(X_M), g \circ f(X)\Big) \end{cases} \tag{3.8}$$

To evaluate these two optimization frameworks, we plot the pairwise distance between latent representations of samples within a random batch, after pre-training, to see if the uniformity term is properly minimized.

Figure 3.12 shows that maximizing uniformity during perturbator optimization leads latent representations to collapse to one point with zero pairwise distance (maxU curve). Minimizing uniformity for both the encoder and the perturbation generator leads to more relevant and useful pairwise distances (minU curve).

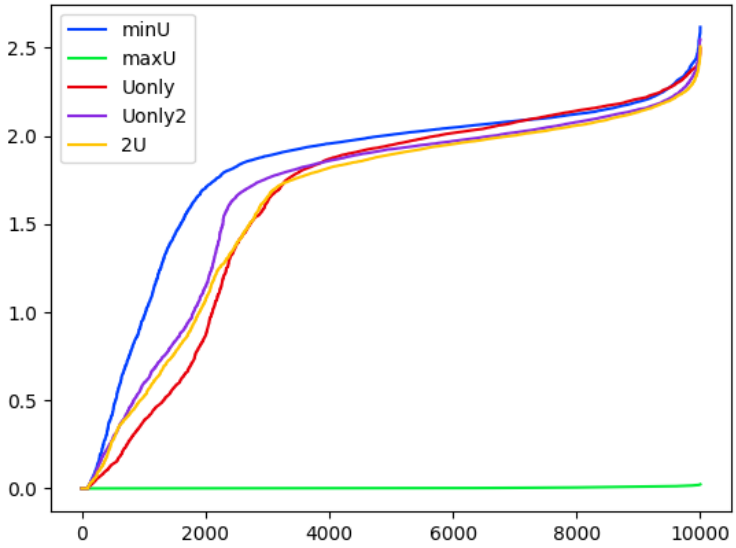


Figure 3.12: Pairwise distances between latent representations of images of a batch sampled at random after pre-training (the x-axis represents the images for which each pairwise distance, in the y-axis, has been calculated). The maxU curve shows that representations from optimization of Equation (3.7) collapse to one point, while the minU curve from optimization of Equation (3.8) shows spread pairwise distance values.

Figure 3.13 shows that minimizing uniformity generates more relevant images than maximizing it. Minimizing only uniformity generates too redundant images.

These preliminary results of our method application to the alignment and uniformity framework show that minimizing the mutual information might eventually not be the best approach for the perturbation generator as it goes against the uniformity constraint. Later works should focus on a thorough hyper-parameter analysis to find the right balance between alignment and uniformity.

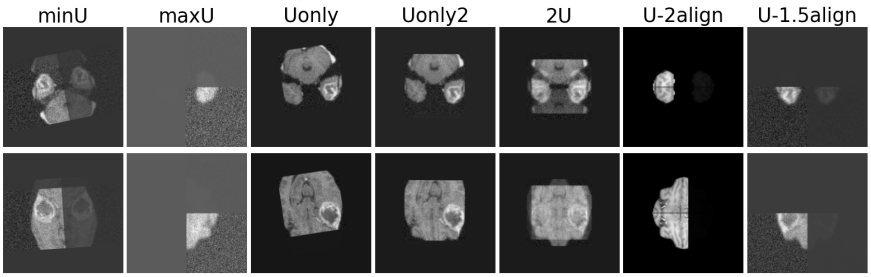


Figure 3.13: Perturbations generated after various optimizations of alignment and uniformity, from left to right: minimizing uniformity, maximizing uniformity, optimizing uniformity only, optimizing uniformity only with updated computation. Three last columns: varying alignment and uniformity weights.

3.4 Conclusion and perspectives

In this chapter, we have investigated how to optimize perturbations applied in contrastive learning pre-training. Some automatic augmentation methods have already been developed both in fully and self-supervised learning, but they are either based on generative networks (which allow for less control on the generated perturbations, thus reducing interpretability), based on perturbation networks applied in the latent space or inapplicable to medical images.

Building on the work of Tian et al. [2020] introducing a "sweet spot" for contrastive learning views, we introduce a perturbation network minimizing the mutual information between views without reducing supervised task performances. Our perturbation network outputs a set of parameters defining the composition of perturbations to apply. To build this network, each perturbation must be differentiable with respect to its parameters. We thus propose a differentiable formulation of the used perturbations. The perturbation generator is then trained to minimize the mutual information between views while the encoder learns to maximize it. A supervision constraint is used to ensure that the generated perturbations are not detrimental to the supervised target task. After pre-training, we obtained good linear evaluation results on two datasets (brain and chest images) and relevant perturbation outputs. Further results on prostate images are detailed in the next chapter.

As shown in Section 3.3.4, the perturbation generator could yet be improved: by increasing variability, generating more perturbed views, or changing the mutual information minimization approach. Methods to overcome these limitations are perspectives for further work and improvements. Another perspective could entail having multiple classification heads to optimize the perturbation generator on multiple supervised target tasks.

Chapter 4

Introducing metadata confidence in contrastive learning: application to prostate cancer lesion detection

Contents

In the previous chapter, we investigated how to optimize perturbations for contrastive learning pre-training through a generic method applicable to many supervised tasks. Some diagnosis tasks are subject to inter and intra-radiologist variability and it can be hard to obtain a consensus label. When applying deep learning to these tasks, for which a ground truth consensus is difficult to find, performances are highly impacted. In this chapter, we investigate how to improve deep learning algorithms performances for prostate cancer lesion detection on Magnetic Resonance Images (MRI) by combining annotation variability and conditional contrastive learning. The methods presented in Chapter 3 will be applied in this context.

Prostate cancer is the second most common cancer in men worldwide. Its early detection is crucial for efficient medical treatment. This detection is often done based on multi-parametric MRI which has proved successful in increasing diagnosis accuracy [Rouvière et al., 2019].

Recently, deep learning methods have been developed to automate prostate cancer detection [Bhattacharya et al., 2021, Saha et al., 2021, Yu et al., 2020]. Most of these methods rely on datasets of thousands of images, where lesions are usually manually annotated and classified by experts. However, lesion annotations at pixel level are costly to obtain and annotations at exam level, such as lesion severity, can be subject to annotator variability. Self-supervised learning approaches, in particular contrastive learning ones, could thus be applied to reduce the annotation cost. Furthermore, conditional contrastive learning approaches, such as those introduced by Dufumier et al. [2021b], could be improved to take annotator variability into account.

In this chapter, we first introduce the clinical context (Section 4.1). We then present existing deep learning methods for prostate cancer lesion detection (Section 4.2.1) and conditional contrastive learning approaches (Section 4.2.2). In Section 4.3, we introduce the private dataset used, the available annotations, and their variability. We expose our proposed conditional contrastive learning method including annotator confidence and the obtained results in Section 4.5. In Section 4.6, we present preliminary results on pseudo-label computation with nearest neighbors and contrastive learning on the decoder. Finally, we apply the method proposed in Chapter 3 to prostate cancer lesion classification (Section 4.7).

4.1 Clinical context

4.1.1 Anatomy

The prostate is a gland of the male reproductive system. It is located under the bladder ahead of the rectum. It is responsible for producing fluid for semen, closing up the urethra up to the bladder during ejaculation, and metabolizing hormones. It can be divided into multiple zones: the peripheral zone (PZ), the transitional zone around the urethra (TZ), the anterior fibro-muscular stroma (AFMS), and a central zone (CZ) around ejaculatory ducts. Figure 4.1 shows a schematic view of the prostate.

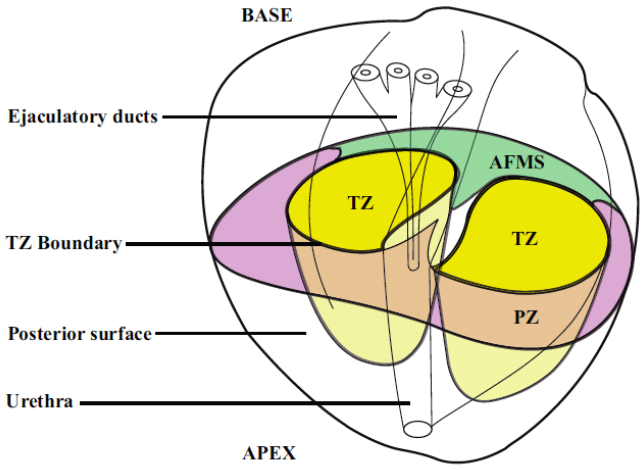


Figure 4.1: Schematic view of the prostate and its different zones (taken from Colin [2012]).

4.1.2 Pathology

Every year, in France, 56800 new cases of prostate cancer are identified. It is the most common cancer in men, ahead of lung cancer (28200 new cases per year) and colorectal cancer (23200 new cases per year) [Bray et al., 2018].

It is a cancer of varying aggressiveness with a slow development. A localized cancer will be asymptomatic, a locally advanced cancer will cause urinary disorders and more locally advanced cancers can lead to kidney failures. In most advanced cancer stages, the disease becomes metastatic leading to bone pain, loss of weight, and loss of general condition.

In France, among patients diagnosed, 9000 will die from prostate cancer each year which makes it a public health issue. French medical and urology institutions thus recommend annual screening for prostate cancer from the age of 50 [Rouvière et al., 2019]. This begins with a prostate-specific antigen (PSA) test and a digital rectal examination (DRE). If, during these examinations, the urologist suspects the presence of cancer, further tests will be carried out, including an MRI examination. The contribution of MRI to the early detection of prostate cancer is twofold: 1. compared with clinical-biological examinations alone, it enables better classification of patients who will need a biopsy; and 2. by guiding biopsies, it greatly improves the detection of significant cancers while reducing the detection of non-significant cancers, thereby avoiding over-diagnosis [Hugosson et al., 2022].

During the examination, practitioners will seek to identify tumors present in a patient prostate by analyzing multi-parametric MRI (mpMRI) combining a T2-weighted sequence (T2w), a diffusion-weighted image (DWI), apparent diffusion coefficient (ADC) mapping and a dynamic contrast-enhanced (DCE) sequence (a detailed explanation of these sequences will be given in Section 4.3).

The performance and interpretation of prostate MRI have been standardized by the Prostate Imaging Reporting and Data System (PI-RADS) committee [Turkbey et al., 2019, Weinreb et al., 2016]. Each tumor is assigned a PI-RADS score, reflecting the

probability of the lesion being clinically significant. This score ranges from 1 (very low probability) to 5 (very high probability). Identification of a lesion for which the radiologist has assigned a score greater than 4 will lead to the planning of a biopsy. Anatomopathological examination is essential to identify all the characteristics of the tumor (exact nature, composition, degree of aggressiveness).

Doctors will be able to indicate the malignant or benign nature of the tumor and its grade only after this examination has been performed. It is thus quite common for some practitioners to have different opinions when it comes to assigning PI-RADS scores for a given case. Kasivisvanathan et al. [2018] showed that interpretation efficiency largely depends on radiologists' experience. Although the latest versions of PI-RADS recommendations have reduced inter-annotator variability, it remains relatively high ($\kappa = 0.57$ according to Turkbey et al. [2019]).

Diagnostic support from prostate MRI for cancer detection remains a major challenge to make diagnoses more reliable. Due to the growing prevalence of prostate cancer and the latest PI-RADS recommendations, the number of prostate MRIs has increased, which led to the development of deep learning algorithms for computer-aided diagnosis.

4.2 Related works

4.2.1 Deep learning for prostate cancer lesion detection

Existing methods for prostate cancer detection with deep learning can be split into two families: classification approaches and combined approaches that aim to classify and detect lesions in MRI.

Classification approaches

In this section, we give two examples of classification approaches without concern for completeness.

Reda et al. [2019] build seven different convolutional neural networks evaluating seven diffusion-weighted sequences with different b-values: diffusion sequences rely on water molecules motion within tissues, this motion is random and positively correlated with the signal loss of the sequence of which b is a variable [Huisman, 2003]. Each network outputs two probabilities: one for benign and the other for malignant prostate. The seven probabilities are then concatenated and fed to a support vector machine, which makes the final decision.

Yoo et al. [2019] combine neural networks and machine learning algorithms. Their solution first seeks to detect cancerous lesions on axial slices of the input volume. Slices are fed to five identical fully convolutional neural networks working in parallel. For each patient, probabilities are generated for each slice by each model. Among this set, the top five are kept. Five sets of first-order statistical features are then built and fed to a decision tree algorithm for feature selection. The selected features are then given to a random forest classifier outputting the final prostate classification among normal or abnormal.

The main limitations of classification approaches are that they do not provide interpretable supports for radiologists, such as lesion localization or pixel-wise segmentation, and are thus of less clinical value.

Approaches combining classification with detection or segmentation methods often output lesion segmentation masks which are of better use for radiologists.

Combined methods: classification and detection

In this section, we cite some methods combining classification and detection approaches for automatic prostate cancer detection on MRI, but without being exhaustive. Results by Schelb et al. [2019] paved the way for the use of neural networks for prostate cancer lesion detection. They showed that a U-Net trained with T2w and diffusion MRI to generate prostate cancer probability maps achieved similar performances to clinical PI-RADS assessment. Since then, many works have been proposed which can be split based on the network architecture used.

Feature pyramid networks

Feature pyramid networks [Lin et al., 2017] (FPN) have been developed to detect objects at different scales with optimal memory cost. An encoder is used as backbone. Feature maps from each encoder level are extracted and then combined, through skip connections (concatenation of encoder feature maps to corresponding decoder ones), with up-sampled feature maps from lower levels. Generated feature maps are then fed to a detector such as a region proposal network [Ren et al., 2015] for object detection. Feature pyramid networks can also be used in segmentation models for mask predictions.

Yu et al. [2020] use a ResNet50 [He et al., 2016] as backbone and add an instance and a semantic segmentation branch to the FPN. Features from different levels, generated with the FPN, are used as input to a semantic segmentation branch: feature maps at different resolutions are up-sampled, normalized, and fed to a Squeeze-And-Excitation block [Hu et al., 2018] before generating the final semantic segmentation. The Squeeze-And-Excitation block models dependencies between the channels of the feature maps, it starts by squeezing spatial information at the channel level through global average pooling. The generated descriptor is then fed to two fully connected layers and a nonlinearity. The instance segmentation branch is an extension of the Mask R-CNN [He et al., 2017]: features generated by the ResNet backbone are fed through an attention module and are then used to generate region proposals. The attention module is defined to integrate global and local features from both instance and semantic segmentation branches. The model is optimized to minimize a combination of a segmentation loss function on the semantic segmentation output and a classification loss function on the detection outputs of the region proposal network.

U-Net

Many works have used the famous and successful U-Net architecture [Ronneberger et al., 2015] for prostate cancer lesions segmentation. Proposals range from very simple adaptations of the U-Net architecture to more complex ones.

Schelb et al. [2019] train a base 2D U-Net and compare its performances to PI-RADS clinical assessment. They build on their work by simulating the clinical deployment of an automated prostate cancer detection network [Schelb et al., 2020] using an ensemble of 16 2D U-Nets.

Saha et al. [2021] introduce a prior defining cancer spatial prevalence into 3D U-Net training: a lesion spatial prevalence heatmap is computed and added as a supplementary channel for network input. They use an adaptation of the 3D U-Net++ [Zhou et al., 2018]: residual blocks are used in the encoder before activation layers, and decoder dense and deep supervision blocks are dropped to have a lighter architecture. They introduce an attention mechanism through channel-wise Squeeze-and-Excitation [Hu et al., 2018] in residual blocks and grid-attention gates [Schlemper et al., 2018] before skip-connection layers. Grid-attention gates learn coefficient weighting the pixels of feature maps according to their relevance to the task. A patch-wise lesion residual classifier is trained in parallel. Detection and classification outputs are fused into a decision function to aggregate predictions and reduce false positives: patches classification is used to identify false positives in the detection map and to suppress them.

These approaches with the U-Net architecture do not necessarily combine classification and detection approaches, but the model used generates lesion segmentation masks that are readily interpretable by clinicians.

Edge detection networks

The third family of deep learning methods for prostate lesion segmentation is based on edge detectors such as holistically-nested edge (HED) ones [Xie and Tu, 2015]. In HED, side-output layers are introduced after each convolution block, and supervision is imposed at each side-output level. A weight fusion layer is added to combine outputs at different scales. This method has been applied straightforwardly to prostate cancer lesion detection by Sumathipala et al. [2018].

Bhattacharya et al. [2021] build on this architecture introducing a "Correlated Signature Network" to classify and locate indolent from aggressive prostate cancer. T2w, ADC, diffusion, and histopathology images are fed to a pre-trained VGG-16 network [Simonyan and Zisserman, 2014] for feature extraction. These features are concatenated and used to train a correlation network to learn a shared representation from MRI and histopathology. This representation is then given to a modified HED network: T2w, ADC sequences, and feature maps are fed to three parallel branches and fused later to generate segmentation maps.

Other examples of combined methods for prostate cancer lesion detection and segmentation have been reported in various reviews [Bhattacharya et al., 2022, Rouvière et al., 2022, Sushentsev et al., 2022].

The performances reported by these different works open up the prospect of their use in routine clinical practice. For these studies, databases of several thousand examinations have been enriched with dense annotations (prostate segmentation and lesion segmentation), enabling the development of models for lesion detection and segmentation using deep neural networks. However, Hosseinzadeh et al. [2021] show that larger cohorts are still needed to develop tools whose performances are compatible with clinical routine.

Building large annotated datasets with lesion annotations at pixel level is quite costly, which advocates for the development of self-supervised methods to take advantage of unannotated data.

Contrastive learning for prostate cancer lesion detection

As introduced in Section 2.2.2, contrastive learning methods have been introduced as part of self-supervised learning to train a model to learn similarity, in the latent space, between images. These methods have been applied to prostate lesion detection problems.

Fernandez-Quilez et al. [2022] apply the base simCLR [Chen et al., 2020a] (explained in details in Section 3.2) approach on 2D T2w axial MRI. They report better fine-tuning results than random and ImageNet [Deng et al., 2009] initializations with different fractions of labeled data. This approach does not use multi-parametric MRI although it has been shown to increase diagnostic relevance.

Gutiérrez et al. [2022] build sequence-specific parallel encoders and apply the contrastive learning framework to prostate lesions regions of interest. The neural network is then fine-tuned to classify lesion clinical significance. Having three encoders (one for each sequence among T2w, ADC, and diffusion) is computationally heavy and would not apply to whole 3D volumes. This method also needs to know the lesion region of interest beforehand.

4.2.2 Conditional contrastive learning

To improve latent representation quality, some works have proposed to condition contrastive learning with class labels or weak metadata.

Khosla et al. [2020a] propose to condition pair sampling with class labels, when available. Rather than defining positive pairs with two perturbed views of the same image, positive pairs are defined as all the images sharing the same label. This leads to a new loss function:

$$\mathcal{L}_{sup} = \sum_{i \in I} \mathcal{L}_i = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} -\log \left(\frac{e^{sim(x_i, x_p)}}{\sum_{a \in A(i)} e^{sim(x_i, x_a)}} \right) \tag{4.1}$$

where $A(i) = I \setminus i$, I being the set of perturbed images, $P(i) = \{p \in A(i) : y_p = y_i\}$ with y the label of the image.

This method performs well in a fully-supervised setting where class labels are available but is not applicable if class labels are missing. Some works have proposed to use weak metadata to condition contrastive learning.

Tsai et al. [2022] and Dufumier et al. [2021a] introduce a kernel function on metadata y to *condition* positive and negative pairs selection. Using the alignment and uniformity formulation introduced by Wang and Isola [2020] and presented in Section 3.3.4, they introduce the following loss function:

$$\mathcal{L}_w = \underbrace{\frac{1}{N} \sum_{i,j=1}^{N} w(y_i, y_j) d_{ij}}_{\text{Conditional Alignment}} + \underbrace{\log \big(\frac{1}{N^2} \sum_{i,j=1}^{N} (||w||_\infty - w(y_i, y_j)) e^{-d_{ij}} \big)}_{\text{Conditional Uniformity}} \tag{4.2}$$

where $d_{ij} = ||x_i^1 - x_j^2||_2$ (superscripts 1 and 2 indicate that latent representations x_i and x_j come from two perturbed versions of input images i and j), w is a kernel function measuring the degree of similarity between metadata y_i and y_j, $0 \leq w \leq 1$ and $||w||_\infty = w(y_i, y_i) = 1$.

The conditional alignment term brings close together, in the representation space, only samples that have a metadata similarity greater than 0, while the conditional uniformity term does not repel all samples uniformly but weights the repulsion based on metadata dissimilarity. A schematic view of this function is shown in Figure 4.2.

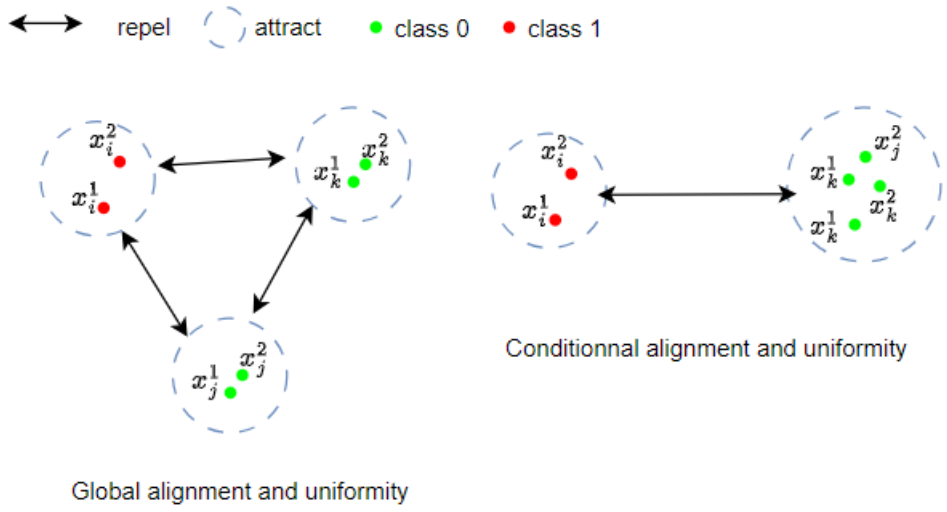


Figure 4.2: Schematic view of loss functions in Equations (3.1) (left) and (4.2) (right).

Class labels and weak metadata can be noisy. As introduced in Section 2.2.2, some works have used noisy labels in contrastive learning approaches and introduced confidence in labels in contrastive loss functions. The existing approaches are usually based on a measure of label noise through loss function values but a priori knowledge on label noise can be available. As introduced in Section 4.1.2, PI-RADS scores can be subject to high inter-annotator variability which can be measured and used within contrastive learning loss function.

With the increasing amount of available prostate MRI, many deep learning methods have been proposed to automate prostate cancer detection and lesion segmentation. The proposed fully-supervised approaches require large annotated datasets with manual lesion annotations at pixel level which are costly to obtain. This advocates for the use of self-supervised learning methods, but only few have been proposed for prostate cancer detection [Gutiérrez et al., 2022, Fernandez-Quilez et al., 2022]. Conditional contrastive learning methods have been proposed to take metadata into account in pre-training but, to the best of our knowledge, none of them use a priori knowledge on metadata variability. We thus propose a conditional contrastive learning method, taking PI-RADS scores variability into account, to pre-train a prostate cancer lesion segmentation model.

4.3 Dataset

4.3.1 MRI sequences

The private dataset that we use contains 2415 multi-parametric MRIs acquired on 1.5T and 3T MRI machines from different manufacturers (66% GE Medical Systems, 28% Siemens, and 5.2% Siemens Healthineers). All MRIs were paired with radiologically-estimated annotations of clinically significant prostate cancer via PI-RADS v2.1. MRIs were interpreted by at least one of twelve radiologists with varying experience. The whole prostate gland was manually segmented. Found lesions were assigned a PI-RADS score and lesions with PI-RADS ≥ 3 were manually segmented. Interpretation was performed with the use of axial T2w, axial diffusion-weighted imaging (DWI) with calculated apparent diffusion coefficient (ADC) maps and high-b-value DWI images (b $\geq$ 1200s/mm2) and dynamic contrast-enhanced imaging (DCE).

T2w sequences give a high-resolution view of the prostate anatomy and are thus called morphological sequences.

DWI sequences show water molecules' motion within tissues. The intensity of each pixel reflects water diffusion at that pixel location: on a DWI sequence a high signal value reflects a high cell density because of less water molecules diffusion. These sequences are acquired with different values of the diffusion variable b depending only on acquisition parameters.

ADC sequences are obtained by taking the difference between two DWI sequences with different b values.

DCE sequences show changes in tissue over time after administration of a contrast agent. In a recent literature review, Castillo et al. [2020] found no performance difference when using the DCE sequence in machine learning algorithms.

We thus use T2w, DWI, and ADC sequences to build our training dataset. Figure 4.3 shows the different sequences used and the bizonal (peripheral and transitional zones) prostate segmentation. Bizonal prostate segmentations are generated by a deep learning algorithm, developed by Incepto, trained on T2w sequences and available prostate manual segmentations.

For each interpreted examination we then have a bi-parametric MRI composed of three MRI sequences, the prostate gland manual segmentation, lesions grading and segmentation when available, and a global examination grading.

