Doctoral Dissertation
Doctoral Program in Computer Enginering (34.th cycle)

Machine learning methods for the analysis and interpretation of images and other multi-dimensional data

Sina Famouri

* * * * * *

Supervisor
Prof. Fabrizio Lamberti

Doctoral Examination Committee:
Prof. Anke Meyer-Baese, Referee, Florida State University
Prof. Anna Vignati, Referee, Università di Torino
Prof. Elisa Syrol Clols, Universitat Politècnica de Catalunya
Prof. Giorgio Leonardi, Università del Piemonte Orientale
Prof. Silvia Anna Chiusano, Politecnico di Torino

Politecnico di Torino
October 28, 2022

I hereby declare that, the contents and organisation of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

. . . . . .

Sina Famouri

Turin, October 28, 2022

Summary

Machine learning, and in particular deep learning-based models such as convolutional neural networks, have consistently demonstrated unprecedented performance in the analysis and interpretation of images and more generally of multidimensional complex data. In this context, fine-grained object recognition and classification, characterized by subtle differences between classes and large variations within classes, is attracting increasing attention not only in the general computer vision community, but also in specialized fields with high performance requirements, such as medical and industrial applications. Since collecting large annotated datasets is expensive and time consuming, especially in the above-mentioned domains, advances in unsupervised, semi-supervised, self-supervised, and transfer learning are key to substantial improvements. In particular, this research focuses on computational diagnosis in the medical domain, which combines the need for fine-grained analysis of subtle disease patterns with the hurdles of developing deep learning models that can efficiently process high-dimensional and multidimensional images.

Despite the contribution of deep learning in the medical field, the success of such models is hampered by data limitations. In terms of data volume, the digitization of medical records generates enormous amounts of data on a daily basis. However, apart from the administrative difficulties in obtaining the data, acquiring large and well-curated datasets is quite costly. Therefore, the typical size of datasets is rather small compared to natural RGB images. Hence, it may be difficult for deep neural networks to generalize beyond initial laboratory testing, limiting clinical application or even requiring further training in local clinical centers.

The goal of this thesis is to explore various ways to reduce deep learning overreliance on large datasets to make it more data efficient without sacrificing robustness. Possible solutions to combat the lack of data in the medical domain include either reducing the cost of data annotation or, alternatively, reducing the amount of annotated data required for training. This research takes a step toward the following objectives: (1) providing methods that makes training robust against noise, (2) incorporating information from different views of the same organ, and (3) developing multi-task, cross-domain training pipelines that exploit self-supervision to lessen reliance on annotations.

For object detection in medical images, achieved results show that, if the reference standard is noisy, the model is not able to correctly quantify the agreement of reference standard and the networks predictions, which consequently affects the training in a negative way. As a result, a novel approach is devised to counteract this effect by changing the criteria used for matching the bounding boxes proposed by the network with the ground truth. This helps to relax the requirements on annotated medical images and makes the training robust to noise, enabling the use of larger automatically curated datasets for training.

Regarding the use of information from different views of the same organ, a registration method has been proposed to align two views of the same organ in mammography images and incorporate this information into object recognition.

Finally, to address the problem of data scarcity, a dataset has been created as a harmonized collection of 17 publicly available datasets covering various medical imaging modalities and body parts. Subsequently, self-supervised pre-training has been applied to this large collection, and the learned representations have been evaluated by transfer learning to different tasks of computer-aided diagnosis. The aforementioned dataset and the model based on it are referred to as MedNet interchangeably in this context. Experimental results have shown that the extracted features are able to discriminate better than ImageNet, which is currently the *de facto* standard for transfer learning in the medical field. However, further investigation has shown that the final performance of the two pre-trained networks is similar after fine-tuning and that ImageNet slightly outperforms MedNet. Thus, this research has examined the complementary roles played not only by the source and target datasets, but also by the self-supervised and target tasks in determining the most effective strategy for training deep neural networks on medical datasets.

Acknowledgements

I would like to thank my supervisor, Prof. Fabrizio Lamberti, for the support provided throughout the PhD. I would also like to express my deepest gratitude to Dr. Lia Morra, who guided me towards the objectives accomplished in this thesis.

This research was partially supported by NVIDIA Corporation through the donation of a Titan Xp GPU.

Additional computational resources were provided by the HPC@POLITO initiative of Politecnico di Torino (http://www.hpc.polito.it)

Contents

List of Tables x

List of Figures xi

1 Introduction 1
1.1 Contributions . . . 4

2 Feasibility of training object detectors with noisy data 7
2.1 Introduction . . . 7
2.2 Background and related work . . . 10
2.2.1 Object detection and the Faster R-CNN architecture . . . 10
2.2.2 Labeling noise in deep neural networks . . . 11
2.2.3 Breast mass detection . . . 13
2.2.4 Metrics and losses for bounding box regression . . . 14
2.3 Materials and methods . . . 16
2.3.1 Dataset . . . 18
2.3.2 Noise modeling . . . 18
2.3.3 Matching criterion . . . 19
2.3.4 Deep network architecture . . . 22
2.3.5 Experimental setup . . . 23
2.3.6 Evaluation . . . 24
2.4 Results . . . 25
2.5 Discussion . . . 30
2.6 Conclusion . . . 31

3 Multi-view lesion detection and registration in mammography 33
3.1 Introduction . . . 33
3.2 Background and related work . . . 35
3.2.1 Deep learning for medical image registration . . . 35
3.2.2 Multi-view lesion detection in mammography . . . 37
3.3 Deep learning methods for lesion detection in co-registered multiple mammography views . . . 38