4.3.2 Manual lesion segmentation

As shown by Turkbey et al. [2019], there is inter-reader variability when grading a lesion even with PI-RADS v2.1 standardization. Manual lesion segmentation can also vary across radiologists. The private dataset shows two kinds of annotation variability: manual lesion segmentation and lesion scoring variability.

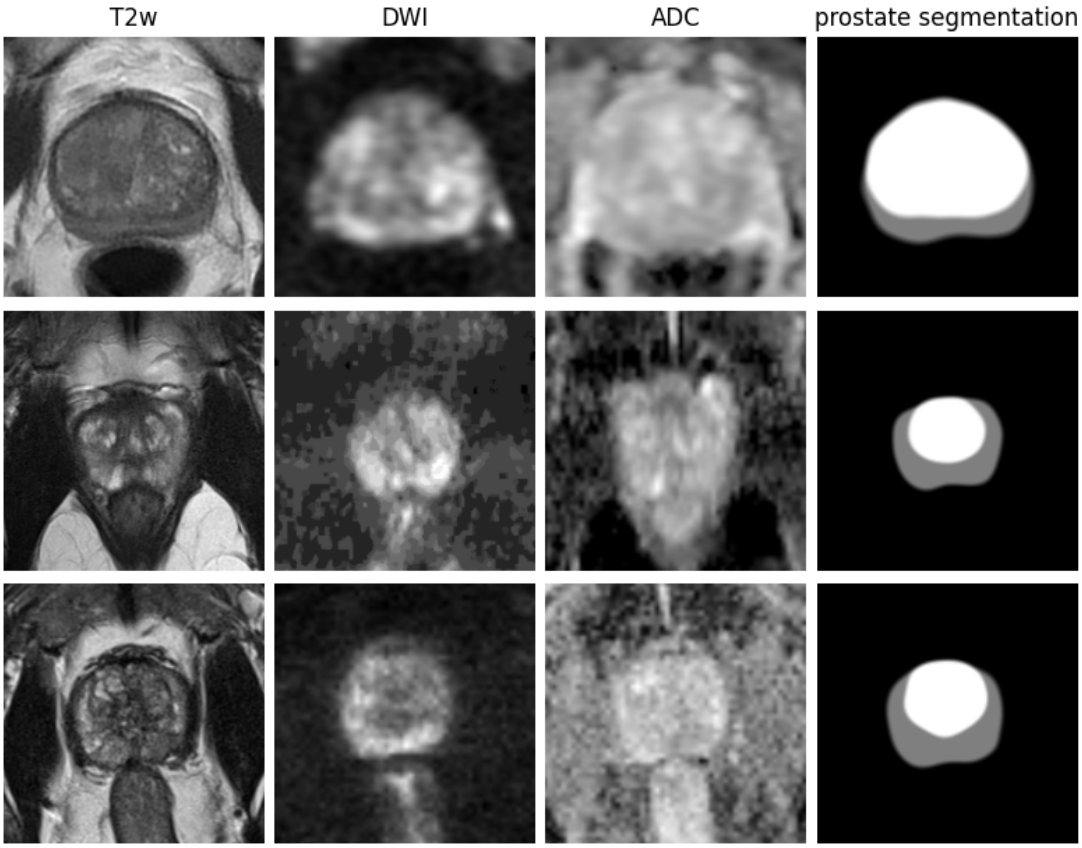


Figure 4.3: Extract of the prostate dataset used with the three MRI sequences (T2w, DWI (b $\geq$ 1200s/mm2), ADC mapping) and the bizonal prostate segmentation (peripheral zone in gray, transitional zone in white).

Manual lesion segmentation variability

Figure 4.4 shows different examinations with different annotations sets. We see that the amount of consensus between radiologists varies depending on the cases. In the first and third columns of Figure 4.4, radiologists have found the same lesion, their manual delineations differ, and the variability is higher in the first example. The second column shows a more problematic example: radiologist A0 has found a lesion while radiologist A1 has not. In the fourth column, radiologist A0 has found one lesion (on the top right) that has not been segmented by other radiologists who agree on the lesion found on the bottom right.

Figure 4.5 shows the average Dice score between annotations available for each exam. We see that 42% of exams have an average annotation Dice strictly below 0.6. We propose to model annotation variability through a redundancy approach: if an exam has been annotated by n radiologists, having n lesion segmentation masks, it is added n times in the dataset with the associated annotation. The rationale here is that, during fully-supervised training, more weight will be given to consistent annotations.

Considering the existing methods in the literature [Warfield et al., 2004, Cardoso et al., 2013, Zhang et al., 2023], this redundancy approach might be overly simplistic. One approach could entail using probability segmentation masks taking annotator consensus

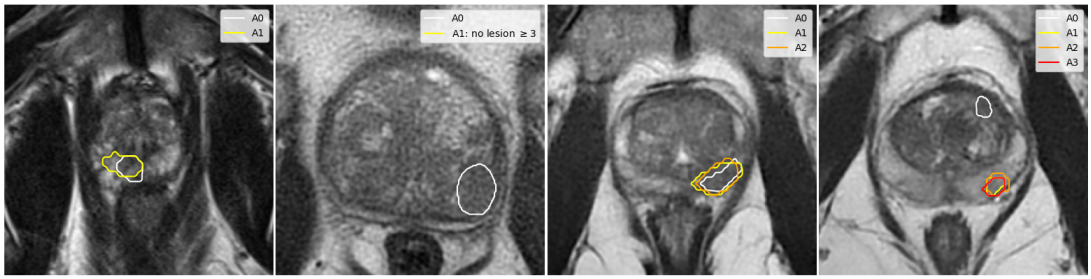


Figure 4.4: Slice from four different prostate examinations with the associated manual lesion segmentations, each contour level is associated with a radiologist Ai. The fourth column shows a case where one annotator (A0) has not found the same lesion as the others.

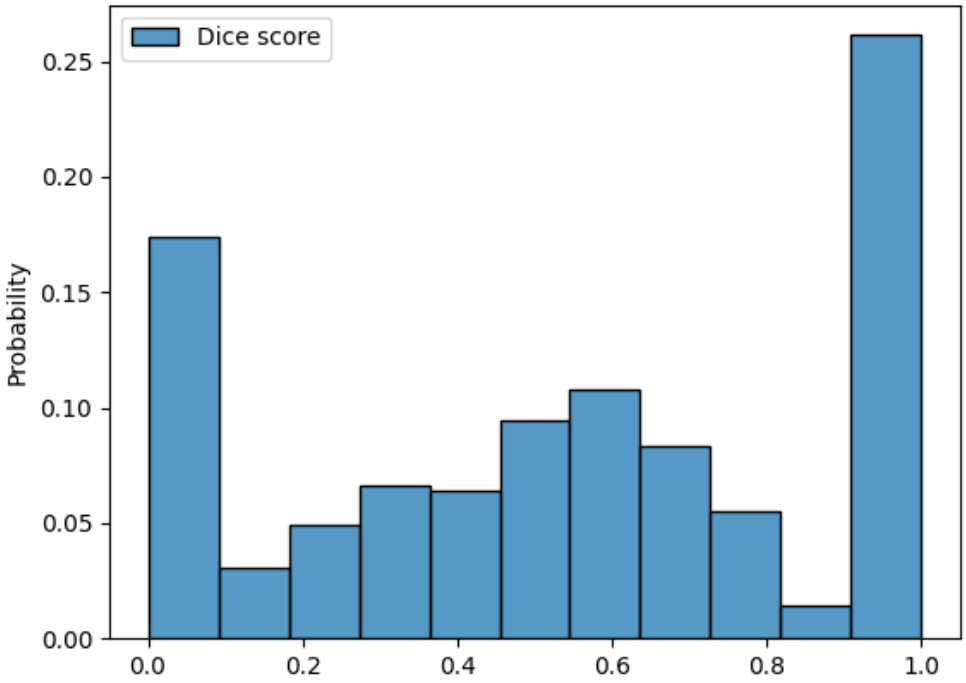


Figure 4.5: Histogram of average annotation Dice score for each exam in the dataset.

into account: each pixel would contain the amount of consistent annotations divided by the total number of annotators. We tried this approach and did not obtain satisfactory results. We do not further investigate how to include the variability of lesion segmentation masks in fully supervised training as we focus on self-supervised learning approaches in this thesis.

Additionally, manually segmenting prostate lesions can be difficult and time-consuming given the size of the lesion (a clinically significant lesion is defined by a volume greater than 0.5 cubic centimeters by Turkbey et al. [2019]) compared to that of the input volume (an average prostate has a volume of 20-25 cubic centimeters [Strand et al.,

2017]). The global PI-RADS score of the examination, which can be extracted from radiologist reports, is a metadata that could be used during pre-training to help reduce annotation costs but is also subject to inter-annotator variability.

Lesion grading variability

As mentioned in Section 4.3.1, the found lesions are assigned a PI-RADS score. A global PI-RADS grading of the exam is defined by the radiologist by taking the highest PI-RADS value of the set of lesions in the exam. This global score does not need the manual delineation of any lesion and can be extracted from radiology reports. As mentioned in Section 4.1.2, PI-RADS grading is subject to inter-annotator variability which exists in our dataset as shown by Figure 4.6.

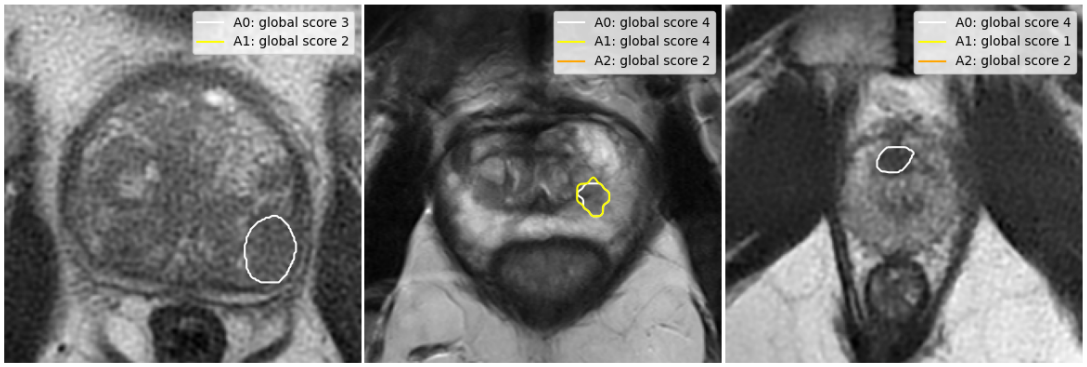


Figure 4.6: Slice from three different prostate examinations with the associated global grading given by radiologist Ai (when the global score is above 3 the manual lesion segmentation is also displayed).

Using global examination grading as a metadata in conditional contrastive pre-training can be useful as it could avoid manual lesion segmentation. However, lesion grading variability should be included in the kernels used in conditional contrastive loss functions.

4.3.3 Metadata analysis

To define an appropriate kernel taking metadata values into account, we perform a data analysis on the private dataset. Given a set of PI-RADS scores, we consider it to be coherent if all its values are either strictly above or below 3. In Figure 4.7, a PI-RADS score is considered coherent if it is part of a coherent set of scores. We see that PI-RADS 3 annotations are never coherent or only annotated by one annotator. A measure of coherence between scores could thus be defined and used to condition contrastive pre-training.

Figure 4.8 shows the distribution of annotators' numbers. We see that most exams are annotated by up to three annotators. The coherence definition could thus be weighted by the number of annotators providing the set of PI-RADS scores thus leading to a confidence measure: a set of coherent PI-RADS scores will be even more confident if provided by a large number of annotators.

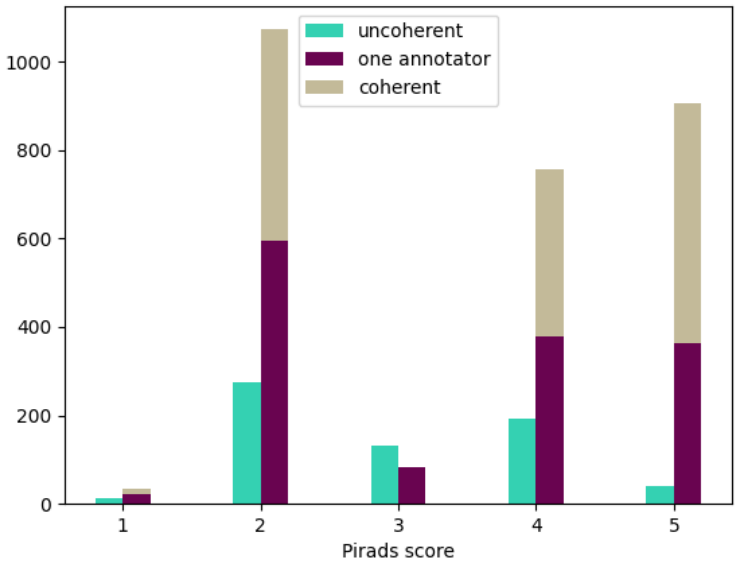


Figure 4.7: Pirads distribution stratified on annotation coherence.

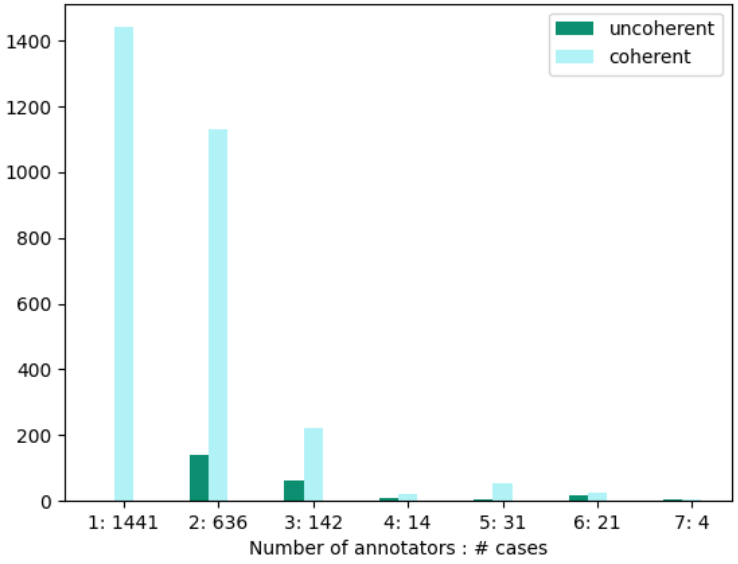


Figure 4.8: Histogram of annotator number.

4.4 Performing contrastive learning on bi-parametric MRI: combination strategy

As mentioned in Section 4.3.1, we are working with a dataset with bi-parametric MRI. With natural images, many works have investigated how to combine multimodal data (especially images and text) in contrastive pre-training. Most of these works propose to train parallel encoders, one for text and one for images, and combine the obtained latent representations in a modality-aware loss function [Yuan et al., 2021, Yang et al.,

2022a, Udandarao et al., 2020].

With medical images, Li et al. [2021b] use a CycleGan [Zhu et al., 2017] to generate images from different vendors. The generated images make the positive pairs in the contrastive loss function, thus exposed to multi-style images. For prostate cancer detection, Gutiérrez et al. [2022] build three parallel encoders taking as input ADC, T2w, and one DWI sequences of prostate lesion. The latent representations are concatenated and used to compute the contrastive loss function.

We thus investigate how to best combine the three modalities, along with the prostate segmentation mask in contrastive pre-training. Following the work of Gutiérrez et al. [2022], we propose to train three modality-specific encoders in parallel, concatenate the vectors of the latent representations, feed this concatenation to a projection head, and apply the contrastive loss function. We also propose to add contrastive loss functions at each encoder level. To take prostate segmentation into account, we update this architecture using prostate segmentation mask as a guide after the first encoder layers. We also pre-train three separate encoders with base contrastive learning [Chen et al., 2020a], freeze each encoder, and add a Dense layer followed by a projection head to optimize the contrastive loss function on the three modalities. Figure 4.9 shows a schematic view of the proposed approaches.

Training encoders in parallel leads to heavy architectures which are bound to create memory issues that cannot be solved by reducing the batch size as the number of negative pairs is important for contrastive learning approaches.

In the following sections, we use the three MRI sequences and the prostate segmentation mask in channels, as it was the best-performing approach when training the segmentation model from scratch. Table 4.1 shows preliminary experiments obtained while training a RetinaUnet [Jaeger et al., 2018] from scratch with different modality fusion strategies: putting all four sequences in channels (4 channels), building a batch with three consecutive slices and performing fusion operations (at some levels in the network, some feature maps are reshaped and fed to a convolutional layer to combine the feature maps of different sequences) in the network (3D fusion), building a batch with three consecutive slices for each modality (2.5D modality), and building a batch with the three modalities for three consecutive slices (2.5D slices).

Method	AUC
4 channels	**0.881**
3D fusion	0.854
2.5D modality	0.869
2.5D slices	0.846

Table 4.1: Results obtained while training a RetinaUnet [Jaeger et al., 2018] from scratch.

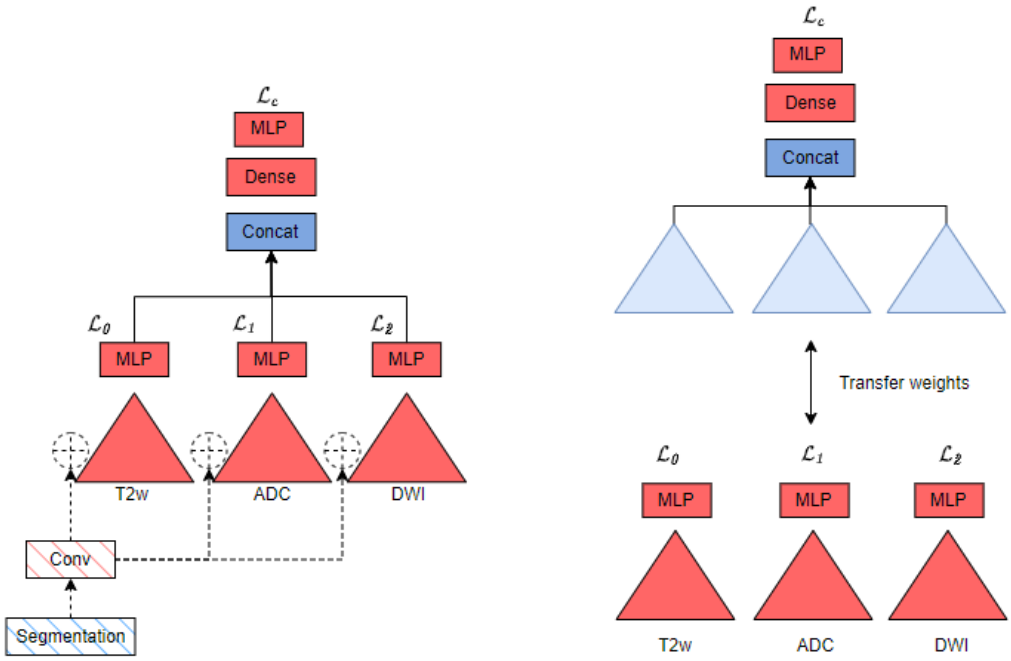


Figure 4.9: Proposed approaches for multi-modal contrastive learning. Left: three modality-specific encoders trained in parallel with (and without) modality-specific contrastive loss functions $\mathcal{L}_{0,1,2}$ followed by a concatenation of latent representations, and a new projection head for contrastive loss computation on concatenation $\mathcal{L}_c$. Dotted and hatched components represent the segmentation guide approach. Right: three modality specific encoders are separately trained with the base simCLR [Chen et al., 2020a] approach, they are then frozen and followed by a projection layer to perform contrastive learning on the concatenation of latent projections.

4.5 Conditional contrastive learning for prostate cancer detection

As shown in the previous section, prostate MRI examinations are assigned a PI-RADS score defining a malignancy level. This score is subject to low inter-reader reproducibility, as shown in Figure 4.4 and by Smith et al. [2019]. Some examinations, as those from the PI-CAI [Saha et al., 2022] dataset, are classified using biopsy results. This kind of classification is usually considered more precise (hence often taken as ground truth) but is also costly to obtain and presents a bias since only patients with high PI-RADS scores undergo a biopsy. Building a generic and automatic lesion detection method must therefore deal with the diversity of classification sources, radiology or biopsy, and the variability of classifications for a given exam.

We thus propose to build a conditional contrastive learning framework taking metadata such as PI-RADS and biopsy scores into account while also considering the variability of these scores.

4.5.1 Contrastive learning framework

We follow the conditional contrastive learning framework introduced in Section 4.2.2 and apply it to PI-RADS and biopsy scores that can have high variability.

To simplify the problem and homogenize PI-RADS and biopsy scores, we decide to binarize both scores, following clinical practice and medical knowledge [Epstein et al., 2015, Turkbey et al., 2019]. We set $y = 0$ for PI-RADS 1 and 2, and $y = 1$ for PI-RADS 4 and 5. We do not consider PI-RADS 3, since it has the highest inter-reader variability [Greer et al., 2019] and low positive predictive value [Westphalen et al., 2020]. This means that all exams with a PI-RADS 3 are considered deprived of metadata. For each exam i, a set of y values is available, noted $\mathbf{y}_i$. The number of annotations may differ among subjects. For a biopsy result (defining an ISUP classification [Epstein et al., 2015]), we set $y = 0$ if ISUP ≤ 1 and $y = 1$ if ISUP ≥ 2. Table 4.2 summarizes these definitions.

	$y = 0$	$y = 1$	$y = \varnothing$
PI-RADS	1,2	4,5	3
ISUP	ISUP ≤ 1	ISUP ≥ 2	

Table 4.2: Link between metadata value (PI-RADS and ISUP) and y score. PI-RADS 3 examinations are considered deprived of metadata ($y = \varnothing$).

To take advantage of the entire dataset, we also consider unannotated data for which metadata are not provided. When computing the loss function on an exam without metadata (no $\mathbf{y}$ associated), we use the standard alignment and uniformity function as defined by Wang and Isola [2020]. This leads to the following contrastive loss function:

$$\mathcal{L}_w = \begin{array}{l} \left.\begin{array}{l} \frac{1}{|A|}\sum_{i\in A}\Big(\sum_{j\in A} w(\mathbf{y}_j,\mathbf{y}_i)||x_i^1 - x_j^2||\Big) \\ +\log\big(\frac{1}{|A|^2}\sum_{i,j\in A}(1-w(\mathbf{y}_i,\mathbf{y}_j))e^{-||x_i^1-x_j^2||}\big) \end{array}\right\}\text{with metadata} \\ \left.\begin{array}{l} +\frac{1}{|U|}\sum_{i\in U}\Big(||x_i^1 - x_i^2||\Big) + \log\Big(\frac{1}{|U|^2}\sum_{\substack{i,j\in U\\ i\neq j}} e^{-||x_i^1-x_j^2||}\Big) \end{array}\right\}\text{without metadata} \end{array} \tag{4.3}$$

where A (resp. U) is the subset with (resp. without) associated $\mathbf{y}$ metadata.

Since the number of annotations may be different between two subjects i and j, we cannot use a standard kernel, as the RBF in [Dufumier et al., 2021b]. We would like to take metadata confidence, namely agreement among annotators, into account. In the following, we propose a kernel w to that end.

4.5.2 Including annotation confidence in kernel definition

Our measure of confidence is based on the discrepancy between the elements of vector $\mathbf{y}$ and their most common value (or majority vote). For exam i, if y_i is the most common value in its metadata vector $\mathbf{y}_i = [y_{i0}, y_{i1}, ...y_{in-1}]$ with n the number of available

scores[1], confidence c is defined as:

$$c(\mathbf{y}_i) = \begin{cases} \epsilon & \text{if} \quad n = 1 \\ 2 \times \left(\frac{\sum_{k=0}^{n-1} \delta(y_{ik}, y_i)}{n} - \frac{1}{2} \right) & \text{if} \quad n > 1 \end{cases} \tag{4.4}$$

where δ is the Dirac function and $\epsilon = 0.1$[2]. $c(\mathbf{y}_i) \in [0, 1]$, 0 is found when an even number of opposite scores is obtained and the majority vote cannot provide a decision. In that case, the associated exam will be considered as deprived of metadata.

Given this definition of confidence, we propose different kernels with increasing levels of complexity. We propose to see how different levels of confidence will impact fine-tuning performances after pre-training with different kernels. The three following subsections will present the different proposed kernels. We first introduce a simple kernel taking only highly confident metadata into account. We then use latent representation values to refine kernel definition on highly confident metadata. Finally, we define a kernel using the whole confidence definition from Equation (4.4) which conditions contrastive learning with metadata coherence and annotators number. Table 4.3 summarizes the similarities and differences between these approaches.

High confidence kernel

In this section, we propose to define a kernel only on examinations with metadata of maximum confidence.

As shown in Figure 4.8, most of the available annotations are either provided by one annotator or can be considered coherent (except for cases with six or seven annotators but they only represent 1% of the database). We start by considering only exams with confidence equal to 1 (defined in Equation (4.4)) or annotated by one annotator for which we also set $c(\mathbf{y}_i) = 1$ in this analysis.

Defining $c_{ij} = \min(c(\mathbf{y}_i), c(\mathbf{y}_j))$, we set the following kernel:

$$w_\delta(\mathbf{y}_i, \mathbf{y}_j) = \begin{cases} 1 & \text{if } i = j & \text{(exam against its own transformed version)} \\ \lambda\delta(c_{ij}, 1) & \text{if } y_i = y_j \text{ and } i \neq j & \text{(different exams, same majority vote)} \\ 0 & \text{if } y_i \neq y_j \text{ and } i \neq j & \text{(different exams, different majority vote)} \end{cases} \tag{4.5}$$

where $\lambda < 1$ is a parameter setting the amount of attraction of different exams with the same majority vote. We use $\lambda = 0.8$ in practice.

For two given exams i and j, the proposed method is interpreted as follows:

- an exam is perfectly aligned with its own transformed version;

[1] in the remainder of this manuscript, score and metadata indicate the binarization output of PI-RADS and biopsy scores without distinction

[2] The maximum number of metadata available for an exam is $n = 7$, the minimal achievable confidence value is thus $c = 2(4/7 - 1/2) > 0.14$. We set ϵ so that the confidence for $n = 1$ is higher than 0 but less than the minimal confidence when n is odd.

- different exams with the same majority vote and considered confident ($c_{ij} = 1$ i. e. $c(y_i) = c(y_j) = 1$ or $n = 1$) are aligned but do not collapse at the same point as $\lambda < 1$;
- different exams with different majority votes are not attracted, they will be repulsed by the second term of Equation (4.3).

Figure 4.10 shows a schematic view of this approach.

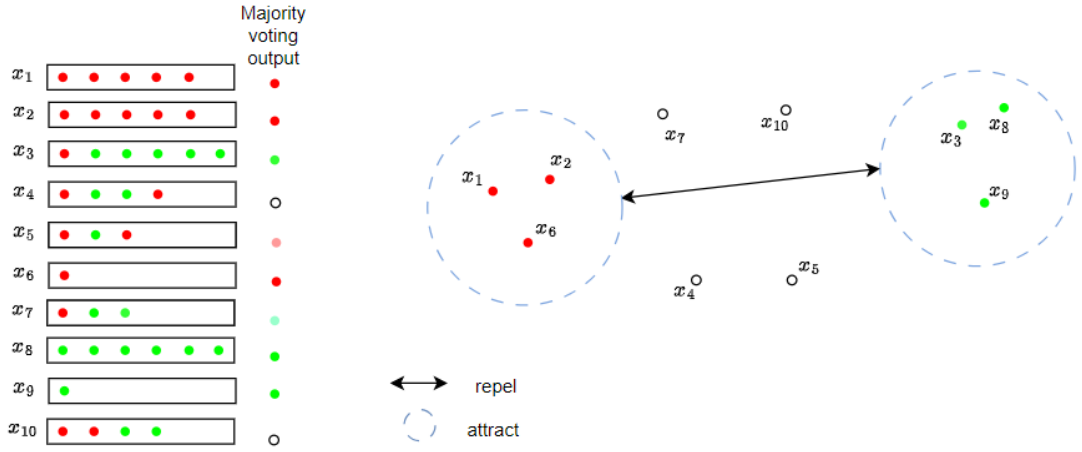


Figure 4.10: Given a set of exams latent representations $(x_i), i \in [1, 10]$, $\mathbf{y}_i$ is represented as a list of colored points. Confidence (c) is represented with color saturation: darker means more confident. Exams with confidence smaller than 1 (lighter coloring) are considered unlabeled and uncolored in the right part of the plot. Exams with a similar confident majority vote y will be attracted. Groups of exams with different y scores are repelled.

This approach might be overly simplistic as alignment and uniformity of latent representations of different exams with the same majority vote is conditioned on only one parameter λ that has to be set and tuned. If λ is too close to 1, different exams with the same majority vote will all collapse at the same point, which amounts to no conditioning. If λ is too close to 0, too few alignment will be performed and latent representations of different exams with the same majority vote will be uniformly repulsed.

Heaviside kernel

To better control the amount of attraction and repulsion between latent representations of images with the same confident majority vote, we introduce another kernel formulation. We propose to condition alignment and uniformity using values of the latent representations of views. This leads to rewrite the second line of Equation (4.5) as follows:

$$w_H(\mathbf{y}_i, \mathbf{y}_j, x_i^1, x_j^2) = \begin{cases} H(||x_i^1 - x_j^2||^2 - \tau)\delta(c_{ij}, 1) & \text{if } y_i = y_j \text{ (different images with and } i \neq j \text{ same majority vote)} \end{cases} \tag{4.6}$$

where H is the Heaviside function. Latent representations of data with the same confident majority votes are attracted only if their distance is above a user-defined threshold τ. We will consider both global and conditional uniformity approaches (introduced in Section 4.2.2). Using conditional uniformity, latent representations of data

with similar confident majority votes are pushed apart provided their distance is below τ. Figure 4.11 shows a schematic view of our approach. We use similar notations and definitions as those introduced in Figure 4.10.

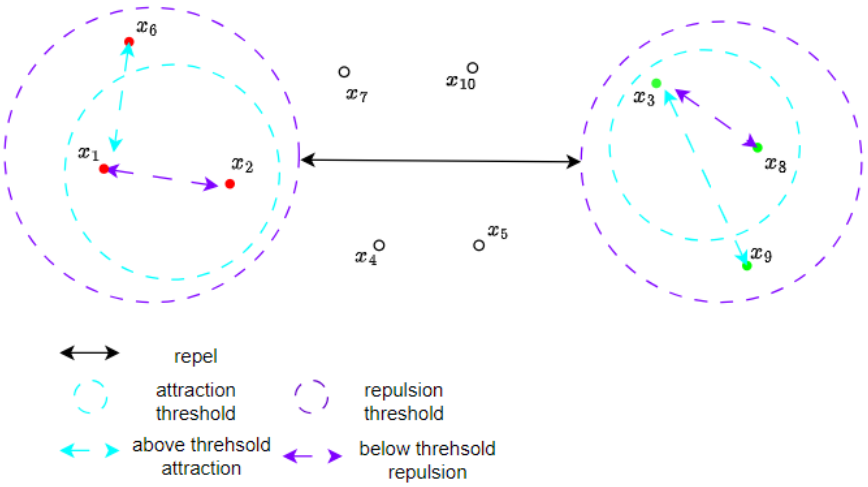


Figure 4.11: Attraction and repulsion between latent representations of examinations with similar confident majority votes are conditioned by attraction and repulsion thresholds.

Parameter τ is chosen such that latent representations of two different views are separated by, at most, an angle of $\pi/3$. This threshold is arbitrarily set which is not convincing, as shown in Section 4.5.4. A search for the correct τ value would need multiple pre-training and fine-tuning experiments, which is time and resource-consuming. In the following section, we propose to define another kernel that does not introduce a supplementary hyper-parameter to be tuned.

Both the aforementioned kernels consider only highly confident samples which amount to ignore inter-reader variability. This is overly simplistic considering clinical reality. We thus propose a final kernel expression taking all cases into account.

Final kernel

Taking all examinations into account rather than only highly confident ones, we get the following expression for different exams and the same majority vote (lines 1 and 3 of Equation (4.5) are unchanged):

$$w(\mathbf{y}_i, \mathbf{y}_j) = \begin{cases} c_{ij} & \text{if } y_i = y_j \text{ and } i \neq j \end{cases}^{\substack{\text{(different exams,} \\ \text{same majority vote)}}} \tag{4.7}$$

where $c_{ij} = \min(c(\mathbf{y}_i), c(\mathbf{y}_j))$. For two given exams i and j, the proposed model is interpreted as follows:

- If both metadata confidences are maximal ($c_{ij} = 1$), $w(\mathbf{y}_i, \mathbf{y}_j)$ will be equal to 1 and full alignment will be computed.
- If either metadata confidence is less than 1, $w(\mathbf{y}_i, \mathbf{y}_j)$ value will be smaller and exams will not be fully aligned in the latent space. The less confidence, the less aligned exams i and j representations will be.

- If confidence drops to zero for either exam, the exam will only be aligned with its own transformed version.

Similarly to decoupled contrastive learning [Yeh et al., 2022], we design w such that the second term of Equation (4.3) does not repel samples with identical majority vote and maximal confidence ($c_{ij} = 1$). Figure 4.12 shows a schematic view of our approach. Previously introduced notations and definitions are unchanged.

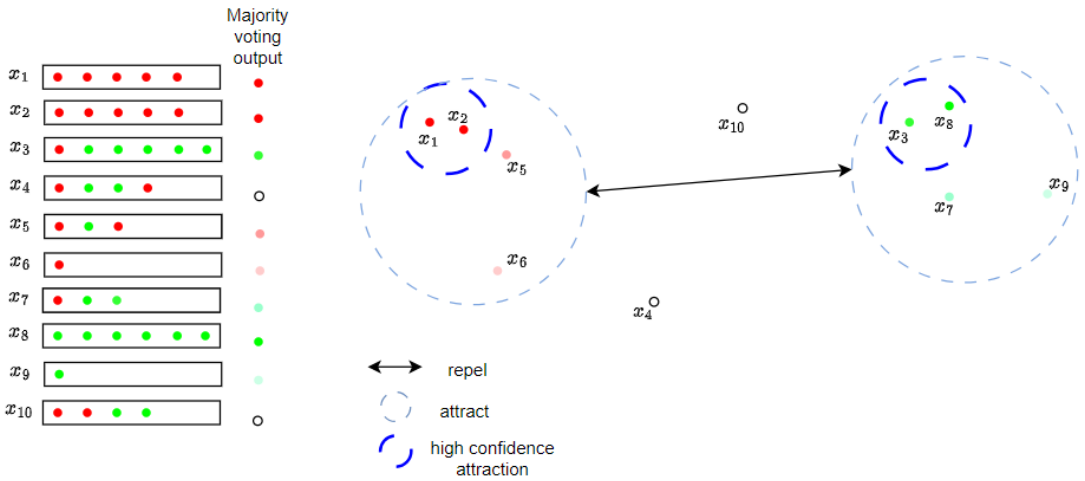


Figure 4.12: Exams such that $c(\mathbf{y}_i) = 0$ (no decision from majority vote) are considered as unlabeled and uncolored. Exams such that $c(\mathbf{y}_{i,j}) = 1$ and $y_i = y_j$, e.g. (x_1, x_2) (resp. (x_3, x_8)), will be strongly attracted while less attracted to patients with $c(\mathbf{y}_i) < 1$, e.g. $x_{5,6}$ (resp. $x_{7,9}$). Groups of exams with different y scores are repelled.

Table 4.3: Summary of the different kernel expressions.

	$i = j$	$y_i = y_j$ and $i \neq j$	$y_i \neq y_j$ and $i \neq j$
w_δ	1	$0.8\delta(c_{ij}, 1)$	0
$w_H(\mathbf{y}_i, \mathbf{y}_j, x_i^1, x_j^2)$	1	$H(\|\|x_i^1 - x_j^2\|\|^2 - \tau)\delta(c_{ij}, 1)$	0
$w(\mathbf{y}_i, \mathbf{y}_j)$	1	c_{ij}	0

4.5.3 Experiments

We perform experiments on both our private and the PI-CAI public dataset [Saha et al., 2022]. Our dataset is composed of 2415 multi-parametric MRI prostate examinations as introduced in Section 4.3. 1397 of them have manual annotations: PI-RADS metadata and manual lesion segmentation provided by up to seven radiologists. Pre-training is performed with data from both datasets on 3915 exams. Fine-tuning is then performed using 1% and 10% of these exams using cross-validation (see **Implementation details** section).

PI-CAI dataset

PI-CAI dataset [Saha et al., 2022] was released for a global challenge of clinically significant prostate lesion detection. It is composed of 1,500 MRIs of patients suspected to have clinically significant prostate cancer. Out of the 1500 cases, 1075 have non clinically significant cancer. Among the 425 cases of clinically significant prostate cancer,

220 have manual lesion segmentation provided by a radiologist. PI-RADS scores are not available for this dataset but a biopsy score (ISUP value [Epstein et al., 2015]) is. We will use the provided ISUP scores as metadata for conditional contrastive learning.

Bizonal prostate segmentations are not provided for this database. We use our internal algorithm to infer bizonal prostate segmentation on T2w PI-CAI sequences.

Implementation details

For both datasets, we apply the following preprocessing. Both DWI and ADC sequences are registered on the T2w sequence using the registration method proposed by Yang et al. [2016]. A crop around an extended prostate region of interest at a resolution of (24,224,224) is applied. Unit normalization is then performed.

We pre-train a 3D U-Net encoder followed by a multi-layer perceptron projection head on our conditional contrastive learning approach. Similarly to nn U-Net [Isensee et al., 2020], the encoder is a fully convolutional network where spatial anisotropy is used (e.g. axial axis is downsampled with a lower frequency since MRI volumes often have lower resolution in this direction). It is composed of four convolution blocks with one convolution layer in each block and takes the four sequences as input in the channel dimension. The projection head is a two-layer perceptron as in [Chen et al., 2020a]. We train with a batch size of 16 for 100 epochs and use a learning rate of 10^{-4}. Following the work of Fernandez-Quilez et al. [2022] on contrastive learning for prostate cancer triage, we use a random sampling of rotation, translation, and horizontal flip to generate the perturbed versions of the images.

To evaluate the impact of contrastive pre-training at low data regime, we perform fine-tuning with 10% and 1% of annotated exams. The contrastive pre-trained encoder is used to initialize the U-Net encoder, the whole encoder-decoder architecture is then fine-tuned on the supervised task. Fine-tuning is performed with 5-fold cross-validation with both datasets using the pre-trained encoder. Using 1% (resp. 10%) of annotated data, each fold has 39 (resp. 269) training data and 12 (resp. 83) validation data. We build a hold-out test set of 500 volumes[3], not used during any training step with data from both datasets to report our results. We also compare fine-tuning from our pre-trained encoder to a model trained from random initialization.

Computing infrastructure. Optimizations were run on GPU NVIDIA T4 cards.

Evaluation measures

The 3D U-Net network outputs lesions segmentation masks which are thresholded, following the dynamic thresholding proposed by Bosma et al. [2023], and of which connected components are computed. For each connected component, a detection probability is assigned as the maximum value of the network output in this component. The output of this post-processing is a binary mask associated with a detection probability per lesion. Figure 4.13 shows an example of this post-processing.

[3]The 100 validation cases on the PI-CAI challenge website being hidden we could not compare our methods to the leaderboard performances.

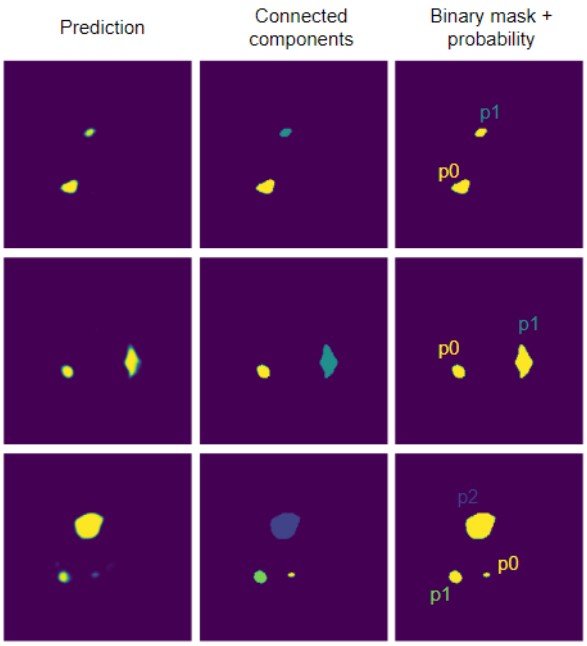


Figure 4.13: Example of our post-processing approach: the network outputs prediction mask (left), connected components are computed (middle) and a binary mask is generated with associated lesion detection probabilities (right).

We compute the overlap between each lesion mask and the reference mask. A lesion is considered a true positive (detection) if the overlap with the reference is above 0.1 as defined by Saha et al. [2021]. This threshold is chosen to keep a maximum number of lesions to be analyzed for AUC computation. Different threshold values are then applied for AUC computation.

As in [Saha et al., 2021, Yu et al., 2020], lesion detection probability is used to compute AUC values at exam and lesion levels, and average precision (mAP). To compute AUC at exam level we take, as ground truth, the absence or presence of a lesion mask, and, as a detection probability, the maximum probability of the set of detected lesions. At lesion level, all detection probabilities are considered and thresholded with different values, which amounts to limiting the number of predicted lesions. The higher this threshold, the lower are sensitivity and the number of predicted lesions, and the higher is specificity.

4.5.4 Results

We compared the proposed methods with different state-of-the-art contrastive learning approaches: alignment and uniformity [Wang and Isola, 2020] (named Unif align in results tables and introduced in Chapter 3), simCLR [Chen et al., 2020a], MoCo v2 [Chen et al., 2020b], nearest neighbor contrastive learning [Dwibedi et al., 2021] (named nnCLR in results tables) and noncontrastive approaches: BYOL [Grill et al., 2020], Barlow Twins [Zbontar et al., 2021] and simSiam [Chen and He, 2020] (all of these methods were presented in Chapter 2). Results with 1% annotated data are presented in Table 4.4. Results with 10% annotated data are in Appendix C.

Comparison with state of the art approaches

For both datasets, we see that including metadata confidence to condition alignment and uniformity in contrastive pre-training yields better performances than previous state-of-the-art approaches and random initialization. The discrepancy between PI-CAI and private mAP is due to the nature of the dataset: the PI-CAI challenge was designed to detect lesions confirmed by biopsy, while our private dataset contains lesions not necessarily confirmed by biopsy. Our private dataset contains manually segmented lesions that might be discarded if a biopsy was performed. The model being fine-tuned on both datasets, PI-CAI exams are overly segmented, which leads to lower mAP values (since our model tends to over-segment on biopsy ground truths). For our clinical application, which aims to reproduce radiologist responses, this is acceptable.

We report significant performance improvement at very low data regime (1% annotated data) compared to existing methods which is a framework often encountered in clinical practice.

Ablation studies

To assess the impact of our approach we perform different ablation studies:

- We remove confidence and use majority vote for kernel computation (Majority voting row in Table 4.4). If two different exams have the same majority vote we set: $w(y_i, y_j) = 0.8$ in Equation (4.7), other w values are kept unchanged. We can see that using the majority vote without taking confidence into account leads to decreased performances.

- We set the confidence of PI-CAI exams to 1 (increasing biopsy confidence, biopsy row in Table 4.4) which amounts to setting $\epsilon = 1$ for PI-CAI exams in Equation (4.4). No particular improvement is observed with this approach.

Comparison between the different proposed kernels

For all of the experienced kernels, we remove the conditioning on uniformity. Exams are uniformly repelled rather than conditioning on metadata similarity for repulsion (which amounts to setting $w(\mathbf{y}_i, \mathbf{y}_j) = 0$ for the second term of Equation (4.3)). The results of these experiments are identified with the GlU prefix in Table 4.4.

We see that for most of the proposed kernels, conditioning the uniformity term with metadata similarity gives better results than performing global uniformity. However, this is not the case with the Heaviside kernel, which is due to the introduction of τ parameter that would need better tuning.

We also see that introducing annotator confidence in kernel definition leads to better performances than when using only high confident exams (with w_δ and w_H kernels).

Results on noncontrastive methods

We see that Barlow Twins, BYOL, and simSiam approaches, which propose different loss definitions and optimization strategies, outperform our approach on some metrics for one dataset or the other. Further work could focus on introducing our proposed confidence kernel along with these methods, as introduced by Tsai et al. [2021].

Table 4.4: 5-fold cross-validation mean AUC and mAP after fine-tuning on PI-CAI and private datasets with 1% of annotated data (standard deviation in parentheses).

Method	AUC exam		AUC lesion		mAP	
	PI-CAI	Private	PI-CAI	Private	PI-CAI	Private
Random init	0.68 (0.06)	0.74 (0.03)	0.73 (0.11)	0.70 (0.05)	0.27 (0.05)	0.62 (0.03)
Unif align	0.66 (0.07)	0.72 (0.01)	0.64 (0.13)	0.68 (0.03)	0.28 (0.07)	0.63 (0.03)
simCLR	0.64 (0.07)	0.73 (0.05)	0.65 (0.08)	0.68 (0.05)	0.22 (0.07)	0.60 (0.06)
MoCo v2	0.63 (0.08)	0.71 (0.04)	0.59 (0.12)	0.64 (0.07)	0.24 (0.10)	0.58 (0.06)
nnCLR	0.57 (0.08)	0.73 (0.05)	0.49 (0.09)	0.62 (0.06)	0.21 (0.05)	0.59 (0.05)
BYOL	0.67 (0.06)	0.72 (0.04)	0.66 (0.16)	0.68 (0.04)	0.26 (0.05)	0.59 (0.04)
Barlow Twins	0.67 (0.06)	**0.76** (0.03)	0.61 (0.09)	0.68 (0.06)	0.25 (0.08)	0.63 (0.04)
simSiam	0.67 (0.06)	0.74 (0.04)	**0.75** (0.06)	0.69 (0.05)	0.29 (0.06)	0.64 (0.05)
$w\delta$ (4.5)	0.66 (0.08)	0.75 (0.04)	0.60 (0.06)	0.67 (0.03)	0.28 (0.09)	0.62 (0.03)
GlU $w\delta$	0.65 (0.05)	0.74 (0.02)	0.60 (0.14)	0.69 (0.03)	0.29 (0.06)	0.64 (0.04)
Majority voting	0.63 (0.06)	0.74 (0.02)	0.62 (0.06)	0.69 (0.04)	0.28 (0.07)	0.61 (0.04)
w_H (4.6)	0.64 (0.08)	0.74 (0.03)	0.62 (0.04)	0.67 (0.02)	0.24 (0.07)	0.62 (0.04)
GlU w_H	0.65 (0.07)	**0.76** (0.01)	0.61 (0.11)	0.70 (0.05)	0.26 (0.07)	**0.65** (0.02)
Biopsy	0.64 (0.06)	0.73 (0.03)	0.69 (0.06)	0.70 (0.03)	0.24 (0.05)	0.62 (0.04)
GlU w	0.60 (0.05)	0.74 (0.03)	0.60 (0.12)	**0.73** (0.02)	0.23 (0.05)	0.63 (0.04)
Ours w (4.7)	**0.70** (0.05)	0.75 (0.03)	**0.75** (0.10)	0.71 (0.03)	**0.30** (0.09)	0.63 (0.04)

4.5.5 Qualitative results

Figure 4.14 shows the impact of our pre-training method on the outputs of the fine-tuned U-Net. Without conditioning, some lesions are missed (false negative cases FN 1, FN 2) and others are falsely detected (false positive cases FP 1, 2, and 3). Adding the conditioned pre-training removes these errors.

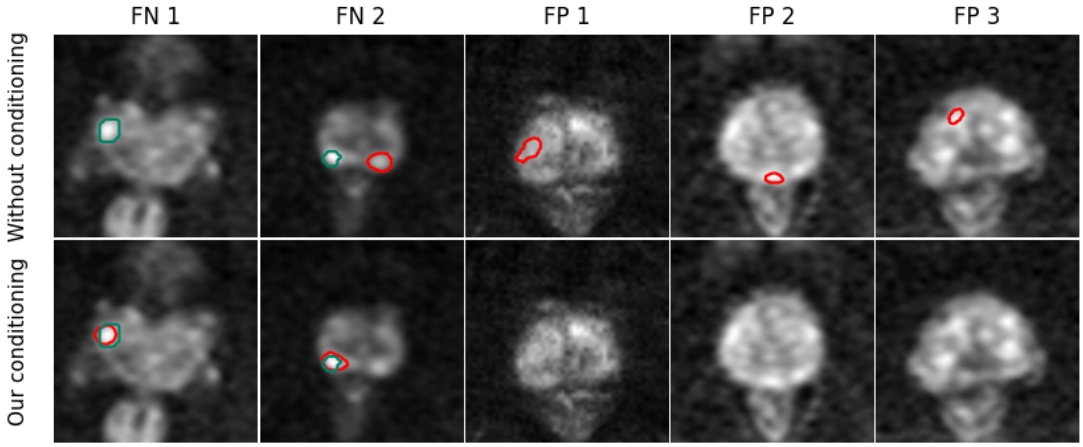


Figure 4.14: Examples of false negative (FN) and false positive (FP) cases (first row) corrected by the proposed method (second row). Reference segmentation: green overlay, predicted lesions: red overlay.

Figure 4.15 shows cases where conditioning pre-training with metadata confidence helps refine the segmentations of lesions predicted by the model. The first and third columns show cases where the predicted lesion is closer to the ground truth after fine-tuning the model pre-trained with our approach. The second column shows a case where

two lesions (among which one false positive) are predicted by the model pre-trained without conditioning. Pre-training the model with metadata conditioning removes the false positive lesion.