3.3.1 Overview . . . 38
3.3.2 Affine registration . . . 38
3.3.3 Loss: affine registration . . . 41
3.4 Elastic registration . . . 42
3.4.1 Loss: elastic registration . . . 43
3.5 Multi-stream object detection . . . 44
3.5.1 Architecture . . . 44
3.5.2 Loss: multi-view object detection . . . 45
3.6 Experimental setup . . . 45
3.6.1 Dataset . . . 45
3.6.2 Transfer learning . . . 46
3.6.3 Pectoral muscle removal . . . 47
3.6.4 Hyperparameter setup . . . 47
3.6.5 Hardware and software setup . . . 49
3.6.6 Evaluation . . . 49
3.7 Results . . . 49
3.7.1 Affine registration . . . 49
3.7.2 Elastic registration . . . 52
3.7.3 Multi-view vs. single-view lesion detection: convergence analysis . . . 55
3.7.4 Multi-view vs. single-view lesion detection: performance analysis . . . 59
3.8 Conclusion . . . 61

4 Self-supervised pre-training for robust representation learning: the MedNet framework 65
4.1 Introduction . . . 65
4.2 Related work . . . 67
4.2.1 Introduction to self-supervised learning . . . 67
4.2.2 Transfer and self-supervised learning in the medical domain 69
4.3 The MedNet dataset . . . 70
4.3.1 Datasets description . . . 70
4.3.2 Sampling and distribution . . . 73
4.3.3 Preprocessing . . . 74
4.3.4 Task specific datasets for transfer learning . . . 75
4.4 Methods . . . 76
4.4.1 Pre-training . . . 76
4.4.2 Transfer learning . . . 78
4.5 Results . . . 82
4.5.1 Representations . . . 82
4.6 Discussion . . . 86
4.7 Conclusion . . . 89

4.8 Future work 90

5 Conclusion 91

Bibliography 93

List of Tables

2.1 Confidence intervals for the AFROC. 28
3.1 FROC curve comparison between Multi-View and Single-View network. 59
4.1 List of the datasets used for pre-training 71
4.2 Datasets used for transfer learning 76
4.3 AUC @ step 1 85
4.4 AUC @ step 2 86
4.5 AUC for one step transfer 86
4.6 Confidence interval 86

List of Figures

2.1 The training framework of Faster R-CNN is summarized in this figure. The noise is injected onto the bounding boxes in the annotations and its impact on the training is highlighted by the red elements. Specifically, the red circle shows where the labeling noise is generated. Faster R-CNN is a two-stage object detector: the RPN filters out candidate regions from the background, while the classifier assigns them the most likely class. Since lesions are rare compared to background, a hard-sampling procedure was added to the framework to avoid overfitting and ensure that more informative region proposals are passed to the classifier. If the bounding boxes are not tightly drawn around the lesion borders, the region proposal that are passed to the classifier may be mislabelled during the training procedure. This is because the region proposals are automatically compared to the reference standard using a matching criterion, such as the IoU, which can lead to incorrect results if the bounding boxes do not match the lesion borders exactly. Figure reproduced from [34]. 17

2.2 Histograms of the noise factor n_{wi} from $\mu = 0$ to $\mu = 3$. All distributions are clipped to the range [0, 5]. Figure reproduced from [34]. 19

2.3 Diameter distribution (histogram) of the noisy bounding boxes for the clean dataset and for each level of noise. Figure reproduced from [34]. 20

2.4 The red boxes show three different bounding box proposals for a lesion. If IoU is used as the matching criterion, the proposals would be scored as $C > B > A$, where the highest score indicates higher similarity to the reference standard. However, based on the distance between centroids, the order would be $A > C > B$. Figure reproduced from [34]. 21

2.5 Comparison of the FROC curves with and without hard sample mining. Figure reproduced from [34]. 25

2.6 Clean (left) and noisy (right) reference standard box compared with ROI proposals generated by the RPN (in green). Many FPs overlap with the noisy reference standard box and, hence, may be mislabelled as TPs by a suboptimal matching criterion. Figure reproduced from [34]. 26
2.7 Average number of anchors per lesion labeled as positive examples in the first iteration of RPN training. Results are compared for Exp_IoU, IoU, and Centroid criteria for the clean dataset (blue) and for increasing noise (from yellow to purple). The number of positive anchors (and thus noise) increases with more relaxed matching criteria and increases more than linearly with the amount of noise. All scales are logarithmic. Figure reproduced from [34]. 27
2.8 FROC curves with 95% confidence interval (calculated by bootstrapping). From left to right the IoU, Centroid, and Exp_IoU criterion were used. The latter is more tolerant towards noise with comparable performance across all levels of noise. Figure reproduced from [34]. 27
2.9 Area under the FROC curves for the IoU (orange), the Centroid inside the bounding box (green) and the Exp_IoU criteria as a function of the noise level. Figure reproduced from [34]. 28
2.10 Examples showing the clean and noisy reference standard annotations (red) vs. Faster R-CNN predictions (green), with IoU matching criterion. From left to right the level of noise increases: it can be seen how the number of false positives detected by Faster R-CNN increases as well. Figure reproduced from [34]. 29
3.1 Architecture of the proposed affine registration network. The feature extraction backbone is the ResNet50 network up to the Conv4_x blocks [47]. Weight sharing between the CC and MLO views reduces the parameters count and prevents overfitting. Figure reproduced from [33]. 39
3.2 Calculation of the overlap mask for the MSE loss. Unregistered (red box) and registered (green box) CC views are shown in (a) and (b). The shaded blue area is included in the calculation of the loss (b). In (c) the registered CC, fixed MLO and overlap mask are shown superimposed. It can be noticed how the margin of the CC view aligns with the pectoral muscle, outside of the overlap area.. . . . 41
3.3 Deformable registration framework. 43