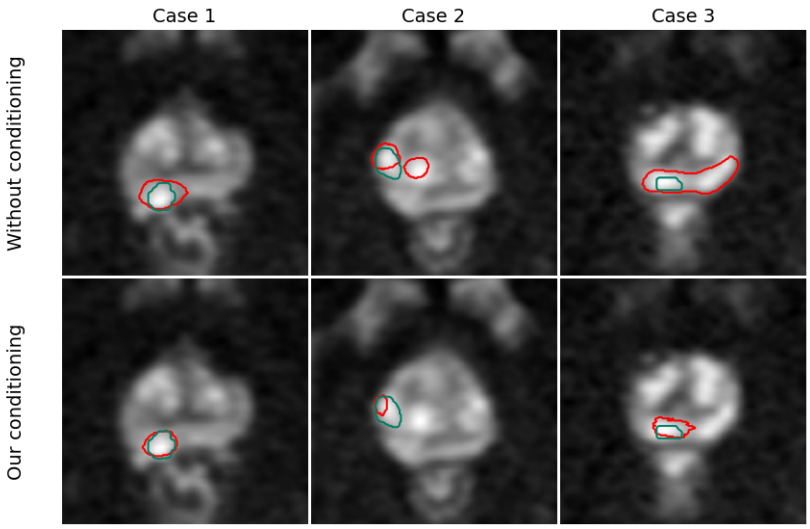


Figure 4.15: Example cases where conditioning with PI-RADS and biopsy scores helps refine lesions segmentation predictions. Reference segmentation: green overlay, predicted lesions: red overlay.

To evaluate the quality of the learned representation we perform dimensionality reduction of the latent representations of the training and test set data. These two sets of data are fed to the trained encoder, and the latent representations before the projection layer are used to fit a 2D tSNE projection. Each point in Figures 4.16 and 4.17 shows the latent projection of a data point on a 2D plane.

On the training set, Figure 4.16 shows a better separation between classes with our approach and global uniformity than Barlow Twins or simCLR.

On the test set the separation is less clear. This separation in the latent space of the encoder does not translate into large AUC differences after fine-tuning.

This might be due to the fine-tuning process which trains the full encoder-decoder architecture. The structure learned in the encoder latent space during pre-training ends up being altered while training the full architecture on the segmentation task adding the multiple layers decoder.

This limitation, along with the work of Chaitanya et al. [2023], advocates for performing contrastive pre-training on the decoder (see Section 4.6.2).

4.6 Two ideas for improvement: preliminary results

4.6.1 Computing pseudo labels with nearest neighbors

To improve the previously introduced confidence kernel (Equation (4.7)), we propose to build on existing approaches on contrastive learning with nearest neighbors [Dwibedi et al., 2021] and pseudo labels [Bošnjak et al., 2023].

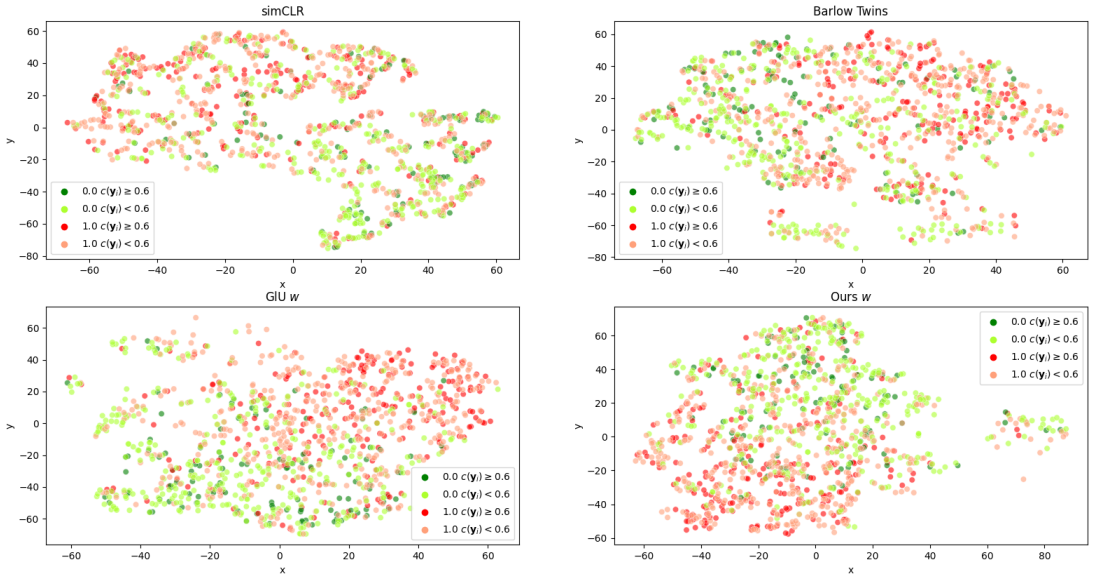


Figure 4.16: tSNE projection of private training set data for different approaches (subjects deprived of metadata are not shown for better readability). Dark (respectively light) green and red points represent subjects with high (respectively low) confidence.

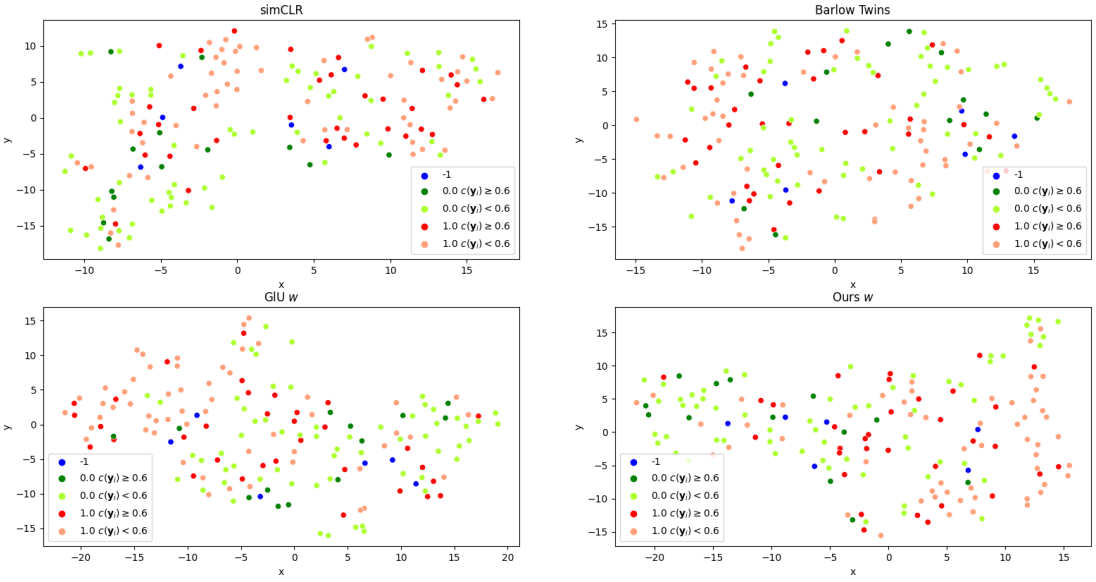


Figure 4.17: tSNE projection of private test set data for different approaches. Dark (respectively light) green and red points represent subjects with high (respectively low) confidence -1 labels represent subjects deprived of metadata.

We propose to compensate for the lack of annotations for a patient by using the annotations of the nearest neighbors in the latent space. We build a support set S of 100 cases with 50 confident cases such that $c(\mathbf{y}_i) = 1$ and $y = 0$ and 50 confident cases such that $c(\mathbf{y}_i) = 1$ and $y = 1$.

During training, we redefine the metadata list for cases with one metadata only ($n = 1$ in Equation (4.4)). Given y_{i0} the score for patient i and x_i its projection in the latent space, we define the new scores list as $\mathbf{y}_i = [y_{i0}, y_{n_0}, y_{n_1}, ...y_{n_m}]$ where n_j are the nearest neighbors of patient i from the support set in the latent space: $n_j = NN(x_i, S)_j$. We use $m = 6$ nearest neighbors to get a new set of 7 scores, 7 being the maximum number of annotators in the database.

For cases with one annotator, we get a new majority vote from pseudo labels and the vote confidence becomes:

$$c(\mathbf{y}_i) = \left\{ \; 2 \times \left(\frac{\sum_{k=0}^{n-1} \delta(y_{ik}, y_i)}{n} - \frac{1}{2} \right) \right. \tag{4.8}$$

where y_{ik} are either true scores or pseudo-labels from nearest neighbors.

Figure 4.18 shows a schematic view of this approach (we follow the same definitions and annotations as in Figure 4.12).

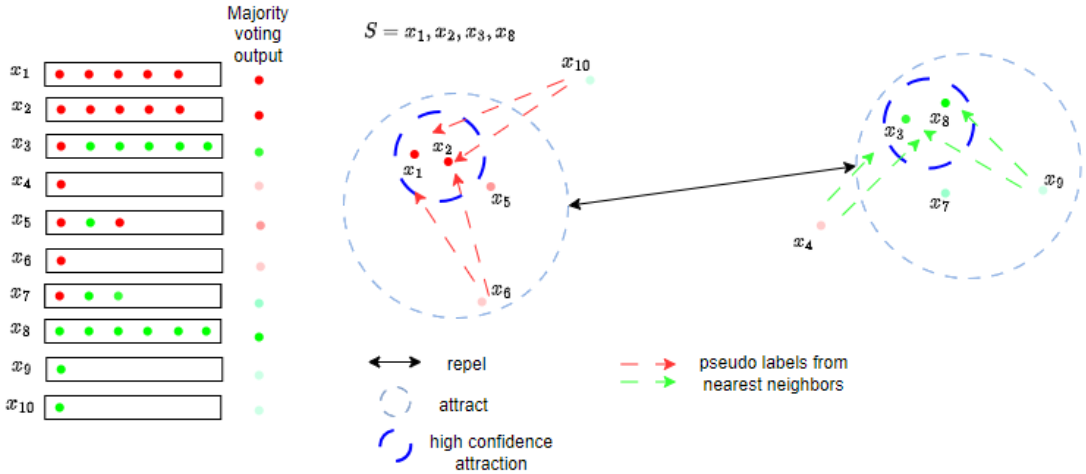


Figure 4.18: Latent representations of exams with only one annotation: x_4, x_6, x_9, x_{10} will be compared to latent representations of confident exams from the support set $S = [x_1, x_2, x_3, x_8]$. Pseudo labels are defined by taking the majority vote of 2 (in this example) nearest neighbors from the support set (represented as dotted arrows).

We compare our method with nearest neighbors with MoCo v2 [Chen et al., 2020b] and nnCLR [Dwibedi et al., 2021] approaches from the literature and the previously proposed approach with confidence kernel (presented in Section 4.5). We also propose to design the memory bank used in MoCo and the support set of nnCLR to contain only patients with scores of maximum confidence. These experiments are ablation studies to evaluate the impact of computing pseudo labels rather than designing a specific memory bank with confident samples.

Table 4.5 shows fine-tuning results with 1% of annotated data (results with 10% of annotated data are in Appendix C).

We see that using a better-designed support set (respectively memory bank) using only confident samples with nnCLR (respectively MoCo) approach improves the AUC at lesion level. Furthermore, the proposed nearest neighbor approach does not yield better

Table 4.5: 5-fold cross-validation mean AUC and mAP after fine-tuning on PI-CAI and private datasets with 1% of annotated data with nearest neighbors approach (standard deviation in parentheses).

Method	AUC exam		AUC lesion		mAP	
	PI-CAI	Private	PI-CAI	Private	PI-CAI	Private
Random init	0.68 (0.06)	0.74 (0.03)	0.73 (0.11)	0.70 (0.05)	0.27 (0.05)	0.62 (0.03)
MoCo	0.63 (0.08)	0.71 (0.04)	0.59 (0.12)	0.64 (0.07)	0.24 (0.10)	0.58 (0.06)
MoCo confident	0.60 (0.05)	0.71 (0.03)	0.63 (0.13)	0.70 (0.04)	0.19 (0.05)	0.57 (0.05)
nnCLR	0.57 (0.08)	0.73 (0.05)	0.49 (0.09)	0.62 (0.06)	0.21 (0.05)	0.59 (0.05)
nnCLR confident	0.57 (0.06)	0.71 (0.04)	0.56 (0.13)	0.68 (0.06)	0.20 (0.05)	0.59 (0.04)
Ours w (4.7)	**0.70** (0.05)	**0.75** (0.03)	**0.75** (0.10)	**0.71** (0.03)	**0.30** (0.09)	**0.63** (0.04)
Ours w (4.7) with nn	0.67 (0.06)	0.74 (0.03)	0.66 (0.13)	0.69 (0.04)	**0.30** (0.09)	0.61 (0.05)

results than using constant confidence for data with one annotation (defined in Equation (4.4)) although the data of the support set are quite well clustered. Figure 4.19 shows a tSNE projection of the latent representations of data in the support set during training: the latent representations of healthy and pathological cases making the support set are well separated during training, especially in epochs 40 to 70.

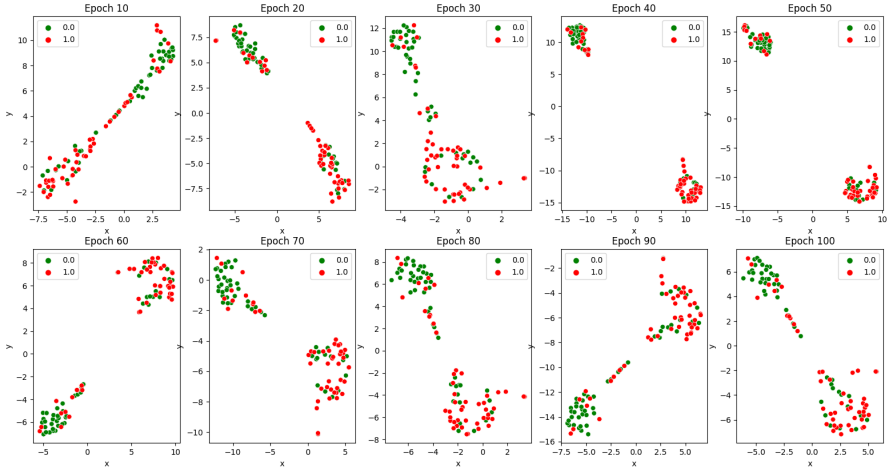


Figure 4.19: tSNE plot of latent representations of data from the support set along training epochs.

To improve the performances of this approach, further works could focus on support set design and optimization strategies:

- The size of the support set: we take 100 cases arbitrarily but a throughout analysis of the impact of the support set size should be performed.
- The selected cases for the support set: the 100 confident cases are chosen arbitrarily in the dataset but some other cases might be better prototypes of confident cases. We expect the impact of the support set composition to reduce as its size grows.

- Computing pseudo labels using nearest neighbors in the latent space is strongly dependent on the representation quality and clustering in the latent space. To improve the reliability of pseudo labels computed with nearest neighbors, a two-stage training strategy could be envisaged: during a few epochs, a conditional contrastive pre-training could be performed using only confident cases with confidence above a certain threshold.

- The distance of each nearest neighbor could be taken into account when defining the new pseudo label and its confidence which would lead to:

$$c(\mathbf{y}_i) = 2 \times \left(\frac{1}{\sum_{k=1}^{m} (||x_i - n_k|| + 1)} \left(\delta(y_{i0}, y_i) + \frac{\sum_{k=1}^{m} \delta(y_{n_k}, y_i)}{||x_i - n_k|| + 1} \right) - \frac{1}{2} \right) \quad (4.9)$$

  where $n_k = NN(x_i, S)_k$ the k-th nearest neighbor in the latent space.

Preliminary results for some of these propositions are presented in Table 4.6. These results are obtained by pre-training the network with the proposed approaches on a subset of the dataset composed of 794 cases, fine-tuning on 324 cases, and testing on 48 cases. We see that changing the support set composition or its size has an important impact on some metrics which advocates for further investigations on its design.

Method	AUC exam	AUC lesion	mAP
$\|S1\| = 100$	0.71	0.64	0.47
$\|S2\| = 100$	0.79	0.57	0.49
$\|S1\| = 300$	0.71	0.51	0.52
$\|S1\| = 50$	0.70	0.58	0.46
$\|S1\| = 10$	0.76	0.56	0.56

Table 4.6: Preliminary results changing the support set with fixed size ($S1$ and $S2$ rows) and changing the support set size once fixed (first, third, fourth, and fifth rows).

4.6.2 Adding contrastive pre-training to decoder

As previously mentioned, pre-training an encoder with contrastive learning for a target segmentation task might not be powerful enough. We propose, as a second perspective for prostate lesion detection with contrastive learning, to investigate the approaches proposed by Chaitanya et al. [2020], Zheng et al. [2021] and Chaitanya et al. [2023] performing contrastive pre-training at the decoder level.

Zheng et al. [2021] (HSSL) combine contrastive and reconstruction loss functions. Encoder features at different scales are concatenated before being fed to a projection head after which the contrastive loss function is computed. The decoder is trained to reconstruct the unperturbed input image. Figure 4.20 shows a schematic view of this approach. To enforce a consistency constraint, we only use one perturbed version of the input image, the second view being the unperturbed input image. A balance has to be found between the batch size, which needs to be large enough for contrastive learning, and the number of decoder blocks to be able to use before having memory issues. We train HSSL with one decoder block which allows us to use a batch size of 4. The reconstruction is thus learned on downsampled volumes. When fine-tuning the model for lesion segmentation, we initialize the encoder and selected decoder blocks with the learned weights, and the remaining decoder blocks are trained from scratch.

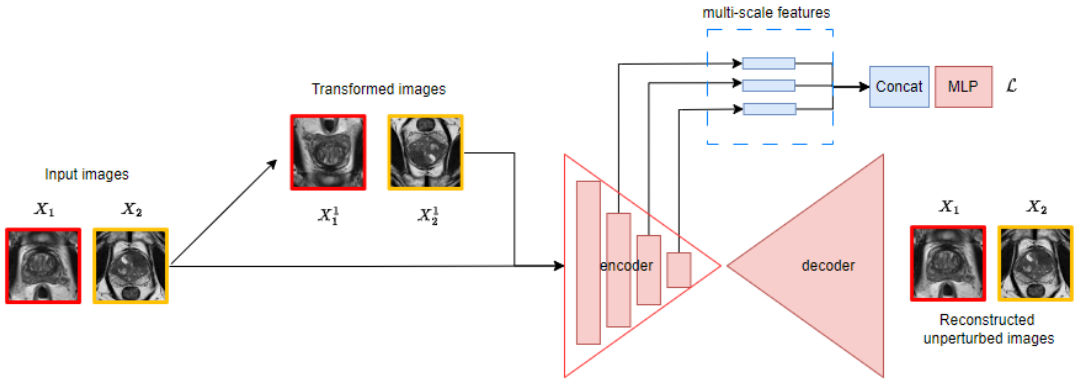


Figure 4.20: HSSL approach [Zheng et al., 2021].

Chaitanya et al. [2020] propose to combine global and local contrastive loss functions applied to partitions of the input volume. The input volume is split into a partition of four sub-volumes along the axial axis, one slice of each sub-volume is sampled at random to build the input batch. The input batch is perturbed with crop, brightness, and contrast transformations to create two perturbed versions which are fed to the encoder, followed by a projection head. The global contrastive loss function is defined to bring closer latent representations of the two perturbed versions of one sub-volume while pushing apart latent representations of different perturbed sub-volumes. The weights learned during this first pre-training stage are used to initialize the encoder for a second stage during which the decoder is pre-trained. During this stage, the input volume is split into sub-volumes, and the input batch is perturbed with brightness and contrast transformations and fed to the encoder-decoder architecture. A projection head is added after a set decoder block. Thirteen patches of fixed size are extracted from the obtained feature maps. A local contrastive loss function is defined to bring closer latent representations of the same patches in the feature maps obtained from perturbed versions of the same input volumes while pushing apart other patches. Figure 4.21 shows a schematic view of this approach.

The model used for this approach is a 2D U-Net taking 2D slices sampled from 3D volumes partition. After pre-training, we fine-tuned the model on 2D slices from the 10% and 1% annotated 3D volumes. The 2D slices are chosen such that they intersect the prostate and that there is a balance between the lesion and non-lesion slices.

In 2D, Chaitanya et al. [2023] build on the local contrastive loss approach by adding a segmentation branch to the contrastive learning framework. A 2D U-Net architecture is used, a contrastive and a segmentation blocks are added after the decoder model. The 2D U-Net followed by the segmentation block is trained for a few epochs on the available annotated data, the trained segmentation model is then used to generate pseudo labels for unannotated data. Joint training is then performed on unannotated and annotated data (see Figure 4.22):

- Annotated data are used to keep training the segmentation model;
- Both annotated and unannotated data are perturbed with random brightness and contrast transformations and fed to the contrastive model;

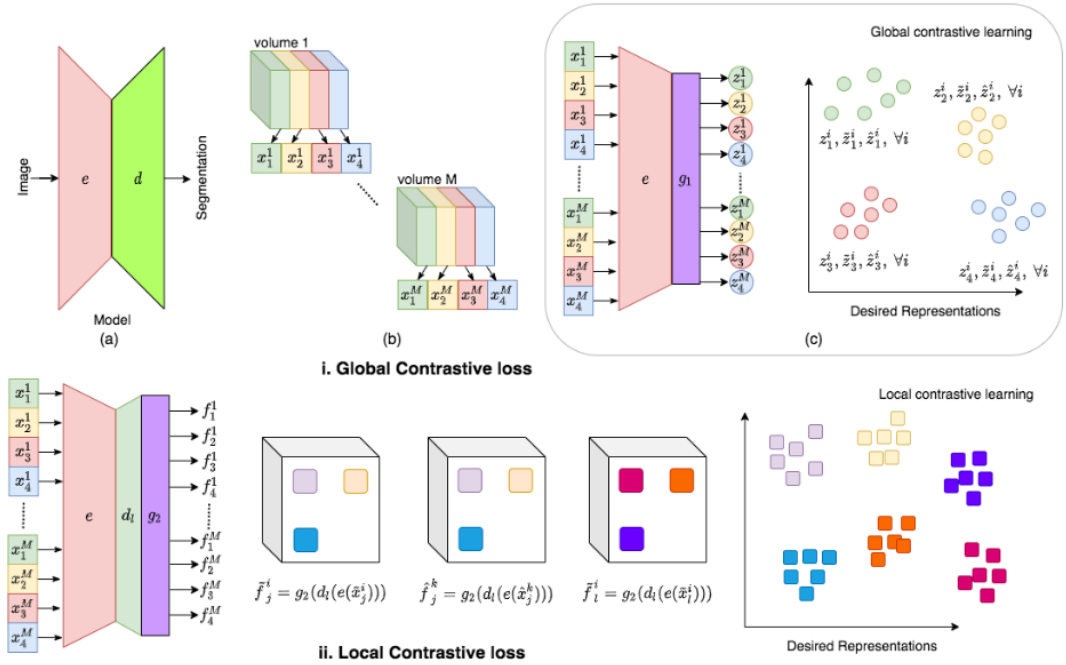


Figure 4.21: Local global contrastive pre-training approach (figure from [Chaitanya et al., 2020]).

- The obtained feature maps are then masked by either available reference segmentation or pseudo label;
- A local contrastive loss function is then applied to bring closer latent representations of pixels from the same class and push apart latent representations of pixels of different classes;
- Pseudo labels are updated during training.

The trained segmentation model is then evaluated on a hold-out test set.

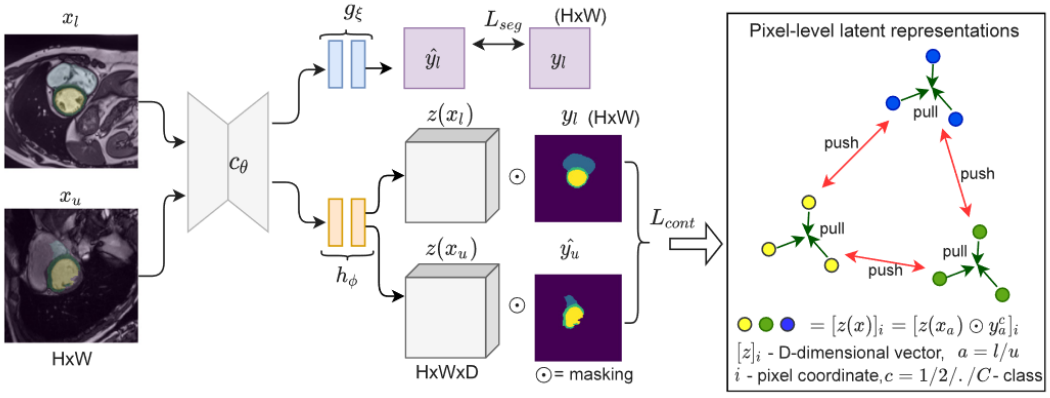


Figure 4.22: Pseudo-label local contrastive learning approach (figure taken from Chaitanya et al. [2023]).

The two aforementioned approaches have led to high performances at very low data regime. We compare the performances obtained by these two approaches to that of a 2D U-Net trained from scratch.

We also propose to apply the local contrastive loss function to a partition of 3D feature maps rather than randomly sampling slices from sub-volumes of 3D volumes to have a 3D model, thus taking advantage of previous experiments. We initialize the encoder with the weights learned by the confidence contrastive pre-training and train the first decoder level with the local loss function applied on patches from sub-volumes of feature maps. Figure 4.23 shows the feature maps partition approach, and patches on feature maps are then computed as in Figure 4.21.

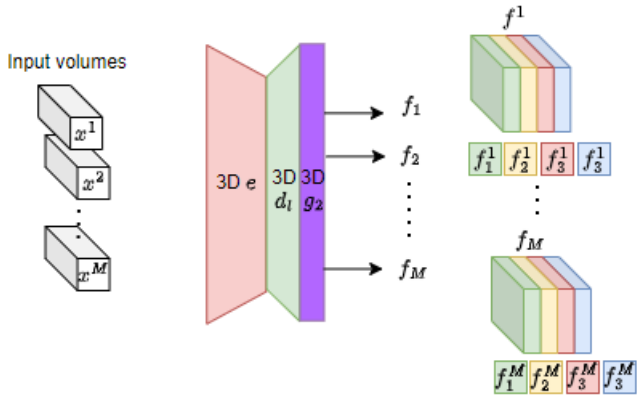


Figure 4.23: Feature maps partition approach.

As a preliminary work, we apply the contrastive loss function with confidence kernel presented in Section 4.5 at the decoder level: rather than applying the projection head after the encoder, we apply it after the first decoder layer. We also apply the uniform align contrastive loss function [Wang and Isola, 2020] after the first decoder layer, initializing the encoder with the weights from the confidence approach pre-training. Results of these approaches are shown in the second part of Tables 4.7 and 4.9. We see that applying base contrastive learning at the decoder level after pre-training the encoder with the confidence kernel does not improve performances compared to applying it to the decoder only.

We compare the contrastive approaches of the literature with Models Genesis [Zhou et al., 2019b] (which pre-trains a U-Net to reconstruct images transformed with perturbations specific to medical images, as introduced in Section 2.1).

Applying contrastive learning to the decoder, we are aiming to improve segmentation performances. To evaluate the impact of the tested methods, we report segmentation metrics: Dice Coefficient on every volume and Dice Coefficient on volumes having at least one lesion (true negative and false positive exams are not considered for this metric). Dice results on the 10% database are reported in Appendix C.

On the 1% database (results in Tables 4.7 and C.3), HSSL, Models Genesis and our confidence approach have similar performances. On the 10% database (results in Table 4.9), HSSL and Models Genesis approaches lead to small performance improvements

compared to our approach. We also train the HSSL architecture with our confidence contrastive loss function (row HSSL with w (4.7) in both Tables 4.7 and 4.9) which does not lead to performances improvement. Further works should focus on the balance between contrastive and reconstruction loss terms and adding contrastive learning in the decoder along with image reconstruction. Pre-training the full U-Net architecture rather than selecting some decoder blocks could also help improve performances but is bound to lead to memory issues. Parallel training and contrastive learning without negative pairs approaches could be investigated to get around this limitation.

Results on the 2D dataset are shown in the last section of Tables 4.7 and 4.9. The approach on feature maps partition presented in Figure 4.23 did not yield any results and was very unstable to optimize. One possible explanation is that instability is due to random sampling of slices from sub-volumes of feature maps.

Table 4.7: 5-fold cross-validation mean AUC and mAP after fine-tuning on PI-CAI and private datasets with 1% of annotated data and decoder contrastive pre-training (standard deviation in parentheses).

Method	AUC exam		AUC lesion		mAP	
	PI-CAI	Private	PI-CAI	Private	PI-CAI	Private
Random init	0.68 (0.06)	0.74 (0.03)	0.73 (0.11)	0.70 (0.05)	0.27 (0.05)	0.62 (0.03)
HSSL	0.67 (0.05)	**0.78** (0.03)	0.67 (0.09)	0.70 (0.05)	0.30 (0.05)	0.67 (0.04)
Models Genesis	0.68 (0.07)	0.77 (0.06)	0.63 (0.17)	0.69 (0.07)	**0.31** (0.07)	**0.68** (0.04)
HSSL with w (4.7)	0.67 (0.06)	0.75 (0.03)	0.67 (0.10)	0.70 (0.02)	0.28 (0.09)	0.66 (0.04)
Ours w (4.7)	**0.70** (0.05)	0.75 (0.03)	**0.75** (0.10)	0.71 (0.03)	0.30 (0.09)	0.63 (0.04)
Ours w (4.7) decoder	**0.70** (0.02)	0.75 (0.02)	0.73 (0.08)	**0.72** (0.04)	0.30 (0.05)	0.63 (0.02)
Decoder w (4.7) init	0.65 (0.08)	0.75 (0.03)	0.63 (0.14)	0.70 (0.05)	0.25 (0.07)	0.64 (0.04)
Random 2D init	0.59 (0.08)	0.73 (0.06)	0.57 (0.12)	0.64 (0.09)	0.20 (0.08)	0.56 (0.10)
Local global	0.44 (0.09)	0.59 (0.08)	0.56 (0.09)	0.58 (0.03)	0.09 (0.03)	0.36 (0.08)

Table 4.8: 5-fold cross validation Dice and Dice lesion after fine-tuning on PI-CAI and private datasets with 1% of annotated data and decoder contrastive pre-training (standard deviation in parentheses).

Method	Dice		Dice lesion	
	PI-CAI	Private	PI-CAI	Private
Random init	0.05 (0.01)	0.19 (0.01)	0.29 (0.05)	0.32 (0.01)
HSSL	**0.06** (0.01)	**0.22** (0.01)	0.32 (0.06)	**0.36** (0.02)
Models Genesis	0.05 (0.02)	0.21 (0.03)	0.29 (0.10)	0.34 (0.04)
HSSL with w (4.7)	0.04 (0.02)	0.18 (0.04)	0.25 (0.09)	0.30 (0.07)
Ours w (4.7)	0.06 (0.01)	0.21 (0.02)	**0.33** (0.06)	0.35 (0.03)
Ours w (4.7) decoder	0.05 (0.01)	0.19 (0.03)	0.29 (0.06)	0.32 (0.06)
Decoder w (4.7) init	0.04 (0.02)	0.17 (0.03)	0.22 (0.10)	0.28 (0.06)
Random 2D init	0.03 (0.01)	0.13 (0.03)	0.15 (0.05)	0.21 (0.05)
Local global	0.02 (0.01)	0.10 (0.03)	0.09 (0.04)	0.16 (0.05)

The local-global approach proposed by Chaitanya et al. [2020] did not lead to the same performance improvement on our dataset. We hypothesize that this is due to the size of the objects we want to segment. In the original paper, the pre-training approach is

Table 4.9: 5-fold cross-validation mean AUC and mAP after fine-tuning on PI-CAI and private datasets with 10% of annotated data and decoder contrastive pre-training (standard deviation in parentheses).

Method	AUC exam		AUC lesion		mAP	
	PI-CAI	Private	PI-CAI	Private	PI-CAI	Private
Random init	0.78 (0.04)	0.80 (0.02)	0.77 (0.06)	0.73 (0.02)	0.38 (0.04)	0.68 (0.02)
HSSL	0.82 (0.02)	**0.83** (0.01)	0.77 (0.04)	0.72 (0.02)	0.35 (0.03)	0.70 (0.03)
Models Genesis	**0.83** (0.01)	**0.83** (0.01)	0.81 (0.02)	**0.74** (0.03)	**0.41** (0.03)	**0.73** (0.02)
HSSL with w (4.7)	0.80 (0.02)	0.80 (0.01)	0.77 (0.04)	0.71 (0.00)	0.36 (0.03)	0.69 (0.02)
Ours w (4.7)	0.80 (0.02)	0.80 (0.02)	**0.82** (0.04)	**0.74** (0.02)	0.40 (0.04)	0.70 (0.01)
Ours w (4.7) decoder	0.79 (0.05)	0.81 (0.01)	0.74 (0.07)	0.72 (0.04)	0.38 (0.02)	0.69 (0.03)
Decoder w (4.7) init	0.78 (0.03)	0.79 (0.03)	0.77 (0.06)	0.75 (0.03)	0.35 (0.02)	0.68 (0.03)
Random 2D init	0.74 (0.03)	0.80 (0.02)	0.71 (0.02)	0.72 (0.02)	0.33 (0.04)	0.66 (0.03)
Local global	0.67 (0.03)	0.74 (0.02)	0.71 (0.02)	0.72 (0.03)	0.22 (0.03)	0.59 (0.02)

applied for whole organ segmentation: prostate or heart. When selecting patches from the feature maps for the local contrastive loss function computation, there is a high probability that it intersects the region of interest for segmentation (as the organ to segment is present in most of the volume slices). A prostate lesion being considered clinically significant above 0.5 cubic centimeters, it can be present on only two or three slices over the whole volume and thus absent from most of the selected feature maps patches.

Similar conclusions can be drawn from the pseudo label approach [Chaitanya et al., 2023]. In contrast to what is done in the original paper, our problem only has two classes: lesion and background. In 2D, this leads to a highly imbalanced problem as lesions are only present in some slices: to avoid this issue, we run our experiments on a small dataset of 17012 2D slices each containing a lesion. As this is preliminary work, no cross-validation was performed on this approach, results are presented on the hold-out test set of 500 cases previously used but pre-training has been done on only one fold.

Table 4.10 shows that the pseudo label approach does not lead to detection metrics improvement but Table 4.11 shows that segmentation performances are higher. Figure 4.24 shows the evolution of the generated pseudo labels on the cases considered as unannotated during pre-training. The last row shows the reference lesion segmentation. We see that during training pseudo labels tend to get closer to the reference segmentation.

Figure 4.25 shows cases where the model did not manage to learn relevant pseudo labels which mostly concern small lesions.

To improve this approach to segment small objects in a two-class setting, further works could focus on introducing a weighted contrastive loss function at the pixel level, considering prostate, lesion, and background pixels, and assigning a higher weight to the alignment of lesion pixels. Annotation variability could also be dealt with by choosing experts labeled cases for the labeled set and annotation confidence could also be introduced at the pixel level in the contrastive loss function.

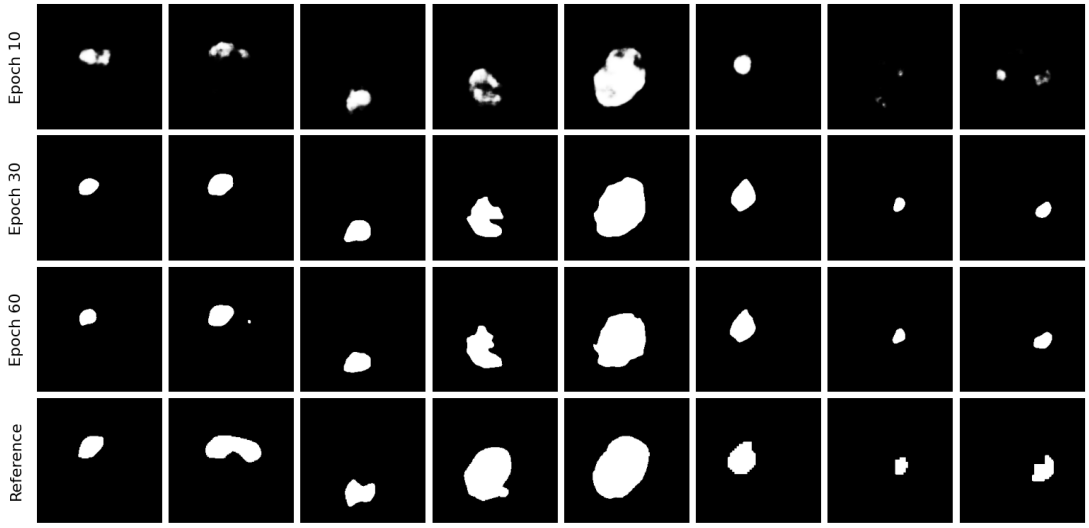


Figure 4.24: Examples of pseudo labels at different training epochs (10, 30, and 60, first three rows) and reference segmentation (last row). Each column shows a different patient slice.

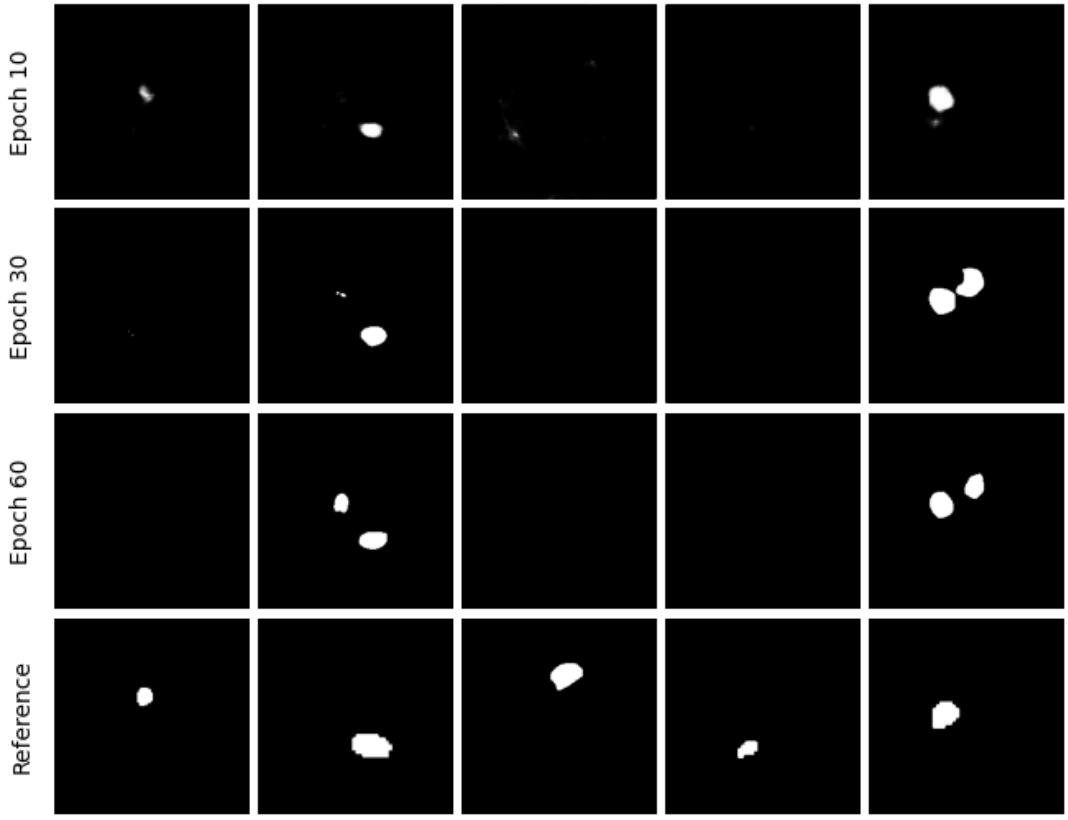


Figure 4.25: Examples of pseudo labels failure cases at different training epochs (10, 30, and 60, first three rows) and reference segmentation (last row). Each column shows a different patient slice.

Table 4.10: Comparison of the pseudo label approach with random initialization.

Method	AUC exam		AUC lesion		mAP	
	PI-CAI	Private	PI-CAI	Private	PI-CAI	Private
Random 2d init lesion	0.65	0.78	0.80	0.77	0.22	0.65
Pseudo label	0.52	0.60	0.25	0.35	0.07	0.24

Table 4.11: Comparison of dice metrics of the pseudo label approach with random initialization.

Method	Dice		Dice lesion	
	PI-CAI	Private	PI-CAI	Private
Random 2d init lesion	0.06	0.22	0.31	0.36
Pseudo label	**0.21**	**0.32**	0.23	**0.40**

4.7 Perturbations optimization for contrastive learning pre-training

We now apply the methods presented in Chapter 3 to the prostate lesion detection task. We optimize the perturbation generator and the encoder with some amount of supervision to classify prostate 2D slices as healthy or pathological. We thus build a 2D dataset by taking slices of 5510 3D volumes: a slice is selected if it intersects the prostate and such that there is a balance between lesion and nonlesion slices. A slice is considered pathological if it intersects a lesion segmentation mask, otherwise, it is defined as healthy. As in the experiments described in Chapter 3, we use 10% of supervision to guide the perturbation generator optimization. This leads to a dataset of 50618 2D slices (10% of which are considered annotated) among which 14968 are kept as a hold-out test set.

We apply a similar split and evaluation strategy as introduced in Section 3.3.3: data for the pre-training step is split into training and validation sets, and three optimizations are performed changing the 10% supervision set for variability analysis. Linear evaluation is performed on the hold-out test set. We perform linear evaluation using the weights obtained at different pre-training epochs. We train the encoder on the classification task in a fully supervised manner with 10% and 100% labeled data. We perform linear evaluation on the frozen encoder with the hold-out test set and report the obtained AUC as horizontal lines.

We see that the approach introduced in Chapter 3 translates well to the prostate lesion classification problem. Pre-training the encoder with 10% of supervision while optimization perturbations generates better latent representations than training from scratch with 10% of annotations.

Figure 4.27 shows the obtained perturbations after pre-training on three different sets of supervised data. We see that depending on the supervised sets the obtained perturbations are not the same. Sets 0 and 1 give relevant perturbations almost often including lesion pixels. The optimization with supervised set 2 is an example of the mode collapse of M: the perturbation generator did not manage to get out of this local

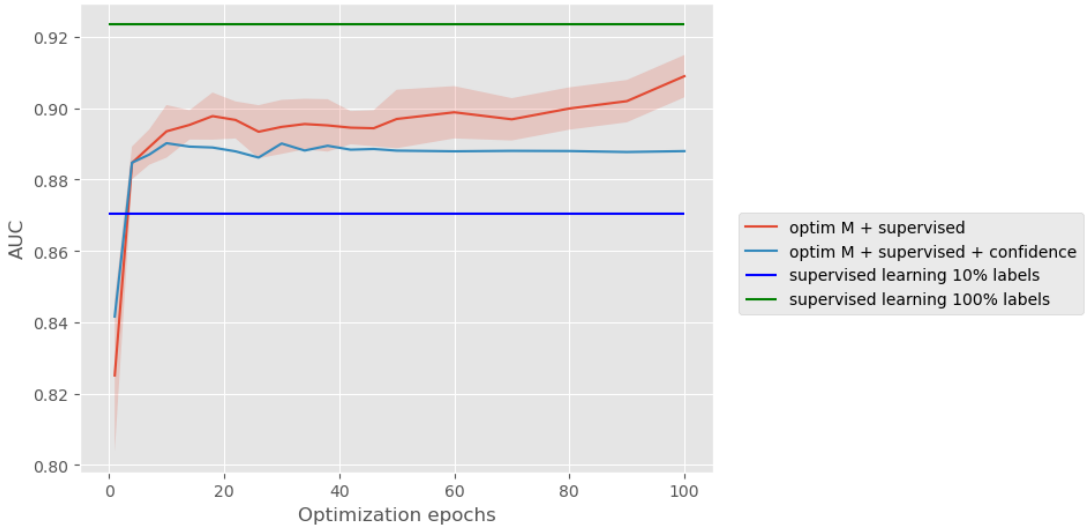


Figure 4.26: Linear evaluation results across optimization epochs.

minima but the encoder managed to get enough information from the unperturbed image to obtain sufficient classification performances.

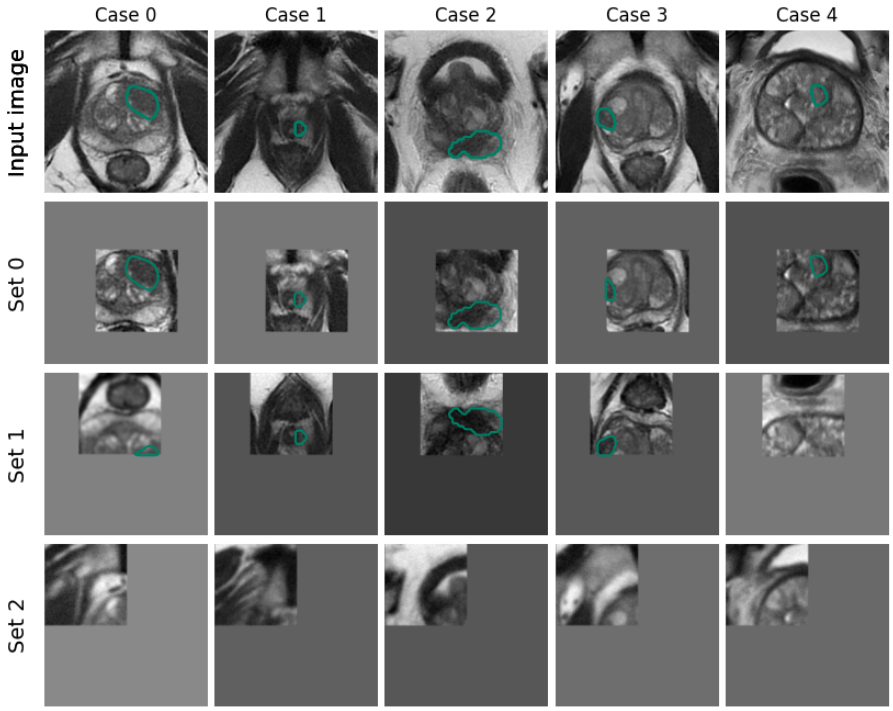


Figure 4.27: Examples of generated perturbations on five different cases after optimization on three different supervised sets. Reference lesion segmentation is displayed in green overlay.

Applying confidence conditional contrastive learning on perturbation generator optimization

We perform encoder and perturbation generator optimization using the confidence contrastive loss introduced in Section 4.5. As we are using the alignment and uniformity framework, as introduced in Section 3.3.4 we are only minimizing alignment for the perturbation generator. Figure 4.26 shows that introducing confidence in the perturbation optimization framework does not lead to better latent representation (as these are preliminary results, optimization on only one supervision set is performed hence the lack of standard deviation).

Further works should thus investigate how to improve the perturbation generator optimization when working in the alignment and uniformity framework with metadata confidence.

4.8 Conclusion

In this chapter, we have introduced metadata confidence based on agreement among annotators into contrastive learning for prostate cancer lesion detection. Fully supervised deep learning methods have already been developed to automate prostate cancer detection but they are based on large datasets where lesions are usually manually segmented and classified by experts. These annotations are costly to obtain while PI-RADS score at the exam level is readily available from radiology reports. However, these annotations are subject to high inter and intra-annotator variability [Turkbey et al., 2019]. Additionally, some examinations may be classified using a biopsy result which is often taken as ground truth but is much more costly to obtain. In this context, building a deep learning algorithm to automatically detect cancerous lesions on prostate MRI necessitates taking into account annotations variability and source (biopsy or radiological).

Following the conditional contrastive learning framework introduced by Dufumier et al. [2021a] who propose to weight alignment and uniformity based on metadata similarity, we introduce a kernel based on classification (PI-RADS or biopsy) confidence. We measure confidence as the agreement among annotators for each exam. We then define different kernels with increasing levels of complexity with respect to confidence inclusion. We pre-train the encoder of a 3D U-Net with this conditional contrastive learning framework. After pre-training, we fine-tuned the whole encoder-decoder architecture on the lesion detection task with 1% and 10% annotated data. With 1% annotated data we show substantial improvements with the most complex kernel using confidence compared to other contrastive learning approaches from the literature: on two datasets, we report an AUC increase between 3 and 4% at both exam and lesion levels and an mAP increase between 3 and 7%. With 10% annotated data, our proposed method performs similarly to noncontrastive approaches but still outperforms contrastive methods from the literature. Further work could focus on combining the confidence approach with noncontrastive methods. We also show that the proposed perturbation generator introduced in Chapter 3 leads to better latent representation quality than fully supervised learning with 10% annotated data on the prostate cancer classification task. The perturbations found are well centered on the tumor region of interest except for one supervision fold which underlines the need to avoid the perturbation generator mode collapse.

Nevertheless, it seems that structuring the encoder latent space only is not enough to keep a satisfactory class separation after fine-tuning: the structure learned by the encoder is transformed when fine-tuning the full architecture with multiple decoder layers. We present preliminary results applying self-supervised and contrastive pre-training on the decoder. The experiments show that adding self and contrastive pre-training to the decoder helps improve performances during fine-tuning. The local and pseudo-label approaches from the literature did not yield the expected improvements. Further works could thus investigate combining local pseudo-label approaches with annotation confidence, pre-training with more medically linked perturbations, and including other patients' metadata such as age or PSA level.

As developed in Section 4.6.1, further works could also focus on improving the confidence-based kernel using nearest neighbors in the latent space by working on support set design or including nearest neighbor distance in the kernel.

Chapter 5

Conclusions and Perspectives

5.1 Summary of the thesis contributions

The recent growth in the amount of available medical images has gone in hand with the development of deep learning algorithms for many diverse applications. The performances of a deep learning model are strongly linked to the quality and size of the labeled datasets available for training. However annotating medical data is costly, time-consuming, often requires expert knowledge, and can be subject to annotator variability.

In this thesis, we have proposed methods to reduce the annotation workload while keeping high algorithm performances in medical imaging. We have investigated self-supervised and contrastive learning methods which propose to take advantage of the available unannotated data during pre-training. These methods introduce auxiliary tasks that do not need supplementary labels.

In Chapter 3, we have introduced a new perturbation generator guided by some amount of supervision to optimize the perturbations used in contrastive pre-training. This generator is not based on generative models and is trained along the encoder. To build this generator, we have proposed a differentiable framework for perturbations. We have proven that contrastive pre-training with perturbation optimization and some amount of supervision leads to better latent representation quality than usual methods using perturbations sampled at random within a fixed list with fixed parameters. We have shown the performances of our approach at low annotated data regime on three different datasets. After optimization, we see that centered crop is the most relevant perturbation for contrastive pre-training and is often found to be centered on the region of interest.