3.4 Multi-view Faster R-CNN architecture. The network takes as input the MLO and the co-registered CC image. A backbone with shared weights computes the feature maps, which are then fed to the RPN and classifier heads. The region proposals output by the RPN are fed to a Non-Maximums Suppression layer and then to a Dual-view Region Pooling layer which combines features from both views. The region proposals are then classified by a dual-output classifier head which outputs separate classification and regression parameters for each view. 44
3.5 Evolution of the loss during training: MSE (a) and GIoU (b) . . . 49
3.6 Histogram of the GIoU loss for the test set when incorporating a backbone pre-trained on ImageNet (a) and on single-view object detection (b). 50
3.7 Affine registration examples: the MLO and registered CC views are shown overlapped. The MLO bounding box is shown in red, the CC in blue, before (cyan blue) and after rectification. 51
3.8 Histogram of the GIoU loss for the test set for affine (a) and elastic (b) registration. 52
3.9 Examples of successful elastic registration (test set). The MLO and registered CC views are shown separately and then overlapped. The MLO bounding box is shown in red, the CC in blue, after rectification. 53
3.10 Examples of unsuccessful elastic registration (test set). The MLO and registered CC views are shown separately and then overlapped. The MLO bounding box is shown in red, the CC in blue, after rectification. 54
3.11 Monitoring of Multi-View Training (CC view on the left column and MLO view on the right column): Detector Losses (a,b), RPN Losses (c,d) and Mean Overlapping Bounding Boxes with Ground Truth (e). 56
3.12 Single-View Network Training FROC @ 20, 40, 60, 80, 100 epochs: full curve (a) and truncated at 2FPs/image (b). 57
3.13 Multi-View Network Training FROC @ 20, 40, 60, 80, 100 epochs: full curve (a) and truncated at 2FPs/image (b). 57
3.14 Model convergence analysis: train set FROC curves of Multi-View network and Single-View network at epochs 10, 20, 30, 40, 50, 60. . 58
3.15 FROC curves on the CBIS-DDSM test set. 59
3.16 Predictions of the multi-view network with a single class vector for both views on an example in which the registration was not able to align the lesion bounding boxes. 60
3.17 Single-View Network: Mass detection Results on CBIS-DDSM. . . . 61
3.18 Multi-View Network: Mass detection results in CBIS-DDSM. 62
4.1 Portions of the dataset occupied by each modality is balanced. . . . 74

4.2 The organization of the methodology and experiments for evaluation of the pretrained weights is illustrated. The pretrained weights features shown in yellow have been used for visualization via TSNE and also transfer learning at the first step. 77
4.3 Illustration of Model Genesis framework. The images will randomly go through at most three transformations. $\tilde{X}$ is the transformed version of X. during training the model tries to minimize the reconstruction error between X' and X. 78
4.4 Examples of MRI images in the test set. 79
4.5 Examples of CT images from the test set. 80
4.6 Examples of X-Ray images from the test set. 81
4.7 T-SNE visualization of the representations colored by modality . . . 83
4.8 The T-SNE embeddings shows that the model is able to cluster similar images together as well as clustering them by modality. Each cluster is marked and on the border of the image example images from each cluster is included. 84
4.9 Filters of the first convolution layers of the encoder 87
4.10 Filters of the middle convolution layers of the encoder 87
4.11 Filters of the final convolution layers of the encoder 88
4.12 Reconstruction of patches from the Luna16 dataset that contain nodules 89

Chapter 1

Introduction

In the last ten years, deep learning has led to major breakthroughs in many fields, including computer vision [67], medical imaging [84, 109], text analysis, cyber-security [135, 60], and many others. In particular, this thesis focuses on computer-aided diagnosis in the medical domain, which combines the need for fine-grained analysis of subtle disease patterns with the hurdles of designing deep learning models that can efficiently process high-dimensional and multi-dimensional images.

Computer-aided diagnosis and, more generally, applications of deep learning in radiology have been at the center of the attention of many deep learning practitioners [109, 84]. Not only were deep learning-based systems shown to systematically outperform conventional machine learning systems, they have also started to rival human radiologists as well [109]. The healthcare sector, and radiology in particular, is thus poised to immensely benefit from the deep learning revolution. Key expected benefits include decreasing costs [5, 20], tackling the radiologist shortage [66, 103, 53, 20], and enabling personalized medicine [38, 109, 121, 41].

There are several reasons underpinning deep learning success in medical imaging. First, the volume of data produced in healthcare and medicine is enormous and deep learning models scale effectively to large datasets without saturating performance. Second, deep learning is capable of automatically learning a compact and hierarchical representation by directly feeding raw data to a deep network [32, 109]. This approach bypasses the need for hand-crafted feature engineering, and has been demonstrated to obtain more discriminative features with minimal design time. Furthermore, such representations can be easily transferred to different domains, datasets or tasks by simply initializing a deep neural network with pre-trained weights.

Despite the effectiveness of deep neural networks and the astonishing amount of data collected on a daily basis, several issues still prevent to reap the full potential of deep learning in healthcare and medicine. In radiology, one such issue is that the most common paradigm for training deep neural networks, i.e., supervised learning,

requires large and well-curated training sets. However, collecting such datasets is challenging and costly due to many technical, legal, and administrative barriers [62, 75]. While some of these barriers may be overcome through the collaboration of hospitals, research centers and hardware/software vendors, the cost of data annotation remains a fundamental barrier to deep learning adoption. To obtain such annotations, images must be compared against an established reference standard, which could be an independent device or procedure (such as biopsy or follow-up) or could be obtained based on the consensus opinion of several radiologists. This costly and time consuming data presentation is one of the main limitations to building supervised deep-learning networks for medical imaging tasks [32]. As a result, the typical size of datasets is rather small with respect to natural RGB images. As as a result, it may be difficult for deep neural networks to generalize beyond the initial laboratory testing, which limits clinical adoption [109], or may require even require further training at local institutions [25].