This perturbation generator can still be improved to avoid mode collapse by adding noise during training for instance. Further works could also look at this generator through the augmentation lens to use the contrastive pre-training to find augmentations to be used in fully supervised training.

In Chapter 4, we have then investigated how to take metadata variability and confidence into account in contrastive pre-training applied to the specific problem of prostate cancer detection. For prostate cancer grading, different metadata can be available: PI-

RADS scores, ranging from 1 to 5 and readily available from radiology reports but subject to high inter and intra-annotator variability, and biopsy scores, more precise but also more costly to obtain and subject to bias as only patients with high PI-RADS scores undergo a biopsy. We have introduced a new kernel for conditional contrastive learning based on annotation confidence. We have proposed conditional contrastive kernels with increasing levels of complexity and defined confidence as agreement among annotators. We have shown that the confidence approach improves detection performances compared to other methods from the literature and to simpler proposed kernels.

We have presented preliminary works to improve the confidence contrastive learning kernel. To compensate for the heterogeneous number of annotators for each patient, we have proposed to use metadata from latent space nearest neighbors as pseudo labels for patients with only one metadata value. This approach can be improved by further investigating the choice of nearest neighbors, the structure of the latent space with highly confidence samples as a first training stage, and extending the pseudo label computation to patients with metadata of smaller confidence.

In Section 4.5.5, we have shown that pre-training the encoder with contrastive learning was not sufficient to obtain the highest performances on the segmentation task as fine-tuning an encoder-decoder architecture alters the structure learned by the encoder during pre-training. We have presented preliminary results to improve performance on the segmentation task. We have used the proposed confidence contrastive loss function at the decoder level showing the benefits of decoder pre-training. We have proposed to use existing approaches from the literature to add contrastive learning on the decoder through multi-scale, reconstruction, local contrastive loss functions, and pseudo-label approaches. The reconstruction approach proposed by Zhou et al. [2019b], introducing perturbations more specific to medical imaging, has led to the best performances.

The pseudo label approach proposed by Chaitanya et al. [2023] has yielded Dice score and qualitative segmentation improvements although it did not improve AUC or mean average precision. This approach has to be further investigated to better define the local loss function for small object segmentation. Annotation confidence could also be added to this pre-training by choosing the most confident cases (e.g. annotated by experts) for the first segmentation stage and introducing a confidence kernel for the local contrastive loss.

5.2 Perspectives

5.2.1 Combining perturbation generator and confidence approach

The two main contributions of this thesis, namely the perturbation generator and the confidence kernel, could be further investigated in joint decoder pre-training with local loss functions and noncontrastive approaches. The perturbation generator introduced in Chapter 3 could be used in a decoder pre-training approach for segmentation tasks. We showed in Chapter 4 that noncontrastive approaches such as BYOL [Grill et al., 2020], simSiam [Chen and He, 2020] and Barlow Twins [Zbontar et al., 2021] were performing well on encoder pre-training, and further improvements could be obtained by combining these approaches with the confidence kernel.

5.2.2 Improving the perturbation generator

In Chapter 3, we showed that optimizing a perturbation generator for contrastive pre-training was beneficial as some perturbations were more relevant to the subsequent supervised task.

In line with the more medically oriented perturbations proposed by Zhou et al. [2019b], further research could focus on generating and optimizing more realistic perturbations such as metal artifacts, organ size, or image acquisition noise to increase fine-tuned model robustness.

Anatomical invariances such as organ size could be learned using simple transformations such as zooming in or out to simulate size differences. More advanced methods such as spatial transformer networks could be used to learn a generator of organs of different sizes.

Metal artifacts (when present in images) and image acquisition noise could be generated using diffusion models trained to learn and generate the noise distribution in the dataset. Perturbations could also be learned directly in the MRI k-space to generate MRI sequences from a perturbed k-space.

5.2.3 Building on multi-modal conditional contrastive pre-training with confidence

In Chapter 4, we showed that including metadata confidence in contrastive pre-training was beneficial to improve performances. Given that pre-training the decoder is most beneficial to segmentation performances, further works could investigate how to include some amount of supervision with segmentation annotations while taking their variability into account. Confidence in segmentations could be defined considering the expertise of the annotator, when known, or also defined based on the amount of agreement at the pixel level.

Many metadata are available from radiological reports, in Chapter 4 we have focused on pathology score at the exam level. Following the research on multi-modality combining text and images done on natural images and the work proposed by Bosma et al. [2021], future works should combine image contrastive latent space with as much information from reports as possible: age, sex, pathology, clinical data... The information extracted from reports could be used to introduce a more general knowledge of pathology and the patient's potential anteriorities.

Contrastive learning also seems a great candidate to include knowledge of indirect signs between different pathologies. For instance, a knee medial meniscus tear is often associated with an anterior cruciate ligament lesion. This a-priori medical knowledge is used in practice by radiologists when providing their diagnosis and should be included in contrastive pre-training.

Some recent works on prostate cancer lesion detection [Schelb et al., 2020, Duran et al., 2022] have shown great performances using ensemble approaches: training multiple models in parallel and averaging their predictions. Contrastive pre-training could be applied to these approaches to pre-train the different models and to see the performance gain while reducing the number of models used.

5.2.4 Contrastive pre-training in a concrete medical and industrial setting

Finally, the contributions made by this thesis are part of a specific medical and industrial context in which annotations are costly to obtain but, when building a partnership with a hospital, a large amount of unannotated medical data and patient metadata can be available. We showed that this low annotated data regime framework benefited from contrastive pre-training with optimized perturbation and metadata inclusion.

Having a performing framework to pre-train models on unannotated data allows to use images from different clinical sites and increases robustness to different MRI machines for deployment without needing to mobilize doctors to annotate. These pre-training approaches could also be included in an active learning framework where a model would be pre-trained on unannotated data and fine-tuned with the small initial amount of data available. In the development phase, when doctors are asked to annotate an increasing amount of images, the pre-trained model could be fine-tuned on these new annotations to find the most accurate and sufficient cases to annotate to obtain good performances.

Appendix A

Appendix: Latent space clustering

A.1 Perturbations clustering

Generation (see Section 2.1) and context based (see Section 2.2) pretext tasks introduce different perturbations applied to the input image which will then be processed by a neural network. Finding optimal pretext tasks, hence perturbations, for self-supervised learning implies finding an importance order between them. We thus tried to create perturbations clusters. The rationale is that perturbations with similar parameters shall be closer than ones with different parameters. Looking at one perturbation with different parameter values, latent representations should be ordered as the parameter values. For example, if we are looking at three additive Gaussian noises with different σ values, if $\sigma_1 < \sigma_2 < \sigma_3$, we expect latent representation distances to be similarly ordered. We started by investigating perturbations used in contrastive learning [Chaitanya et al., 2023, Chen et al., 2020a, Dufumier et al., 2021b]: Gaussian blur, Gaussian noise and flip. Once a perturbations clustering is build, it should be possible to move within the perturbations clustered space in order to find the most relevant ones for pre-training. Similar to contrastive learning approaches (presented in details in Section 3.2), we build loss functions in the latent space encoding perturbation proximity. We train an encoder to bring closer representations in the latent space of similar perturbations.

We propose to model the perturbations proximity in the latent space with a parameter p encoding how similar two perturbations are.

The parameter p can be defined as $1 - \Delta\alpha$ where $\Delta\alpha$ is the absolute value of the difference between the shared parameters of two perturbations of the same set (for instance $|\sigma_1 - \sigma_2|$ for two blurs with standard deviations σ_1 and σ_2).

Considering three perturbations T_1, T_2 and T_3 where T_1 and T_2 share a parameter (for instance two blurs with different σ values), the proximity matrix between tasks is the following:

$$\begin{matrix} 1 & p & 0 \\ p & 1 & 0 \\ 0 & 0 & 1 \end{matrix}$$

We constraint two latent projections z_i and z_j to be normalized and such that: $\|z_i - z_j\|^2 = 4(1 - p_{ij})$ where p_{ij} is an element of the proximity matrix defined above (row i column j). Note that using a simpler approach where $\|z_1 - z_2\| = (\frac{1}{p} - 1)$ would not work if $p = 0$.

This leads to the following optimization problem:

$$\begin{aligned} \min \quad & z_i \cdot z_j - (2p_{ij} - 1) \\ \text{s.t.} \quad & \|z_i\| = 1 \end{aligned} \tag{A.1}$$

The encoder generating latent representations z_i, z_j is optimized through gradient descent to minimize Equation (A.1). To evaluate the performance of proximity training, we compute a 3D UMAP [McInnes and Healy, 2018] projection of latent representations of the validation set before and after training. The right side of Figure A.1 shows that a decent clustering is obtained with two perturbations. However we see that a separation between the two perturbations already exists without needing to train a model. We obtain this clustering when computing a forward pass on a randomly initialized model as shown on the left side of Figure A.1.

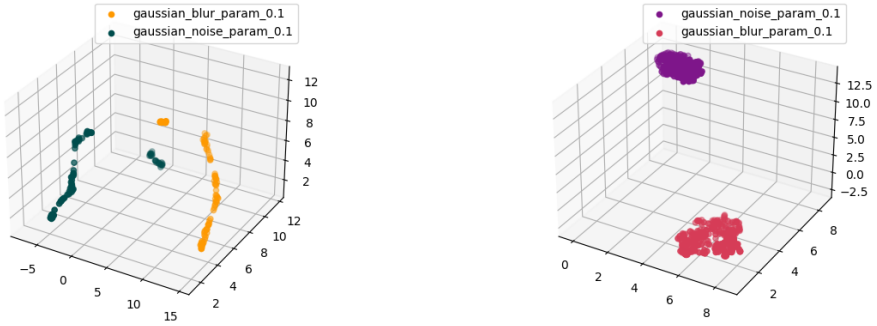


Figure A.1: UMAP projection of latent representations before (left) and after (right) training (the three axis are the three UMAP components).

To get rid of existing a priori separation between perturbations, we performed pre-training with two supplementary gaussian blur and noise and also included the unperturbed input image in pre-training. Figure A.2 shows that latent representations of non perturbed images clusters are close to that of additive noise.

To help the model learning meaningful representations of the data, while learning perturbation clustering in the latent space, we couple clustering with reconstruction learning. Using a U-Net [Ronneberger et al., 2015] to reconstruct the input image and extracting the encoder latent representation with a multi-layer perceptron (MLP) as in simCLR [Chen et al., 2020a], the optimization problem is the following:

$$\begin{aligned} \min \quad & z_i \cdot z_j - (2p_{ij} - 1) + MSE(x_i, f(x_i)) \\ \text{s.t.} \quad & \|z_i\| = 1 \end{aligned} \tag{A.2}$$

where x_i is one image of the batch being compared to one other image x_j, z_i, z_j are the projections in the latent space of the encoder output for x_i and x_j, $f(x_i)$ is the reconstruction of x_i generated by the U-Net, and MSE is the mearn square error

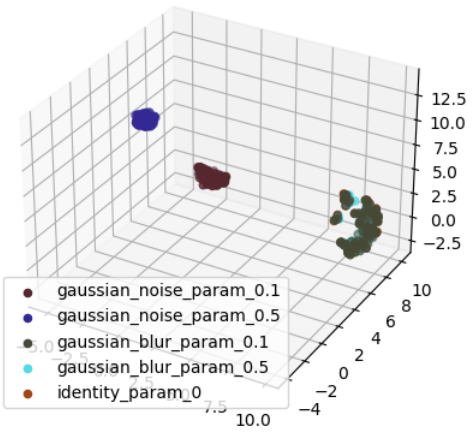


Figure A.2: Results on the validation set with more perturbations: two Gaussian noises, two Gaussian blurs and the unperturbed input image (identity_param_0 on the plot).

function.

Adding reconstruction in training does not help going over a priori distances between clusters. Due to memory constraints and model size, we had to use a smaller batch size, which is detrimental to contrastive loss functions that rely on a large number of negative samples.

More experiments were carried out to learn Gaussian blur parameter order and clustering by introducing more perturbations, but without successful results.

A.2 Linear combination of perturbations coupled with reconstruction

Once clustering is learned, we propose to move inside the learned latent space to find the most relevant perturbations for pre-training. Our hypothesis is that clustering can be used to find an optimal combination of perturbations in the latent space. We optimize a linear combination of perturbations to apply to the data at the input image and latent space levels. Using the clustered perturbations, we sample a linear combination of the latent representations of perturbations and feed them to the segmentation decoder. When learning linear combination of the latent space representations of perturbations in 3D, the memory cost increased drastically as the number of perturbations increased. Clustering and reconstruction were learned using five additive noises, crop, inpainting and identity perturbations. Figure A.3 shows satisfactory clustering results as inpainting and identity are clustered together and additive noises are logically ordered.

Despite clustering satisfactory results, the learned weights did not allow us to fine-tune the network on the segmentation task. Also, the benefit of clustering learning for the subsequent supervised task remains unclear.

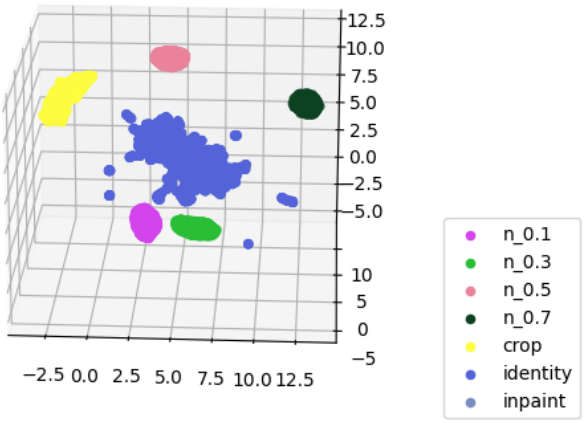


Figure A.3: Clustering results of latent representations of four additive noises, crop and inpainting perturbations,"identity" is the latent representation of unperturbed images.

To assess the performance of clustering pre-training we fine-tuned a U-Net for tumor segmentation using encoder learned weights. We compared the tested approaches to existing encoder pre-training methods in contrastive learning: simCLR [Chen et al., 2020a] and Models Genesis [Zhou et al., 2019a], as well as fully supervised training with augmentations. simCLR pre-training performed much better than any other method. We thus abandoned the clustering approach to learn a perturbation generator to be optimized for simCLR pre-training.

Appendix B

Appendix: Conditioned encoder and GAN training

B.1 Conditioned encoder training

To learn the optimal perturbations for self-supervised pre-training we propose to train a conditional perturbation generator. The proposed architecture is shown in Figure B.1.

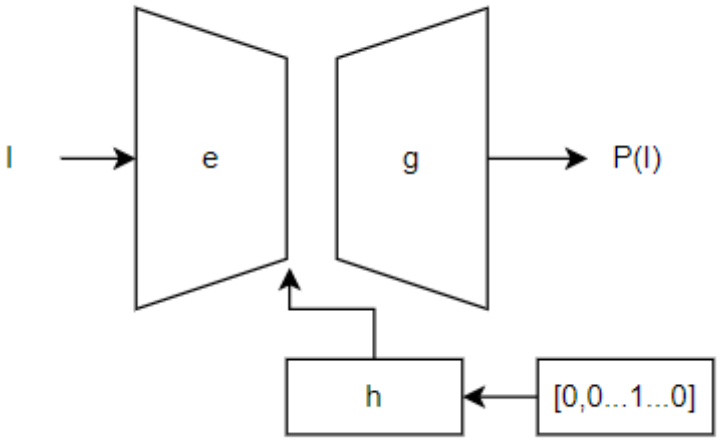
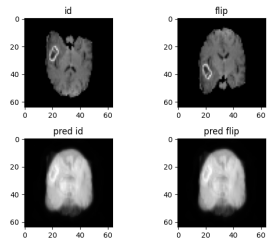
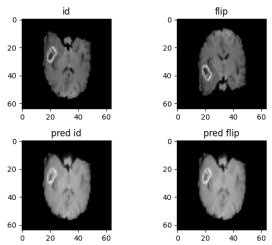
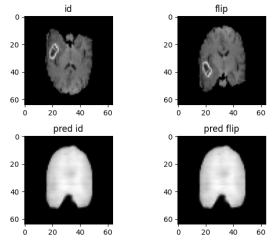
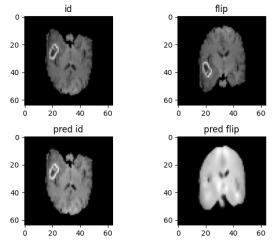


Figure B.1: First perturbation generator architecture.

An encoder e encodes the content of the input image I. A vector defining the perturbation to generate is embedded through h, taking as input a binary vector indicating which perturbation or perturbation composition to apply. The perturbed image is generated by g taking I and h output as inputs. For h both Keras embedding layers and a fully convolutional network were tested. The generator was also trained with and without skip connections.

Table B.1 sums up the different experiments carried out and their results on 3D volumes of the BraTs dataset. For the first experiments, we trained the network to generate images perturbed by horizontal and vertical flips and centered crop of fixed size.

Table B.1: Experiments results of perturbation generation without GAN.

Method name	Observation	
Id + flip, l2	not learning	
Id + flip, l1	not learning	
Id + flip, without skip connections	not learning lacking details	
Id + flip, instance normalization	better learning, lacking details	

When using batch normalization (BN), batch normalization weights were not encapsulating perturbations properly, as shown in the right side of Figure B.2. Because every image in a batch can be perturbed with a different perturbation, computing the normalization weights over the whole batch cannot encapsulate individual perturbations. Figure B.2 (left) shows that batch normalization works when recomputing its weights at inference. We thus replaced it by instance normalization. The flip was better learned and the model distinguished better identity from flip.

To increase the amount of training data we moved from 3D volumes to 2D slices. Volumes were split along the axial axis to get 2D slices, and only slices with less than 80% of black pixels were kept.

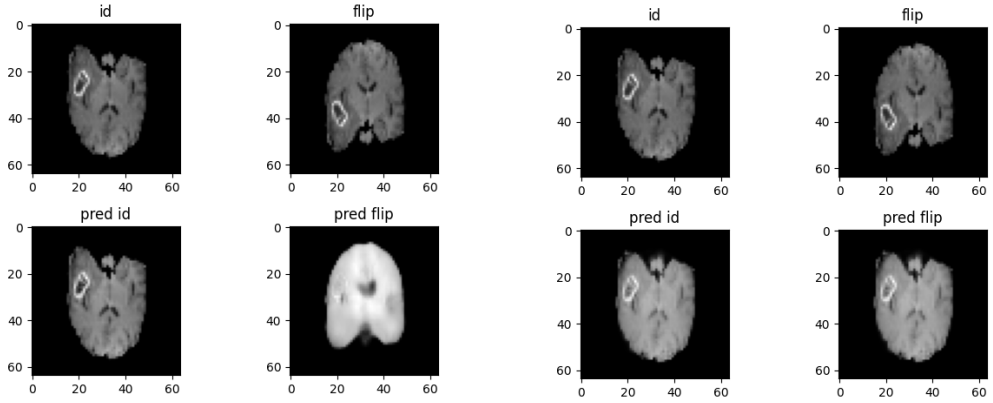


Figure B.2: Examples of generated images with different batch normalization strategies : recomputing BN weights at prediction (left), using learned BN weights during prediction (right)

On the 2D database, we changed the generator depth and the embedding architecture. Playing on the embedding architecture did not change the results much. Figures B.3 shows that a deeper network generates better results.

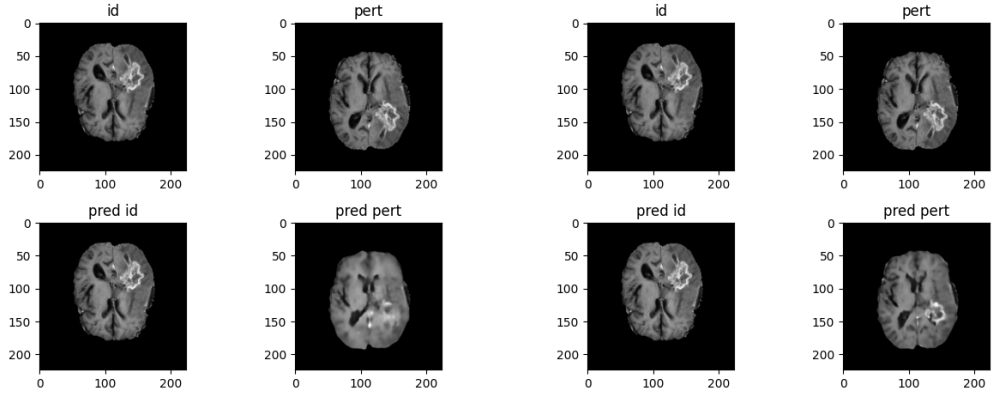


Figure B.3: Generated images examples with network of depth 4 (left) and 5 (right)

A simple auto-encoder was trained to learn identity and flip generation. To optimize perturbations for self-supervised pre-training, we need to be able to generate multiple perturbations and compositions thereof. We thus trained an auto-encoder to learn crop and flip perturbations. Figure B.4 shows that when training the auto-encoder to learn perturbations composition the generated flip is lacking details. The image is much smoother and high frequency details are lost. Figure B.5 shows that the crop perturbation is properly generated.

B.2 GAN training

To improve the perturbation generator performances, we added a discriminator after image generation. Figure B.6 shows that introducing a discriminator increased generated images quality when learning the composition of flip and crop perturbations.

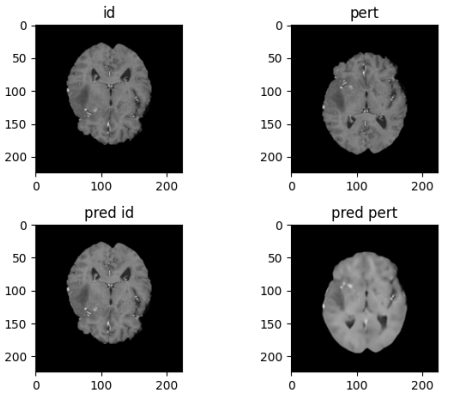


Figure B.4: Autoencoder flip generation

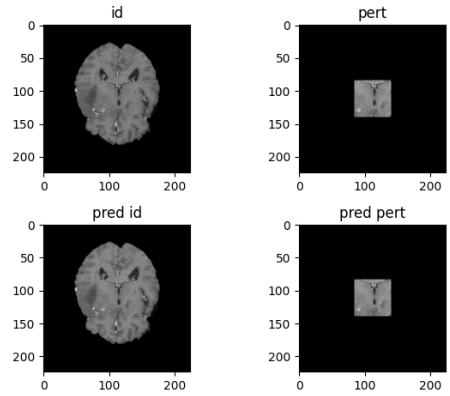


Figure B.5: Autoencoder crop generation

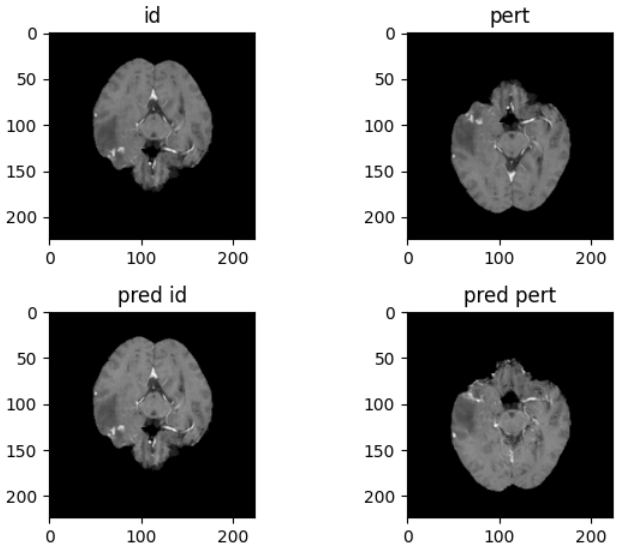


Figure B.6: Flip generation with GAN.

B.2.1 Perturbation regularization

For the subsequent optimization of the perturbation generator we need to be able to sample a perturbation conditioning among those learned in perturbation generator training. To that end, every perturbation has to be regularized as a function of a parameter $\lambda \in [0, 1]$. Perturbations such as blur and Gaussian noise are already dependent of a continuous parameter.

The crop perturbation has been regularized as follows: a crop perturbation is defined by two λ values, one for the center position of the crop as a proportion of the full image size and another one for the crop size as a proportion of the full image size.

The flip has been regularized using a projection formulation. The flip perturbation is a function of a λ value indicating the "amount" of flip performed and the axis a of the flip: $Flip_{\lambda,a}(x) = x - 2\lambda\langle x, a\rangle a$. When $\lambda = 0$ no perturbation is applied, when $\lambda = 1$ we obtain the full flip around the specified axis. If $\lambda \in]0, 1[$ the resulting image is as shown in Figure B.7.

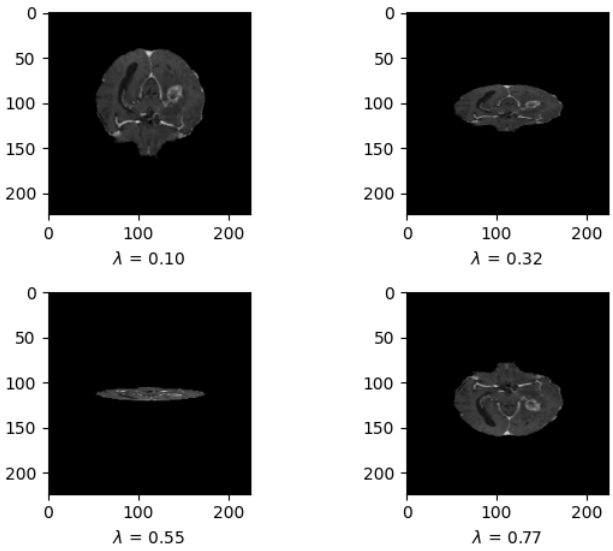


Figure B.7: Impact of λ value on flip output.