Another important issue to highlight is the high variability of medical images, in terms of modality, resolution, acquisition protocols, noise, post-processing and many other factors. Not to mention that research and development is constantly pushing new devices on the market, which would require deep neural networks to be consistently updated. While, in theory, deep neural networks could learn to be robust to variations in the acquisition protocol, in practice it is difficult to collect sufficiently large datasets that capture the entire spectrum that a deep neural network may encounter in the clinic. In order to build robust deep neural networks at acceptable cost, it is first necessary to reduce the cost and effort required for data annotation. Then, one must analyze whether, and how, a deep neural network can more easily generalize to additional modalities or acquisition protocols.

As a motivating example, one could consider one of the most studied computer-aided detection/diagnosis applications, which is screening mammography [89]. Breast cancer, according to the American Cancer Society, is the most diagnosed cancer with more the 250 thousand cases per year in the USA and more than two million all over the world. Mammography screening has been shown to reduce mortality by enabling early diagnosis and treatment. Yet, screening programs are expensive and challenging to set-up and maintain. The very low prevalence (roughly four or five cases out of 1000 screened women) and the variety of diagnostic signs associated to breast cancer makes screening mammography extremely challenging to read and, for the same reasons, very large-scale datasets are required to train effective deep neural networks. Screening programs requires radiologists with several years of experience to maintain the required levels of sensitivity and specificity. Most of the radiologists' time, however, is spent ruling out negative cases, whereas the relatively high recall rate (between 3% and 15% depending on the screening program) implies that most of the women recalled for further workup are actually negative, increasing costs and anxiety associated to screening and reducing women's compliance. It comes at not surprise that computer-aided detection and diagnosis systems for

mammography have been intensively investigated since the 1980's [89], and that many authors are advocating the use of deep learning to lessen the radiologists' workload [104, 138]

The goal of this thesis is to investigate different ways in which deep learning over-reliance on large-scale datasets could be reduced, in order to make it more data efficient without sacrificing robustness. Possible solutions to tackle data starvation in the medical field are either reducing the cost of data annotation or, alternatively, lowering the amount of annotated data needed for training.

The cost of data annotation can be reduced by automatically mining reports produced by radiologists during clinical practice, which include not only free text report, but also structured reports, image annotations (e.g., bounding boxes) or lesion measurements [55, 136]. These approaches are certainly cheaper to implement than collecting ad-hoc annotations, however, they may introduce noise in the reference standard. Consequently, deep neural networks will be trained on noisy data which may affect the results, especially, in the case that the noise is frequently seen in the data. As a further matter, the evaluation will also be hindered by this fact since the test set is not curated and gold standard may not be available. Therefore, there is a need for methods that relax the requirements for annotating medical images and make the training robust against noise to be able to use larger automatically curated datasets.

On the other hand, to lower the amount of data needed for training, two methods are possible: defining networks with stronger inductive biases, that incorporate prior knowledge about the medical task at hand, or exploiting transfer learning (usually from domains/tasks in which data is abundant to domains/tasks in which data is scarce).

As an example of the former approach, it could be considered the standard screening mammography exam, in which two views are acquired per breast, namely the Cranio-Caudal (CC) and Mediolateral-Oblique (MLO). For each view, the organ is projected into a 2D image and diagnostic information is scattered in both images. The radiologist locates suspicious areas on both views by roughly triangulating from the two views. This increases diagnostic confidence as false positives due to tissue overlap are likely to disappear in the contralateral view. Few deep learning-based computer-aided diagnosis systems are able to locally combine information from the two views in a similar way as the radiologist would do. However, to achieve this goal there is a need for methods that can find a mapping between corresponding regions in both images and then merge information from both views to help diagnosis. This task is challenging since, at its best, it can be trained in a semi/weakly supervised fashion, since the breast lacks distinctive anatomical landmarks and hence a reference standard is very difficult to establish.

Finally, transfer learning has been one of the key enablers of deep learning [13], and medical imaging is no exception [109]. Transfer learning allows to harness the foundational models, trained on large scale datasets, that can be then adapted to

a variety of different tasks with minimal modifications [13].

In the medical domain, the most common approach based on current literature is to transfer learning from large-scale RGB datasets, such as ImageNet, which are commonly available in most deep learning frameworks [21]. Despite the huge domain-gap between medical images and natural images, there is an ongoing debate as to whether transfer learning from ImageNet can actually improve the performance of deep neural networks [21, 94, 51]. Some practitioners argue that using ImageNet does not necessarily boost performance, but only improves convergence [94]. In any case, transfer learning from ImageNet constrains practitioners to architectures that work well on the RGB domain, which are not necessarily optimal for high-dimensional medical imaging. For instance, many medical modalities are volumetric, and would benefit from the use of 3D convolutions [143].

Transfer learning from the medical domain would lessen these issues, but a medical dataset comparable in size and variety to ImageNet is not available. Furthermore, since medical modalities differ vastly in terms of tasks, diseases and so forth, annotation would be extremely difficult to harmonize. This leads to the possible solution which is self-supervised learning. Self-supervised methods provide the chance of learning representations without relying on labels, thus enabling to pre-train the network on larger datasets. However, given that the task that the network is trained on is synthetically generated, the representations learnt through self-supervision may not necessarily be adequate for solving clinically relevant tasks. Finding a suitable combination of datasets and self-supervision tasks would be a crucial first steps towards building foundational models for the medical domain, that can be easily transferred to a variety of different target tasks and domains.

1.1 Contributions

Based on the aforementioned problems that are slowing down utilization of deep learning for medical images, this research focus on providing and analyzing solutions to tackle these challenges. The objectives of this thesis are focused on three main pillars: (1) providing methods that makes training robust against noise, (2) incorporating information from different views of the same organ, and (3) developing multi-task, cross-domain training pipelines that exploit self-supervision to lessen reliance on annotations.

The focus of Chapter 2 is investigating the first pillar. In this research, a quantitative approach has been provided for the evaluation of noisy annotations in breast lesion detection by injecting varying degrees of noise in images, enlarging bounding boxes surrounding masses. It has been shown that labeling noise propagates through the training procedure due to imperfect matching between the network predictions and the reference standard. A novel matching criterion has been proposed to counter this effect. This activity opens new opportunities to use bookmarks that

are collected routinely in clinical practice. The proposed approach allows relaxing annotation criteria and achieve robust training of object detectors for automatic breast mass detection.

The second pillar is investigated in Chapter 3. As said, in a standard mammography study, two views are acquired per breast, the CC and the MLO. Due to the projective nature of 2D mammography, tissue superposition may both mask or mimic the presence of lesions. Therefore, integrating information from both views is paramount to increase diagnostic confidence for both radiologists and computer-aided detection systems. This emphasizes the importance of automatically matching regions from the two views. In this research, a deep convolutional neural network (CNN) for the registration of mammography images is proposed. The network is trained to predict the affine transformation that minimizes the mean squared error between the MLO and the registered CC view. However, due to the complex nature of the breast glandular pattern, deformations due to compression and the paucity of natural anatomic landmarks, optimizing the mean squared error alone yields suboptimal results. Hence, semi-supervised techniques leveraging lesion annotations are proposed.

Following on the third pillar, self-supervision has been utilized for pre-training on a large scale medical imaging dataset in Chapter 4 and has been studied from various perspectives. First, the MedNet dataset was established as an harmonized collection of 17 publicly available datasets covering various medical imaging modalities and body parts. Then, a self-supervised pre-training framework has been applied to this large scale collection, and the learned representations were evaluated by transfer learning to different computer-aided diagnosis tasks. Experimental results have shown that the extracted features are able to discriminate better than ImageNet, which is currently the de-facto standard for transfer learning in the medical domain. Further investigations showed that, however, the final performance of both pre-trained networks are similar after fine-tuning, and ImageNet marginally surpasses MedNet. The chapter thus investigates the complementary role played not only by the source and target datasets, but also by the self-supervised and target tasks, in determining the most effective strategy.

Chapter 2

Feasibility of training object detectors with noisy data

Work described in this chapter was originally presented in [34].

2.1 Introduction

Data starvation is often mentioned as one of the key obstacles to the application of deep learning in radiology [62]. Several key obstacles to large-scale data collection have been identified in the literature [62, 75]. Many technical and administrative barriers, including privacy concerns, prevent data from being easily exchanged among clinical institutions. Yet, unlike other medical specialties, radiology departments have been extensively digitized in the past years, and thousands, if not millions, of images are routinely stored in hospital picture archiving and communication systems (PACS); DICOM tags facilitate the navigation of existing archives and allow to retrospectively assemble medium-to-large scale datasets. This is not true, of course, for all branches of radiology, as certain diseases are rare thus putting intrinsic limitations to the number of samples that can be retrieved. However, for applications such as cancer screening, raw data is often abundant.

Images alone, however, are not sufficient to train a deep neural network for Computer Aided Detection/Diagnosis (CAD) systems. High quality annotations must be included to identify the disease status of a subject or a candidate lesion according to an established reference standard, which should be as accurate as possible. The reference standard is usually determined from an independent device or procedure, that is often recorded in a separate data silo. In breast imaging, the reference standard is typically established based on biopsy or histopathology for the presence of cancer, whereas for negative cases it is generally established by biopsy or follow-up over a period of two years or longer. Although CAD systems can be trained from case-level only, a lesion-level reference standard is usually established

by marking lesions on the image. The availability of lesion-level annotations, besides being useful in the assessment of CAD performance, was also shown to lead to higher performance [27]. Hence, there is an interest in reducing the time and effort needed to collect such annotations at scale.

The starting point for this research is that information about lesion location and characteristics is routinely collected by radiologists in PACS and reading systems in the form of text reports, various types of bookmarks, and lesion measurements [62, 136, 85]. This information could be used to automatically determine an approximate reference standard at limited additional cost. However, compared to human-annotated datasets, quality usually comes at the expense of quantity, and it is important to assess the potential impact on the performance of deep neural networks trained on an imperfect reference standard [55, 136, 123]. It is expected that the increasing adoption of structured reporting by radiologists will further facilitate this practice in the future [62].

Some authors have used text mining to automatically extract labels from free-text reports [55]. Unfortunately, only patient-level labels can normally be obtained in this way. Besides text reports, reporting workstations typically provide drawing tools such as bounding boxes (ellipses or squares), arrows, lines, or diameters that radiologists can use to mark and measure specific lesions [136, 85]. A study conducted at the NIH Clinical Center found that the number of CT scans with such bookmarks jumped after 2015. Bookmarks were frequently presented as ellipses (8.4%) or lesion diameters (46%) [136]. The recently published DeepLesion dataset, which contains over 32,000 lesions identified on CT images using diameter measurements, demonstrates the potential of this approach [136]. Such annotations can be used to train object detection networks that enable both detection and localization of lesions compared to image-level classifiers [102, 136].

Mining strategies are attractive, but inevitably lead to some degree of noise in the reference standard [39, 58]. For example, it is not necessary for all lesions mentioned in the report to be explicitly annotated [136], and reports from individual radiologists may suffer from large inter-rater variability [108, 7, 65]. Annotations collected to train and test machine learning models are usually recorded using two- or three-dimensional bounding boxes drawn as close as possible to the lesion boundaries [85] or by segmenting the lesion [7]. Based on the experience gained in other studies focusing on the conduction of large-scale breast screening trials [6], bookmarks recorded in clinical practice need not be as precise and may serve purposes other than annotation of the lesion (e.g., identification of the area selected for biopsy or further workup).