Training a GAN with such a perturbation formulation did not yield satisfactory results. Figure B.8 shows that when trying to generate only a flipped image the generator outputs a cropped image. Figure B.9 shows that a single crop is not properly generated. Training the network to generate perturbations composition was detrimental to single perturbation generation.

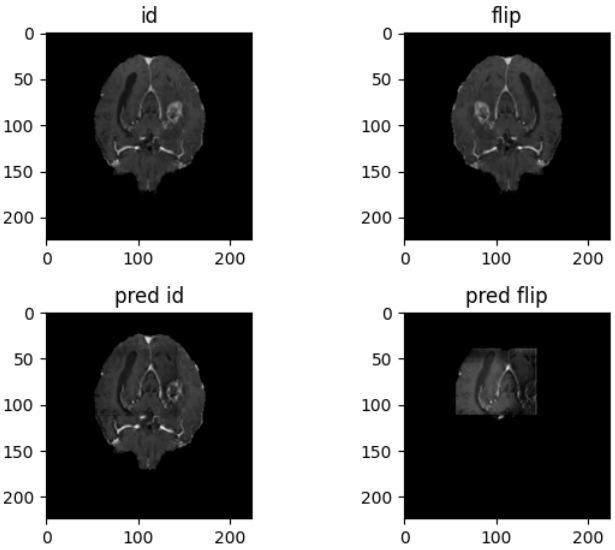


Figure B.8: Example of generated flip from regularized GAN trained to generate flip, crop and perturbation composition.

To avoid the burden of training a generative model for perturbation generation we propose to train a perturbation function generator rather than a perturbed image generator as developed in Section 3.3.

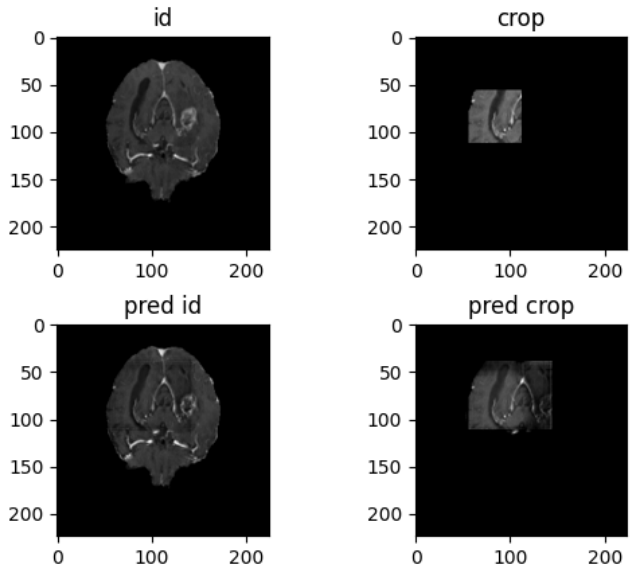


Figure B.9: Example of generated crop from regularized GAN trained to generate flip, crop and perturbation composition.

Appendix C

Appendix: Supplementary results for Chapter 4

Table C.1: Results with 10% annotated data: 5-fold cross validation mean AUC and mAP after fine-tuning on PI-CAI and private datasets.

Method	AUC exam		AUC lesion		mAP	
	PI-CAI	Private	PI-CAI	Private	PI-CAI	Private
Random init	0.78 (0.04)	0.80 (0.02)	0.77 (0.06)	0.73 (0.02)	0.38 (0.04)	0.68 (0.02)
simCLR	0.77 (0.04)	0.80 (0.03)	0.71 (0.03)	0.70 (0.02)	0.36 (0.03)	0.67 (0.02)
Unif Align	0.79 (0.05)	0.79 (0.01)	0.76 (0.04)	0.73 (0.01)	0.39 (0.05)	0.69 (0.02)
MoCo v2	0.77 (0.01)	0.79 (0.01)	0.74 (0.05)	0.72 (0.02)	0.36 (0.05)	0.68 (0.01)
nnCLR	0.78 (0.06)	0.81 (0.03)	0.76 (0.07)	0.73 (0.02)	0.39 (0.03)	0.67 (0.04)
BYOL	0.81 (0.03)	0.80 (0.01)	0.78 (0.04)	0.72 (0.03)	0.37 (0.02)	0.67 (0.02)
Barlow Twins	**0.82** (0.03)	0.80 (0.02)	**0.83** (0.03)	**0.76** (0.02)	**0.44** (0.03)	0.69 (0.02)
simSiam	**0.82** (0.01)	0.81 (0.02)	0.82 (0.02)	0.73 (0.03)	0.41 (0.04)	**0.71** (0.01)
w_δ (4.5)	0.80 (0.02)	0.78 (0.03)	0.75 (0.03)	0.72 (0.02)	0.40 (0.03)	0.67 (0.02)
GlU w_δ	0.79 (0.06)	**0.82** (0.02)	0.76 (0.06)	0.74 (0.03)	0.39 (0.03)	0.70 (0.01)
Majority voting	0.75 (0.04)	0.79 (0.03)	0.75 (0.05)	0.74 (0.02)	0.38 (0.05)	0.67 (0.03)
w_H (4.6)	0.79 (0.03)	0.80 (0.00)	0.76 (0.04)	0.71 (0.05)	0.41 (0.02)	0.67 (0.01)
GlU w_H	0.79 (0.02)	0.79 (0.02)	0.76 (0.04)	0.73 (0.02)	0.40 (0.05)	0.69 (0.02)
Biopsy	0.77 (0.02)	0.80 (0.02)	0.75 (0.06)	0.73 (0.03)	0.36 (0.02)	0.70 (0.03)
GlU w (4.7)	0.78 (0.02)	0.81 (0.01)	0.74 (0.03)	0.72 (0.02)	0.39 (0.04)	0.70 (0.00)
Ours w (4.7)	0.80 (0.02)	0.80 (0.02)	0.82 (0.04)	0.74 (0.02)	0.40 (0.04)	0.70 (0.01)

Table C.2: 5-fold cross validation mean AUC and mAP after fine-tuning on PI-CAI and private datasets with 10% of annotated data with nearest neighbors approach (standard deviation in parentheses)

Method	AUC exam		AUC lesion		mAP	
	PI-CAI	Private	PI-CAI	Private	PI-CAI	Private
Random init	0.78 (0.04)	0.80 (0.02)	0.77 (0.06)	0.73 (0.02)	0.38 (0.04)	0.68 (0.02)
MoCo v2	0.77 (0.01)	0.79 (0.01)	0.74 (0.05)	0.72 (0.02)	0.36 (0.05)	0.68 (0.01)
MoCo v2 confident	0.76 (0.03)	0.80 (0.01)	0.76 (0.02)	0.72 (0.02)	0.39 (0.04)	0.69 (0.02)
nnCLR	0.78 (0.06)	0.81 (0.03)	0.76 (0.07)	0.73 (0.02)	0.39 (0.03)	0.67 (0.04)
nnCLR confident	0.76 (0.07)	0.79 (0.01)	0.76 (0.08)	0.75 (0.03)	0.38 (0.03)	0.69 (0.01)
Ours w (4.7)	**0.80** (0.02)	0.80 (0.02)	**0.82** (0.04)	**0.74** (0.02)	**0.40** (0.04)	**0.70** (0.01)
Ours w (4.7) with nn	0.79 (0.04)	**0.81** (0.03)	0.74 (0.07)	0.73 (0.03)	0.37 (0.02)	0.69 (0.03)

Table C.3: 5-fold cross validation Dice and Dice lesion after fine-tuning on PI-CAI and private datasets with 10% of annotated data and decoder contrastive pre-training (standard deviation in parentheses).

Method	Dice		Dice lesion	
	PI-CAI	Private	PI-CAI	Private
Random init	0.07 (0.00)	0.24 (0.01)	0.39 (0.02)	0.40 (0.02)
HSSL	0.07 (0.00)	0.25 (0.01)	0.39 (0.02)	0.42 (0.01)
Models Genesis	**0.08** (0.00)	0.26 (0.01)	0.42 (0.01)	0.43 (0.01)
HSSL with w (4.7)	0.07 (0.00)	0.25 (0.00)	0.40 (0.02)	0.42 (0.01)
Ours w (4.7)	**0.08** (0.00)	**0.27** (0.01)	**0.44** (0.02)	**0.44** (0.02)
Ours w (4.7) decoder	0.07 (0.01)	0.24 (0.02)	0.38 (0.05)	0.40 (0.03)
Decoder w (4.7) init	0.07 (0.01)	0.25 (0.01)	0.40 (0.04)	0.41 (0.01)
Random 2D init	0.06 (0.00)	0.24 (0.01)	0.35 (0.01)	0.39 (0.02)
Local global	0.06 (0.00)	0.21 (0.01)	0.33 (0.02)	0.35 (0.01)

Bibliography

W. Bai, C. Chen, G. Tarroni, J. Duan, F. Guitton, S. E. Petersen, Y. Guo, P. M. Matthews, and D. Rueckert. Self-supervised learning for cardiac mr image segmentation by anatomical position prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2019. 10

Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *International Conference on Machine Learning (ICML)*, 2009. 18

D. Berthelot, N. Carlini, I. J. Goodfellow, N. Papernot, A. Oliver, and C. Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 18

I. Bhattacharya, A. Seetharaman, C. A. Kunder, W. Shao, L. C. Chen, S. J. C. Soerensen, J. B. Wang, N. C. Teslovich, R. E. Fan, P. Ghanouni, J. D. Brooks, G. A. Sonn, and M. Rusu. Selective identification and localization of indolent and aggressive prostate cancers via corrsignia: an MRI-pathology correlation and deep learning framework. *Medical image analysis*, 75:102288, 2021. 40, 44

I. Bhattacharya, Y. S. Khandwala, S. Vesal, W. Shao, Q. Yang, S. J. C. Soerensen, R. E. Fan, P. Ghanouni, C. A. Kunder, J. D. Brooks, Y. Hu, M. Rusu, and G. A. Sonn. A review of artificial intelligence in prostate cancer detection on imaging. *Therapeutic Advances in Urology*, 14, 2022. 44

M. Blendowski, H. Nickisch, and M. P. Heinrich. How to learn from unlabeled volume data: Self-supervised 3d context feature learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2019. 10

J. S. Bosma, A. Saha, M. Hosseinzadeh, I. Slootweg, M. de Rooij, and H. J. Huisman. Semisupervised learning with report-guided pseudo labels for deep learning–based prostate cancer detection using biparametric mri. *Radiology: Artificial Intelligence*, 5, 2021. 81

J. S. Bosma, A. Saha, M. Hosseinzadeh, I. Slootweg, M. de Rooij, and H. Huisman. Semisupervised learning with report-guided pseudo labels for deep learning–based prostate cancer detection using biparametric MRI. *Radiology: Artificial Intelligence*, 5(5), 2023. 59

M. Bošnjak, P. H. Richemond, N. Tomasev, F. Strub, J. Walker, F. Hill, L. Buesing, R. Pascanu, C. Blundell, and J. Mitrovic. Semppl: Predicting pseudo-labels for better contrastive representations. *International Conference on Learning Representations (ICLR)*, 2023. 63

B. Bozorgtabar, D. Mahapatra, G. Vray, and J.-P. Thiran. Salad: Self-supervised aggregation learning for anomaly detection on x-rays. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2020. 16

F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68, 2018. 4, 41

P. Bridge, A. L. Fielding, P. Rowntree, and A. P. Pullar. Intraobserver variability: Should we worry? *Journal of medical imaging and radiation sciences*, 47 3:217–220, 2016. 17

M. J. Cardoso, K. K. Leung, M. Modat, S. Keihaninejad, D. M. Cash, J. Barnes, N. C. Fox, and S. Ourselin. Steps: Similarity and truth estimation for propagated segmentations and its application to hippocampal segmentation and brain parcelation. *Medical image analysis*, 17 6:671–84, 2013. 18, 48

F. M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi. Domain generalization by solving jigsaw puzzles. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2224–2233, 2019. 11

M. Caron, H. Touvron, I. Misra, H. J'egou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021. 17

J. M. Castillo, M. Arif, W. J. Niessen, I. G. Schoots, and J. F. Veenland. Automated classification of significant prostate cancer on MRI: A systematic review on the performance of machine learning applications. *Cancers*, 12, 2020. 47

K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 12546–12558, 2020. xii, 16, 26, 67, 69, 71

K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu. Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation. *Medical image analysis*, 87:102792, 2023. xii, 16, 63, 67, 68, 69, 72, 80, 83

G. Chartrand, P. M. Cheng, E. Vorontsov, M. Drozdzal, S. Turcotte, C. J. Pal, S. Kadoury, and A. Tang. Deep learning: A primer for radiologists. *RadioGraphics*, 37(7):2113–2131, 2017. 2

L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert. Self-supervised learning for medical image analysis using image context restoration. *Medical image analysis*, 58:101539, 2019. 8

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *37th International Conference on Machine Learning (ICML)*, volume 119, pages 1597–1607, 2020a. xi, 12, 23, 25, 26, 28, 30, 31, 32, 34, 45, 52, 53, 59, 60, 83, 84, 86

X. Chen and K. He. Exploring simple siamese representation learning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15745–15753, 2020. 17, 60, 80

X. Chen, H. Fan, R. B. Girshick, and K. He. Improved baselines with momentum contrastive learning. *ArXiv*, 2020b. 60, 65

H. Cho, J. Seol, and S. goo Lee. Masked contrastive learning for anomaly detection. *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2021. 13

M. Ciortan, R. Dupuis, and T. Peel. A framework using contrastive learning for classification with noisy labels. *International Conference on Data Technologies and Applications*, 6:61, 2021. 19

J. P. Cohen, M. Luck, and S. Honari. Distribution matching losses can hallucinate features in medical image translation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2018. 24

P. Colin. Image-guided focal laser therapies for prostate cancer. 09 2012. x, 41

E. D. Cubuk, B. Zoph, D. Mané, V. Vasudevan, and Q. V. Le. Autoaugment: Learning augmentation strategies from data. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 113–123, 2019. 22

E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le. Randaugment: Practical automated data augmentation with a reduced search space. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3008–3017, 2020. 22

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. ix, 1, 3, 45

T. Devries and G. W. Taylor. Dataset augmentation in feature space. *International Conference on Learning Representations (ICLR)*, 2017. 22

Y. Ding, J. H. Sohn, M. G. Kawczynski, H. Trivedi, R. Harnish, N. W. Jenkins, D. Lituiev, T. P. Copeland, M. S. Aboian, C. Mari Aparici, S. C. Behr, R. R. Flavell, S.-Y. Huang, K. A. Zalocusky, L. Nardo, Y. Seo, R. A. Hawkins, M. Hernandez Pampaloni, D. Hadley, and B. L. Franc. A Deep Learning Model to Predict a Diagnosis of Alzheimer Disease by Using 18F-FDG PET of the Brain. *Radiology*, 290:456–464, 2019. 3

C. Doersch and A. Zisserman. Multi-task self-supervised visual learning. *IEEE International Conference on Computer Vision (ICCV)*, pages 2070–2079, 2017. 11

C. Doersch, A. K. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. *IEEE International Conference on Computer Vision (ICCV)*, pages 1422–1430, 2015. 10

A. Dosovitskiy, P. Fischer, J. T. Springenberg, M. A. Riedmiller, and T. Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:1734–1747, 2014. 11

Q. Dou, D. Coelho de Castro, K. Kamnitsas, and B. Glocker. Domain Generalization via Model-Agnostic Learning of Semantic Features. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. 4

B. Dufumier, P. Gori, J. Victor, A. Grigis, and E. Duchesnay. Conditional alignment and uniformity for contrastive learning with continuous proxy labels. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021a. 15, 45, 76

B. Dufumier, P. Gori, J. Victor, A. Grigis, M. Wessa, P. Brambilla, P. Favre, M. Polosan, C. McDonald, C. M. Piguet, M. Phillips, L. Eyler, E. Duchesnay, and the Alzheimer's Disease Neuroimaging Initiative. Contrastive Learning with Continuous Proxy Meta-data for 3D MRI Classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 12902, pages 58–68. 2021b. 14, 16, 40, 54, 83

B. Dufumier, C. A. Barbano, R. Louiset, E. Duchesnay, and P. Gori. Rethinking positive sampling for contrastive learning with kernel. *ArXiv*, abs/2206.01646, 2022a. 15

B. Dufumier, A. Grigis, J. Victor, C. Ambroise, V. Frouin, and E. Duchesnay. Openbhb: a large-scale multi-site brain mri data-set for age prediction and debiasing. *NeuroImage*, 263, 2022b. 4

A. Duran, G. Dussert, O. Rouvière, T. Jaouen, P.-M. Jodoin, and C. Lartizien. ProstAttention-Net: A deep attention model for prostate cancer segmentation by aggressiveness in MRI scans. *Medical Image Analysis*, 77:102347, 2022. 81

D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9568–9577, 2021. ix, 14, 60, 63, 65

J. I. Epstein, L. Egevad, M. B. Amin, B. Delahunt, J. R. Srigley, and P. A. Humphrey. The 2014 international society of urological pathology (ISUP) consensus conference on gleason grading of prostatic carcinoma: Definition of grading patterns and proposal for a new grading system. *The American Journal of Surgical Pathology*, 40: 244–252, 2015. 54, 59

A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115–118, 2017. 3

R. Feng, Z. Zhou, M. B. Gotway, and J. Liang. Parts2whole: Self-supervised contrastive learning via reconstruction. *Domain adaptation and representation transfer, and distributed and collaborative learning workshop (MICCAI)*, 12444:85–95, 2020. 8

A. Fernandez-Quilez, T. Eftestøl, S. R. Kjosavik, M. G. Olsen, and K. Oppedal. Contrasting axial T2W MRI for prostate cancer triage: A self-supervised learning approach. *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2022. 45, 46, 59

Q. Garrido, Y. Chen, A. Bardes, L. Najman, and Y. LeCun. On the duality between contrastive and non-contrastive self-supervised learning. *International Conference on Learning Representations (ICLR)*, 2022. 16

C. Ge, J. Wang, Z. Tong, S. Chen, Y. Song, and P. Luo. Soft neighbors are positive supporters in contrastive visual representation learning. *International Conference on Learning Representations (ICLR)*, 2023. 14

S. Gidaris, P. Singh, and N. Komodakis. Unsupervised representation learning by predicting image rotations. *International Conference on Learning Representations (ICLR)*, abs/1803.07728, 2018. 10

I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014. 8

P. Goyal, D. K. Mahajan, A. K. Gupta, and I. Misra. Scaling and benchmarking self-supervised visual representation learning. *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6390–6399, 2019. 11

M. D. Greer et al. Interreader variability of prostate imaging reporting and data system version 2 in detecting and assessing prostate cancer lesions at prostate MRI. *American Journal of Roentgenology (AJR)*, pages 1–8, 2019. 54

J.-B. Grill, F. Strub, F. Altch'e, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. Á. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko. Bootstrap your own latent: A new approach to self-supervised learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. ix, 17, 60, 80

H. Guan and M. Liu. Domain adaptation for medical image analysis: A survey. *IEEE Transactions on Biomedical Engineering*, 69:1173–1185, 2021. 3

A. M. Güneş, W. van Rooij, S. Gulshad, B. J. Slotman, M. R. Dahele, and W. F. Verbakel. Impact of imperfection in medical imaging data on deep learning-based segmentation performance: An experimental study using synthesized data. *Medical physics*, 2023. 4

Y. Gutiérrez, J. Arevalo, and F. Martánez. Multimodal contrastive supervised learning to classify clinical significance MRI regions on prostate cancer. *IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1682–1685, 2022. 45, 46, 52

M. U. Gutmann and A. Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *International Conference on Artificial Intelligence and Statistics*, 2010. 11, 12

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 43

K. He, G. Gkioxari, P. Dollár, and R. B. Girshick. Mask r-cnn. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:386–397, 2017. 43

K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick. Momentum contrast for unsupervised visual representation learning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2019. ix, 12, 19

G. E. Hinton and R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313:504 – 507, 2006. 8

M. Hosseinzadeh, A. Saha, P. Brand, I. Slootweg, M. de Rooij, and H. J. Huisman. Deep learning–assisted prostate cancer detection on bi-parametric MRI: minimum training data size requirements and effect of prior knowledge. *European Radiology*, 32:2224 – 2234, 2021. 44

J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu. Squeeze-and-excitation networks. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. 43, 44

X. Hu, D. Zeng, X. Xu, and Y. Shi. Semi-supervised contrastive learning for label-efficient medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2021. 16

J. Hugosson, M. Månsson, J. Wallström, U. Axcrona, S. V. Carlsson, L. Egevad, K. Geterud, A. Khatami, K. Kohestani, C. G. Pihl, A. Socratous, J. Stranne, R. A. Godtman, and M. Hellström. Prostate cancer screening with psa and mri followed by targeted biopsy only. *The New England journal of medicine*, 387 23:2126–2137, 2022. 41

T. A. G. M. Huisman. Diffusion-weighted imaging: basic concepts and application in cerebral stroke and head trauma. *European Radiology*, 13:2283–2297, 2003. 42

T. Huynh, S. Kornblith, M. R. Walter, M. Maire, and M. Khademi. Boosting contrastive self-supervised learning with false negative cancellation. *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 986–996, 2022. 13

F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18:203 – 211, 2020. 59

P. F. Jaeger, S. A. A. Kohl, S. Bickelhaupt, F. Isensee, T. A. Kuder, H. P. Schlemmer, and K. Maier-Hein. Retina u-net: Embarrassingly simple exploitation of segmentation supervision for medical object detection. In *Machine Learning for Health workshop (NeurIPS)*, 2018. xiii, 52

A. Jiménez-Sánchez, D. Mateus, S. Kirchhoff, C. Kirchhoff, P. Biberthaler, N. Navab, M. Á. G. Ballester, and G. Piella. Medical-based deep curriculum learning for improved fracture classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2019. 18

L. Jing and Y. Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43: 4037–4058, 2019. 9

D. Karimi, H. Dou, S. Warfield, and A. Gholipour. Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. *Medical image analysis*, 65:101759, 2019. 18

V. Kasivisvanathan, A. S. Rannikko, M. Borghi, V. Panebianco, L. A. Mynderse, M. H. Vaarala, A. Briganti, L. Budäus, G. O. Hellawell, R. G. Hindley, M. J. Roobol, S. E. Eggener, M. Ghei, A. Villers, F. Bladou, G. Villeirs, J. Virdi, S. Boxler, G. Robert, P. Singh, W. Venderink, B. Hadaschik, A. Ruffion, J. C. Hu, D. J. A. Margolis, S. Crouzet, L. H. Klotz, S. S. Taneja, P. A. Pinto, I. S. Gill, C. Allen, F. Giganti, A. Freeman, S. Morris, S. Punwani, N. R. Williams, C. Brew-Graves, J. J. Deeks, Y. Takwoingi, M. Emberton, and C. M. Moore. MRI-targeted or standard biopsy for prostate-cancer diagnosis. *The New England Journal of Medicine*, 378:1767–1777, 2018. 42

P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020a. 13, 45

P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 18661–18673, 2020b. 30

D. Kim, D. Cho, D. Yoo, and I.-S. Kweon. Learning image representations by completing damaged jigsaw puzzles. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 793–802, 2018. 11

D. P. Kingma and P. Dhariwal. Glow: Generative Flow with Invertible 1x1 Convolutions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018. 24

A. Kolesnikov, X. Zhai, and L. Beyer. Revisiting self-supervised visual representation learning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1920–1929, 2019. 11

M. Koyama, K. Minami, T. Miyato, and Y. Gal. Contrastive representation learning with trainable augmentation channel. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 22

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the Association for Computing Machinery (ACM)*, 60:84 – 90, 2012. 22

G. Larsson, M. Maire, and G. Shakhnarovich. Colorization as a proxy task for visual understanding. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 840–849, 2017. 8

J. Li, R. Socher, and S. C. H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *International Conference on Learning Representations (ICLR)*, 2020a. ix, 18

J. Li, T. Lin, and Y. Xu. Sslp: Spatial guided self-supervised learning on pathological images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2021a. 15

J. Li, W. Qiang, C. Zheng, B. Su, and H. Xiong. Metaug: Contrastive learning via meta feature augmentation. In *International Conference on Machine Learning (ICML)*, 2022a. 23

S. Li, X. Xia, S. Ge, and T. Liu. Selective-supervised contrastive learning with noisy labels. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 316–325, 2022b. 19

W. Li, L. Wang, W. Li, E. Agustsson, and L. V. Gool. Webvision database: Visual learning and understanding from web data. *ArXiv*, abs/1708.02862, 2017. 17

Y. Li, G. Hu, Y. Wang, T. Hospedales, N. M. Robertson, and Y. Yang. Differentiable automatic data augmentation. In A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, editors, *European Conference on Computer Vision (ECCV)*, pages 580–595, 2020b. 27