The present work links the problem of lesion detection using automatically mined annotations, collected from clinical practice, to the effect of noisy annotations on the generalization error of deep neural networks and, in particular, object detection networks. In this context, previous studies have focused on changes in reference standard labels (e.g., missing or mislabeled objects). Here, a different and

independent source of noise is investigated, i.e., that resulting from loosely labeled lesions (i.e., bounding boxes are approximate and include both the lesion and part of the background). As it will be shown in the remainder of this chapter through controlled experiments with increasing levels of noise, this source of noise can affect recognition performance.

Many different architectures have been proposed for object detection [99, 74, 97] and successfully applied in breast imaging [102, 133, 1, 17, 57, 79]. However, all architectures share some common operating principles: they contain one or more classification modules that classify regions of interest (ROIs), denoted by a bounding box, as either background or one of the possible object classes (lesion). These modules are trained by selecting examples of bounding boxes labeled as foreground (true positive examples) or background (true negative examples).

Distinguishing between positive and negative examples is achieved by comparing ROIs with the bounding boxes of the reference standard based on a matching criterion: usually, an Intersection over Union (IoU) threshold is set to determine whether two boxes match. If the reference standard is inaccurate (e.g., if the bounding boxes are larger than the actual object), the match may be incorrect. The classification modules are then trained on noisy labels and recognition performance may suffer. It is this phenomenon that the work reported here attempts to quantify and characterize here.

The work seeks to assess the effect of labelling noise, in the form of loosely annotated lesions, on standard object detectors, and introduce ways in which they can be made more robust to the presence of noise. In summary, the main contributions of the reported research are as follows:

- a noise model is proposed to simulate imperfect reference standards from clean labels available in the CBIS-DDSM (Curated Breast Imaging Subset of DDSM) dataset, a public, well curated screen-film mammography dataset for which lesion pixel-level segmentations are available. This noise model is used to simulate the presence of different levels of noise;

- it is shown that the use of imperfect bounding boxes has an impact on the performance on the classifier module of standard object detectors, and that this impact is mediated by a critical hyperparameter, which is the choice of the matching criterion used to compare the generated ROIs with the reference standard. A new criterion, named Exp_IoU, is proposed to make the network robust in the presence of high levels of noise, favoring examples that are closer to the center of the reference standard bounding box and, thus, more likely to contain the actual lesion;

- it is demonstrated that standard object detectors, such as Faster R-CNN [99], are quite robust in the presence of low to moderate amount of noise. In the

presence of moderate to large noise, a simple yet effective countermeasure consists in the use of alternative matching criteria.

Faster R-CNN is based on an approach called Regions with Convolutional Neural Networks, abbreviated as R-CNN [40]. The idea behind this approach is to find regions, commonly known as region proposals, that might contain an object. Then CNN features are extracted from region proposals to be classified by a classifier. We focus on a variation of R-CNN, known as *Faster R-CNN*, which, as the name suggests, was designed to process region proposals more efficiently. The rest of the chapter is organized as follows. Related works and background on Faster R-CNN are presented in Section 2.2. The noise model, network training, and matching criteria used in this work are illustrated in Section 2.3. Experimental setup and results are reported in Section 2.4, and discussed in Section 2.5.

2.2 Background and related work

2.2.1 Object detection and the Faster R-CNN architecture

Object detectors are deep neural networks which classify and localize object instances by drawing a bounding boxes around them [139]. In radiology, object detectors are trained to identify lesion candidate regions for computer-aided diagnosis. Modern object detectors operate by generating hundreds of potential regions of interest (ROIs), called anchors or anchor boxes, at different locations and with different aspect ratios. In order to do so, a sliding window is passed through the feature map and, at each location, k proposals with different scales and aspect ratios, known as anchor boxes, are generated (a typical value for k is 9). Each ROI is then independently classified as either background or one of the object (or, as in this case, lesion) classes, and only those with high probability of being an actual object/lesion are shown to the radiologist. Anchors can be generated based on a fixed grid (one-stage detector) or by employing a pre-selection mechanism (two-stage detectors). Faster R-CNN [99] is the most common two-stage object detector, whereas single-stage architecture include YOLO [97] and RetinaNet [74].

It is interesting to have a close look at Faster R-CNN, which will be used in various experiments reported later. It consists of two modules operating in cascade: the Region Proposal Network (RPN) and the detector. The two modules share a convolutional backbone that performs feature extraction, and that is typically pre-trained on ImageNet or another large-scale classification dataset. Inside each module, the features computed by the backbone are then input to *regression head*, which regresses the bounding box coordinates, and a *classification head*, which predicts the final class. The RPN performs a binary classification task (background vs. not-background) and outputs an "objectness" score, i.e., the probability of a given anchor actually containing an object. The ROIs with highest objectness score

are then passed on to the classifier, which is thus trained on a more balanced dataset (in a typical image, the background class is largely over-represented).

The RPN and the classifier module tend to predict multiple bounding boxes for the same object. An algorithm called Non-Maximum Suppression (NMS) is used to reduce the number of correlated ROIs: in synthesis, for each group of overlapping bounding boxes, only the one with the highest classification score is retained, discarding the others.

Training of the two modules is performed jointly in an alternating fashion. A single batch is constructed by selecting all anchor boxes within one image. For each batch, the RPN is trained first, then the detector is updated keeping the output of the RPN fixed. Since each module outputs both the bounding box parameters and the anchor class, the loss is defined for both modules as a combination of regression and classification losses [99]:

$$L(\{p_i\},\{b_i\}) = \frac{1}{n_r}\sum_i L_{cls}(p_i, p_i') + \lambda\frac{1}{n_c}\sum_i L_{reg}(b_i, b_i') \tag{2.1}$$

where L_{reg} is the smooth L1 loss for regression, L_{cls} is the categorical cross entropy, b_i is the i^{th} reference standard bounding box, b_i' is the predicted bounding box, p_i' is the predicted probability that b_i contains a lesion, and p_i is the reference standard label. The same loss is used for both the RPN and the detector, but the hyperparameters are optimized separately.

A critical step during training is labeling each anchor and/or ROI as foreground (positive examples) or background (negative examples). Since the anchors/ROIs are generated from the input grid and aspect ratios (for the RPN) or from the RPN output (for the classifier), none precisely coincides with the lesions bounding boxes in the reference standard. Therefore, a matching criterion is needed to compare the predicted bounding boxes with the reference standard based on some notion of overlap. In some training settings, ROIs which marginally overlap with the reference standard as classified as neutral examples, meaning that they should not be considered neither positive nor negative examples; neutral examples, for instance, may occur along the boundaries of an object/lesion. In the standard computer vision literature, the most common criterion for performing this matching is the IoU. As will be shown in Section 2.3.3, in the presence of labeling noise, the selection of a robust matching criterion becomes paramount to ensure proper convergence.

2.2.2 Labeling noise in deep neural networks

Real-world datasets are often affected by noise, which can manifest itself in different forms. In this context, the term noise is specifically referred to labeling noise, which implies that the labels are imperfect or corrupt; in this thesis, noise that may corrupt the network input is not considered. Depending on the intensity and

structure of the noise, a machine learning model may be unable to learn valid and generalizing patterns from the input labels. The term label noise in the literature has been used to encompass a broad range of imperfect or corrupted labels [58]. In medical imaging, it can be difficult to extract accurate and noise-free labels, due to the high inter-observer variability [108, 84], but also due to the fact that pathological signs are often ambiguous in nature, making it challenging to define a clear-cut boundary between normal and abnormal anatomy [58, 108]. Noise can be further amplified if approximate labels are semi-automatically extracted by, e.g., text mining.

Several works in the literature have studied the impact of noise on classification tasks [39, 37, 98, 36, 49]. Given a training set formed by samples $\{x_i, y_i\}$, where y_i is a discrete variable that corresponds to the true class of the sample, labelling noise can be formally defined as a stochastic process which pollutes the labels that are passed to the machine learning algorithm [36]. As a result, the sample label may not reflect the true class of the sample.

Noise in the machine learning field has been traditionally classified as uniform or structured [37, 36]. Uniform noise is typically random and has an equal probability of affecting each given sample. In the case of structured noise, label corruption depends either on the class label (class-dependent noise) or on the input feature (feature-dependent noise), implying that certain inputs have a higher probability of being affected by noise. For example, in the medical domain, negative or borderline results may be affected by higher noise since a biopsy, which would provide a more accurate reference standard, is not performed. It is intuitive to see that machine learning algorithms struggle to recover from structured noise more than from uniform noise. The most common models for uniform and structured noise studied in the literature are label flips and outliers [122]. Label flips refer to samples that have been assigned a wrong class label (the choice of the samples will dictate whether the noise is uniform or structured), whereas outliers are samples that do not belong to any of the classes in the training set. In this thesis, adversarial noise is not considered; the latter consists in intentionally manipulating input data with very small perturbations specifically designed to alter the output of a machine learning model, which leads to sample misclassification [35].

Various works have sought to investigate the effect of labeling noise on the performance of machine learning models, either from a theoretical or experimental standpoint, and have often reached divergent conclusions depending on the noise regime. Based on theoretical results presented in [87], a high capacity model should be robust to several types of random (or uniform) noise, provided that a sufficiently large training set is available. Practically, deep learning methods have indeed shown resilience to labeling noise, provided that a sufficiently large number of clean labels are available. For example, experiments in MNIST in [105] showed that with a 10:1 ratio of noisy to clean labels, at least 2000 clean labels are needed to reach an accuracy of 90%, whereas for a ratio of 50:1, the number of clean labels required

increases to 10,000 to achieve the same performance. However, in practice, labeling noise may follow complex, class-dependent patterns, and the number of clean samples is often limited. Due to their memorization effect, deep neural networks, sooner or later, begin to memorize noisy labeled samples [141, 45].

The effect of labeling noise on more complex models, such as object detector or segmentation models, is less explored in the literature, especially in the medical domain. Furthermore, there are intrinsic differences between general object detection and lesion detection in the medical domain: for instance, lesions have much more ill-defined margins, and thus it is more difficult to obtain precise segmentation or bounding boxes. Lesions are also more rare, thus leading to an extremely high-class imbalance.

2.2.3 Breast mass detection

Deep learning techniques have been extensively investigated in the context of breast imaging; its performance have substantially surpassed those of conventional computer-aided detection systems [89, 84], and in some cases rivalled those of human experts in laboratory settings [64, 70, 102, 133, 113].

Deep learning-based approaches in mammography analysis fall under two main categories, mostly depending on whether the system outputs case-level or lesion-level predictions. Multi-stream CNNs can predict the presence or absence of malignant findings from multi-view mammography images [133, 80]. Such systems can be trained from case-level annotations only, although benefits are still expected when integrating some form of patch-level supervision. Due to their complex structure and their ability to combine information from multiple views, they usually achieve better performance compared to single-view architectures. Some studies have investigated whether such deep learning-based systems can be used in screening programs to reduce radiologists' workload without negatively affecting clinical outcomes [96, 26]. Some of these systems can be also used independently to automatically detect breast cancer in two-dimensional mammograms, with a performance level in laboratory conditions comparable with that of radiologists [104, 110].