Y. Li, G. Hu, Y. Wang, T. M. Hospedales, N. M. Robertson, and Y. Yang. Differentiable automatic data augmentation. In *European Conference on Computer Vision (ECCV)*, 2020c. 22

Z. Li, Z. Cui, S. Wang, Y. Qi, X. Ouyang, Q. Chen, Y. Yang, Z. Xue, D. Shen, and J.-Z. Cheng. Domain generalization for mammography detection via multi-style and multi-view contrastive learning. *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2021b. 52

T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017. 43

G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez. A Survey on Deep Learning in Medical Image Analysis. *Medical Image Analysis*, 42, 2017. 2

T. J. Littlejohns, J. Holliday, L. M. Gibson, S. Garratt, N. Oesingmann, F. Alfaro-Almagro, J. Bell, C. Boultwood, R. Collins, M. C. Conroy, N. Crabtree, N. Doherty, A. F. Frangi, N. C. Harvey, P. Leeson, K. L. Miller, S. Neubauer, S. E. Petersen, J. Sellors, S. Sheard, S. M. Smith, C. L. M. Sudlow, P. M. Matthews, and N. E. Allen. The uk biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. *Nature Communications*, 11, 2020. 4

A. Liu, Z. Huang, Z. Huang, and N. Wang. Direct differentiable augmentation search. *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12199–12208, 2021a. 22

A. Liu, Z. Huang, Z. Huang, and N. Wang. Direct Differentiable Augmentation Search. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12199–12208, 2021b. 27

Q. Liu, Q. Dou, and P.-A. Heng. Shape-Aware Meta-learning for Generalizing Prostate MRI Segmentation to Unseen Domains. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 475–485, 2020. 4

S. Loizillon, S. Bottani, A. Maire, S. Ströer, D. Dormont, O. Colliot, and N. Burgos. Transfer learning from synthetic to routine clinical data for motion artefact detection in brain t1-weighted mri. In *Medical Imaging*, 2023. 4

M. Lubrano, T. Lazard, G. Balezo, Y. Bellahsen-Harrar, C. Badoual, S. Berlemont, and T. Walter. Automatic grading of cervical biopsies by combining full and self-supervision. *European Conference on Computer Vision (ECCV) Workshops*, 2022. 16

D. K. Mahajan, R. B. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. R. Bharambe, and L. van der Maaten. Exploring the limits of weakly supervised pretraining. *European Conference on Computer Vision (ECCV)*, 2018. 17

C. Matsoukas, J. F. Haslum, M. Sorkhei, M. P. Soderberg, and K. Smith. What makes transfer learning work for medical images: Feature reuse & other factors. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9215–9224, 2022. ix, 3

L. McInnes and J. Healy. Umap: Uniform manifold approximation and projection for dimension reduction. *ArXiv*, abs/1802.03426, 2018. 84

B. H. Menze et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2015. 27

Z. Mirikharaji, K. Abhishek, S. Izadi, and G. Hamarneh. D-lema: Deep learning ensembles from multiple annotations - application to skin lesion segmentation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1837–1846, 2021. 18

I. Misra and L. van der Maaten. Self-supervised learning of pretext-invariant representations. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6706–6716, 2019. 12

B. Mustafa, A. Loh, J. von Freyberg, P. MacWilliams, M. Wilson, S. M. McKinney, M. Sieniek, J. Winkens, Y. Liu, P. Bui, S. Prabhakara, U. Telang, A. Karthikesalingam, N. Houlsby, and V. Natarajan. Supervised transfer learning at scale for medical imaging. *ArXiv*, abs/2101.05913, 2021. 3

M. Noroozi and P. Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision (ECCV)*, 2016. 10

D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2536–2544, 2016. 8

M. Patrick, Y. M. Asano, P. Kuznetsova, R. C. Fong, J. F. Henriques, G. Zweig, and A. Vedaldi. On compositions of transformations in contrastive self-supervised learning. *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9557–9567, 2021. 22

J. Peng, P. Wang, C. Desrosiers, and M. Pedersoli. Self-paced contrastive learning for semi-supervisedmedical image segmentation with meta-labels. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 16

A. Perakis, A. Gorji, S. Jain, K. Chaitanya, S. Rizza, and E. Konukoglu. Contrastive Learning of Single-Cell Phenotypic Representations for Treatment Classification. In *Machine Learning in Medical Imaging workshop (MICCAI)*, pages 565–575, 2021. 28

S. Purushwalkam and A. K. Gupta. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 13

M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio. Transfusion: Understanding Transfer Learning for Medical Imaging. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. 3

I. Reda, M. Ghazal, A. M. Shalaby, M. M. Elmogy, A. Aboulfotouh, M. A. El-Ghar, A. S. Elmaghraby, R. S. Keynton, and A. S. El-Baz. Detecting prostate cancer using a cnn-based system without segmentation. *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 855–858, 2019. 42

P. Ren, Y. Xiao, X. Chang, P.-Y. B. Huang, Z. Li, X. Chen, and X. Wang. A survey of deep active learning. *ACM Computing Surveys (CSUR)*, 54:1 – 40, 2020. 2

S. Ren, K. He, R. B. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2015. 43

M. Riva, P. Gori, F. Yger, and I. Bloch. Is the u-net directional-relationship aware? *IEEE International Conference on Image Processing (ICIP)*, pages 3391–3395, 2022. 10

J. Robinson, C.-Y. Chuang, S. Sra, and S. Jegelka. Contrastive learning with hard negative samples. *International Conference on Learning Representations (ICLR)*, 2020. 13

C. Rommel, T. Moreau, and A. Gramfort. Deep invariant networks with differentiable augmentation layers. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 22

O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015. 9, 43, 84

O. Rouvière, P. A. Puech, R. Renard-Penna, M. Claudon, C. Roy, F. Mege-Lechevallier, M. Decaussin-Petrucci, M. Dubreuil-Chambardel, L. Magaud, L. Remontet, A. Ruffion, M. Colombel, S. Crouzet, A.-M. Schott, L. Lemaitre, M. Rabilloud, and N. Grenier. Use of prostate systematic and targeted biopsy on the basis of multiparametric MRI in biopsy-naive patients (MRI-FIRST): a prospective, multicentre, paired diagnostic study. *The Lancet. Oncology*, 20 1:100–109, 2019. 4, 40, 41

O. Rouvière, T. Jaouen, P. Baseilhac, M. L. Benomar, R. Escande, S. Crouzet, and R. Souchon. Artificial intelligence algorithms aimed at characterizing or detecting prostate cancer on MRI: How accurate are they when tested on independent cohorts? - a systematic review. *Diagnostic and interventional imaging*, 2022. 44

A. Sadafi, N. Koehler, A. Makhro, A. Bogdanova, N. Navab, C. Marr, and T. Peng. Multiclass Deep Active Learning for Detecting Red Blood Cell Subtypes in Brightfield Microscopy. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 685–693. Springer International Publishing, 2019. 2

A. Saha, M. Hosseinzadeh, and H. J. Huisman. End-to-end prostate cancer detection in bpmri via 3d cnns: Effect of attention mechanisms, clinical priori and decoupled false positive reduction. *Medical image analysis*, 73:102155, 2021. 40, 44, 60

A. Saha, J. J. Twilt, et al. Artificial Intelligence and Radiologists at Prostate Cancer Detection in MRI: The PI-CAI Challenge. 2022. 53, 58

É. Sarfati, A. Bône, M.-M. Rohé, P. Gori, and I. Bloch. Learning to diagnose cirrhosis from radiological and histological labels with joint self and weakly-supervised pretraining strategies. *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2023. 16

D. Saxena and J. Cao. Generative adversarial networks (gans). *ACM Computing Surveys (CSUR)*, 54:1 – 42, 2020. 8

P. Schelb, S. A. A. Kohl, J. P. Radtke, M. Wiesenfarth, P. Kickingereder, S. Bickelhaupt, T. A. Kuder, A. Stenzinger, M. Hohenfellner, H. P. Schlemmer, K. Maier-Hein, and D. Bonekamp. Classification of cancer at prostate MRI: Deep learning versus clinical pi-rads assessment. *Radiology*, page 190938, 2019. 43

P. Schelb, X. Wang, J. P. Radtke, M. Wiesenfarth, P. Kickingereder, A. Stenzinger, M. Hohenfellner, H. P. Schlemmer, K. Maier-Hein, and D. Bonekamp. Simulated clinical deployment of fully automatic deep learning for clinical prostate MRI assessment. *European Radiology*, 31:302 – 313, 2020. 44, 81

J. Schlemper, O. Oktay, M. Schaap, M. P. Heinrich, B. Kainz, B. Glocker, and D. Rueckert. Attention gated networks: Learning to leverage salient regions in medical images. *Medical image analysis*, 53:197 – 207, 2018. 44

M. S. Seyfiouglu, Z. Liu, P. Kamath, S. Gangolli, S. Wang, T. Grabowski, and L. G. Shapiro. Brain-aware replacements for supervised contrastive learning in detection of alzheimer's disease. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2022. 21

Y. Shi, N. Siddharth, P. H. S. Torr, and A. R. Kosiorek. Adversarial masking for self-supervised learning. *International Conference on Machine Learning (ICML)*, 2022. 23

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*, 2014. 44

C. P. Smith et al. Intra- and interreader reproducibility of PI-RADSv2: A multireader study. *Journal of Magnetic Resonance Imaging*, 49, 2019. 53

D. W. Strand, D. N. Costa, F. Francis, W. A. Ricke, and C. G. Roehrborn. Targeting phenotypic heterogeneity in benign prostatic hyperplasia. *Differentiation; research in biological diversity*, 96:49–61, 2017. 49

Y. Sumathipala, N. S. Lay, B. I. Turkbey, C. P. Smith, P. L. Choyke, and R. M. Summers. Prostate cancer detection from multi-institution multiparametric mris using deep convolutional neural networks. *Journal of Medical Imaging*, 5:044507 – 044507, 2018. 44

N. Sushentsev, N. M. da Silva, M. Yeung, T. Barrett, E. Sala, M. Roberts, and L. Rundo. Comparative performance of fully-automated and semi-automated artificial intelligence methods for the detection of clinically significant prostate cancer on MRI: a systematic review. *Insights into Imaging*, 13, 2022. 44

A. Tamkin, M. Wu, and N. D. Goodman. Viewmaker networks: Learning views for unsupervised representation learning. *International Conference on Learning Representations (ICLR)*, 2020. 23

Y. Tang, Y.-B. Tang, Y. Peng, K. Yan, M. Bagheri, B. A. Redd, C. J. Brandon, Z. Lu, M. Han, J. Xiao, and R. M. Summers. Automated abnormality classification of chest radiographs using deep convolutional neural networks. *NPJ Digital Medicine*, 3, 2020. 27, 28, 30

X. Tao, Y. Li, W. Zhou, K. Ma, and Y. Zheng. Revisiting rubik's cube: Self-supervised learning with volume-wise transformation for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2020. 8

M. Tardy and D. Mateus. Looking for abnormalities in mammograms with self- and weakly supervised reconstruction. *IEEE Transactions on Medical Imaging*, 40:2711–2722, 2021. 8, 9

Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola. What Makes for Good Views for Contrastive Learning? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 13, 21, 22, 23, 24, 27, 28, 37

Y. Tian, X. Chen, and S. Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *International Conference on Machine Learning (ICML)*, 2021. 16

Y.-H. H. Tsai, S. Bai, L.-P. Morency, and R. Salakhutdinov. A note on connecting barlow twins with negative-sample-free contrastive learning. *ArXiv*, 2021. 61

Y.-H. H. Tsai, T. Li, M. Q. Ma, H. Zhao, K. Zhang, L.-P. Morency, and R. Salakhutdinov. Conditional contrastive learning with kernel. *International Conference on Learning Representations (ICLR)*, 2022. 14, 45

B. I. Turkbey, A. B. Rosenkrantz, M. A. Haider, A. R. Padhani, G. Villeirs, K. J. Macura, C. M. Tempany, P. L. Choyke, F. Cornud, D. J. A. Margolis, H. C. Thoeny, S. Verma, J. O. Barentsz, and J. C. Weinreb. Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2. *European urology*, 2019. 4, 41, 42, 47, 49, 54, 76

V. Udandarao, A. Maiti, D. Srivatsav, S. R. Vyalla, Y. Yin, and R. R. Shah. Cobra: Contrastive bi-modal representation algorithm. *ArXiv*, abs/2005.03687, 2020. 52

A. van den Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018. 23, 26

L. van der Maaten and G. E. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008. 31

A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 17

H. Wang, R. Xiao, Y. Li, L. Feng, G. Niu, G. Chen, and J. J. Zhao. Pico: Contrastive label disambiguation for partial label learning. *International Conference on Learning Representations (ICLR)*, 2022a. 19

T. Wang and P. Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. *International Conference on Machine Learning (ICML)*, 2020. 34, 45, 54, 60, 70

X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. ix, 3, 15

Y. Wang, Q. Zhang, Y. Wang, J. Yang, and Z. Lin. Chaos is a ladder: A new theoretical understanding of contrastive learning via augmentation overlap. *International Conference on Learning Representations (ICLR)*, 2022b. 29

S. Warfield, K. H. Zou, and W. M. Wells. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23:903–921, 2004. 18, 48

C. Wei, H. Wang, W. Shen, and A. L. Yuille. Co2: Consistent contrast for unsupervised visual representation learning. *International Conference on Learning Representations (ICLR)*, 2020. 14

J. C. Weinreb, J. O. Barentsz, P. L. Choyke, F. Cornud, M. A. Haider, K. J. Macura, D. J. A. Margolis, M. D. Schnall, F. Shtern, C. M. Tempany, H. C. Thoeny, and S. Verma. Pi-rads prostate imaging - reporting and data system: 2015, version 2. *European urology*, 69 1:16–40, 2016. 41

A. C. Westphalen et al. Variability of the positive predictive value of PI-RADS for prostate MRI across 26 centers: Experience of the society of abdominal radiology prostate cancer disease-focused panel. *Radiology*, page 190646, 2020. 54

Z. Wu, Y. Xiong, S. X. Yu, and D. Lin. Unsupervised feature learning via non-parametric instance discrimination. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018. ix, 11, 12

T. Xiao, X. Wang, A. A. Efros, and T. Darrell. What should not be contrastive in contrastive learning. In *International Conference on Learning Representations (ICLR)*, 2021. 13, 24, 25

S. Xie and Z. Tu. Holistically-nested edge detection. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1395–1403, 2015. 44

Y. Xue, K. Whitecross, and B. Mirzasoleiman. Investigating why contrastive learning benefits robustness against label noise. In *International Conference on Machine Learning (ICML)*, 2022. 18

K. Yan, J. Cai, D. Jin, S. Miao, A. P. Harrison, D. Guo, Y. Tang, J. Xiao, J. Lu, and L. Lu. Sam: Self-supervised learning of pixel-wise anatomical embeddings in radiological images. *IEEE Transactions on Medical Imaging*, 41:2658–2669, 2020. 15

F. Yang, G. Zamzmi, S. Angara, S. Rajaraman, A. Aquilina, Z. Xue, S. Jaeger, E. Papagiannakis, and S. K. Antani. Assessing inter-annotator agreement for medical image segmentation. *IEEE access : practical innovations, open solutions*, 11:21300 – 21312, 2023. 18

J. Yang, J. Duan, S. Tran, Y. Xu, S. Chanda, L. Chen, B. Zeng, T. M. Chilimbi, and J. Huang. Vision-language pre-training with triple contrastive learning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15650–15659, 2022a. 51

K. Yang, T. Zhou, X. Tian, and D. Tao. Identity-disentangled adversarial augmentation for self-supervised learning. In *International Conference on Machine Learning (ICML)*, 2022b. 23

L. Yang, Y. Zhang, J. Chen, S. Zhang, and D. Z. Chen. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 399–407, 2017. 2

X. Yang, R. Kwitt, and M. Niethammer. Fast predictive image registration. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (DLMIA) Workshop (MICCAI)*, 2016. 59

C. Yeh, C. Hong, et al. Decoupled contrastive learning. In *European Conference on Computer Vision (ECCV)*, pages 668–684, 2022. 58

L. Yi, S. Liu, Q. She, A. McLeod, and B. Wang. On learning contrastive representations for learning with noisy labels. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16661–16670, 2022. 18

S. Yoo, I. Gujrathi, M. A. Haider, and F. Khalvati. Prostate cancer detection using deep convolutional neural networks. *Scientific Reports*, 9, 2019. 42

K.-H. Yu, A. L. Beam, and I. S. Kohane. Artificial intelligence in healthcare. *Nature Biomedical Engineering*, 2:719–731, 2018. 2

X. Yu, B. Lou, D. Zhang, D. J. Winkel, N. Arrahmane, M. Diallo, T. Meng, H. von Busch, R. Grimm, B. Kiefer, D. Comaniciu, and A. Kamen. Deep attentive panoptic model for prostate cancer detection using biparametric MRI scans. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2020. 40, 43, 60

X. Yuan, Z. L. Lin, J. Kuen, J. Zhang, Y. Wang, M. Maire, A. Kale, and B. Faieta. Multimodal contrastive training for visual representation learning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6991–7000, 2021. 51

J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning (ICML)*, 2021. 13, 60, 80

D. Zeng, Y. Wu, X. Hu, X. Xu, H. Yuan, M. Huang, J. Zhuang, J. Hu, and Y. Shi. Positional contrastive learning for volumetric medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2021. 15

X. Zhai, A. Oliver, A. Kolesnikov, and L. Beyer. S4l: Self-supervised semi-supervised learning. *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1476–1485, 2019. 11

C. Zhang, K. Zhang, C. Zhang, T. X. Pham, C. D. Yoo, and I. S. Kweon. How does simsiam avoid collapse without negative samples? a unified understanding with self-supervised contrastive learning. *International Conference on Learning Representations (ICLR)*, 2022. 16, 17

L. Zhang, R. Tanno, M. Xu, Y. Huang, K. Bronik, C. Jin, J. Jacob, Y. Zheng, L. Shao, O. Ciccarelli, F. Barkhof, and D. C. Alexander. Learning from multiple annotators for medical image segmentation. *Pattern Recognition*, 138:109400, 2023. 18, 48

R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. In *European Conference on Computer Vision (ECCV)*, 2016. 8

E. Zheltonozhskii, C. Baskin, A. Mendelson, A. M. Bronstein, and O. Litany. Contrast to divide: Self-supervised pre-training for learning with noisy labels. *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 387–397, 2022. 18

H. Zheng, J. Han, H. Wang, L. Yang, Z. Zhao, C. Wang, and D. Z. Chen. Hierarchical self-supervised learning for medical image segmentation based on multi-domain data aggregation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2021. xii, 16, 67, 68

Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang. Unet++: A nested u-net architecture for medical image segmentation. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (DLMIA) Workshop (MICCAI)*, 11045:3–11, 2018. 44

Z. Zhou, V. Sodha, J. Pang, M. B. Gotway, and J. Liang. Models Genesis: Generic Autodidactic Models for 3D Medical Image Analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 11767, pages 384–393. 2019a. 86

Z. Zhou, V. Sodha, M. M. R. Siddiquee, R. Feng, N. Tajbakhsh, M. B. Gotway, and J. Liang. Models genesis: Generic autodidactic models for 3d medical image analysis. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 11767:384–393, 2019b. ix, 8, 9, 70, 80, 81

J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251, 2017. 52

ECOLE
DOCTORALE

Titre : Méthodes et systèmes d'optimisation de la charge d'annotation en imagerie médicale pour les algorithmes d'apprentissage

Mots clés : Apprentissage, Imagerie Médicale, Annotation

Résumé : Ces dernières années, la quantité de données d'imagerie médicale n'a cessé de croître. En 1980, 30 minutes d'acquisition étaient nécessaires pour obtenir 40 images médicales. Aujourd'hui, 1000 images peuvent être acquises en 4 secondes.
Cette croissance de la quantité de données est allée de pair avec le développement de techniques d'apprentissage profond qui ont besoin d'annotations de qualité pour être entraînées. En imagerie médicale, les annotations sont beaucoup plus coûteuses à obtenir car elles nécessitent l'expertise d'un radiologue dont le temps est limité.
L'objectif de cette thèse est de proposer et de développer des méthodes permettant de limiter la charge d'annotation en imagerie médicale tout en maintenant une performance élevée des algorithmes d'apprentissage profond.
Dans la première partie de cette thèse, nous étudions les méthodes d'apprentissage auto-supervisé. Ces méthodes introduisent des sous-tâches de différents types : approches génératives, contextuelle et basée sur l'auto-distillation. Ces tâches sont utilisées pour pré-entraîner un réseau de neurones sans annotations supplémentaires afin de tirer profit des données non annotées disponibles.
La plupart de ces tâches utilisent des perturbations assez génériques, sans rapport avec la tâche supervisée sous-jacente et échantillonnées au hasard dans une liste avec des paramètres fixés. La meilleure façon de combiner et de choisir ces perturbations et leurs paramètres n'est pas encore claire. En outre, certaines perturbations peuvent être préjudiciables à la tâche supervisée objectif. Certains travaux atténuent ce problème en concevant des sous-tâches pour une tâche supervisée spécifique, en particulier dans le domaine de l'imagerie médicale. Mais ces tâches ne se généralisent pas bien à d'autres problèmes. Un équilibre doit donc être trouvé entre l'optimisation de la perturbation ou de la sous-tâche pour un problème supervisé donné et la capacité de généralisation de la méthode.
Parmi les méthodes basées sur le contexte, les approches d'apprentissage contrastif proposent une tâche de discrimination par instance : l'espace latent est structuré suivant la similarité entre différentes instances. La définition de la similarité des instances est le principal défi de ces approches et a été largement explorée. Lorsque des perturbations sont utilisées pour définir la similarité entre les images, les mêmes questions d'optimisation des perturbations se posent.
Nous introduisons un générateur de perturbations optimisé pour le pré-entraînement contrastif guidé par une petite quantité de supervision.
Les annotations de classes et certaines métadonnées ont été utilisées pour conditionner la similarité des instances, mais ces données peuvent être sujettes à la variabilité des annotateurs, en particulier dans le domaine médical. Certaines méthodes ont été proposées pour utiliser la confiance dans l'apprentissage supervisé et auto-supervisé, mais elles sont principalement basées sur les valeurs de la fonction de perte. Cependant, la confiance dans les annotations et les métadonnées est souvent liée à des connaissances a priori du domaine, telles que l'acquisition des données, l'expérience et l'accord entre les annotateurs. Ceci est encore plus pertinent pour les données médicales.
Dans la deuxième partie de cette thèse, nous proposons une fonction de perte contrastive prenant en compte la confiance des annotations pour le problème spécifique de la détection des lésions du cancer de la prostate.
Enfin, nous explorons quelques approches pour appliquer l'apprentissage auto-supervisé et contrastif à la segmentation des lésions du cancer de la prostate.

Title : Methods and frameworks of annotation cost optimization for deep learning algorithms applied to medical imaging

Keywords : Machine Learning, Medical Imaging, Annotation

Abstract : In recent years, the amount of medical imaging data has kept on growing. In 1980, 30 minutes of acquisition were necessary to obtain 40 medical images. Today, 1000 images can be acquired in 4 seconds.
This growth in the amount of data has gone hand in hand with the development of deep learning techniques which need quality labels to be trained. In medical imaging, labels are much more expensive to obtain as they require the expertise of a radiologist whose time is limited.
The goal of this thesis is to propose and develop methods to limit the annotation load in medical imaging while maintaining a high performance of deep learning algorithms.
In the first part of this thesis, we focus on self-supervised learning methods which introduce pretext tasks of various types : generation-based, context-based, and self-distillation approaches. These tasks are used to pre-train a neural network with no additional annotations to take advantage of the available unannotated data.
Most of these tasks use perturbations often quite generic, unrelated to the objective task and sampled at random in a fixed list with fixed parameters. How to best combine and choose these perturbations and their parameters remains unclear. Furthermore, some perturbations can be detrimental to the target supervised task. Some works mitigate this issue by designing pretext tasks for a specific supervised task, especially in medical imaging. However, these tasks do not generalize well to other problems. A balance must be found between perturbation or pretext task optimization for a given supervised problem and the method's generalization ability.
Among context-based methods, contrastive learning approaches propose an instance-level discrimination task : the latent space is structured with instance similarity. Defining instance similarity is the main challenge of these approaches and has been widely explored. When defining similarity through perturbed versions of the same image, the same questions of perturbation optimization arise.
We introduce a perturbation generator optimized for contrastive pre-training guided by a small amount of supervision.
Class labels and metadata have been used to condition instance similarity, but these data can be subject to annotator variability, especially in the medical domain. Some methods have been proposed to use confidence in fully supervised and self-supervised training, but it is mostly based on loss function values. However, confidence on labels and metadata is often linked to a priori domain knowledge such as data acquisition, annotators' experience, and agreement. This is even more relevant for medical data.
In the second part of this thesis, we design an adapted contrastive loss introducing annotation confidence for the specific problem of prostate cancer lesion detection.
Finally, we explore some approaches to apply self-supervised and contrastive learning to prostate cancer lesion segmentation.

Institut Polytechnique de Paris
91120 Palaiseau, France