Training deep learning networks from case-level labels only requires extremely large-scale and enriched datasets, including hundreds of thousands of images and thousands of cancer examples. As an example, the (private) NYU Breast Cancer Screening Dataset contains roughly 186,000 women [134]. Collecting such large datasets may take years. Furthermore, such deep learning-based systems may fail to generalize to a new population. In fact, in an independent replication study [25], the performance of multiple DNNs trained on the NYU dataset dropped significantly, and local retraining was required to mitigate the drop.

For these reasons, there is still interest in complementing case-level predictions with other architectures, such as object detectors, that are trained on lesions-level

annotations, e.g., bounding boxes or segmentation masks [102, 17, 80, 113, 1, 57, 79]. From the radiologist's perspective, providing precise localization of cancerous lesions is advantageous when a system is used as a CAD tool. From a machine learning point of view, by exploiting information on the location of the injury during training, object detectors also have the potential to exploit training data more effectively [113]. Indeed, the DREAM challenge [27], in which only case-level labels were provided, was won by the Faster R-CNN model by Ribli et al. [102, 113], which leveraged external data with bounding box annotations, including public and private sources. Thus, acquiring lesion-level annotations at scale and with a reduced cost is still relevant as it may substantially reduce the amount of data needed to train a deep learning-based system or adapt it to a local population. The work presented in this chapter aims at establishing whether annotation requirements can be relaxed, paving the way for future crowd-sourcing of annotations.

Several architectures for object detection were proposed. Compared to general-purpose object detection benchmarks, medical applications place greater emphasis on accuracy than execution speed because real-time performance is usually not needed. This is evident in mammography applications which employed architectures renowned for their accuracy, such as Faster R-CNN [102, 17, 113, 1] and RetinaNet [80, 57], with few exceptions based on YOLO [79]. In the public InBreast dataset [83], solutions based on Faster R-CNN outperformed those based on RetinaNet, achieving 92% sensitivity at 0.3 False Positive (FP)/image [1]. However, many differences in the experimental settings hinder direct comparison between different studies: for instance, since the InBreast dataset is small, research groups relied on different public and private digital mammography datasets in order to build the training sets. In any case, since both Faster R-CNN and RetinaNet employ similar criteria for matching predicted ROIs with the reference standard, the techniques described in this chapter could be applied to RetinaNet as well.

For the sake of reproducibility, in this research the publicly available CBIS-DDSM dataset was exploited, choosing the Faster R-CNN architecture based on the above considerations. Hyperparameters of the Faster R-CNN architecture were tailored to the specific task, also using information available from previous studies [102, 16]. In the experiments reported in this chapter, Faster R-CNN showed consistent signs of overfitting in the CBIS-DDSM data set, which is in agreement with previous experiments by [17]. The experiments presented in this chapter show that the way that ROIs are selected during training can affect generalization; in particular, the use of hard sample mining was found to increase performance (Section 2.3.4).

2.2.4 Metrics and losses for bounding box regression

A common problem of object detection and CAD systems is how to evaluate whether, and to what extent, an object proposal corresponds to a given bounding

box in the reference standard. The IoU is the most common matching metric in object detection [99, 97, 74, 101]. Its value, bounded between 0 (no overlap) and 1 (complete overlap), measures how tightly the detector output fits the reference standard bounding boxes.

Variants of the IoU have been proposed to generate more robust regression losses that, unlike the commonly used smooth l_p-norms, directly maximize the IoU metric [101, 142]. Since the IoU is only defined when two bounding boxes overlap, and thus does not provide any gradient for the non-overlapping cases, it cannot be used directly as a regression loss. Such variants include the Generalized Intersection over Union (GIoU) [101] and the Distance-IoU [142]. The former changes the mathematical formulation of the GIoU in such a way that, in the case of non-overlapping bounding boxes, yields an approximate distance between the two bounding boxes. The latter combines the IoU with the Euclidean distance between the central points of the proposal and reference standard bounding boxes. Both metrics were used to define regression losses that could be incorporated into existing architectures, with performance gains especially for fast architectures such as YOLO [142, 101].

In the present research, the focus is instead of understanding the impact of systematic biases in the reference standard, and how they may affect the training of the classifier head, rather than improve the training of the regression head. It is expected that, in any case, systematic biases will be reflected in the trained regression parameters. In this respect, the focus is shifted from the regression loss to the matching criterion used to label ROIs during training. The GIoU or Distance-IoU were mainly used as a substitute loss, while the IoU metric was still used as the matching criterion [142, 101]. Specifically, a IoU greater than a certain threshold, such as 0.5 [74] or 0.7 [99] denotes a foreground object, assuming that the reference standard was noise-free.

When a bounding box proposal does not overlap with the reference standard, it can be safely labeled as background. Hence, the GIoU metric [101] is not particularly relevant in this context, since it is not required or useful to yield a non-zero response in the case of non-overlapping bounding boxes, whereas for those who indeed overlap the GIoU has the same formulation of the IoU. On the other hand, in the presence of noise the number of proposals that will overlap with the reference standard is expected to increase significantly and, as shown in the rest of this chapter, matching criteria that incorporate both distance and overlap measures (e.g., [142]) allow for a better differentiation between competing overlapping proposals. In [142], however, this aspect was not investigated.

The present work also draws inspiration from the CAD evaluation field. Since the performance of a CAD system depends not only on whether it can correctly identify a diseased case, but also on whether it can correctly locate the abnormality in the images, recommendations are to pay specific attention to the matching criterion used [92]. The matching rules define the required level of correspondence between the CAD marks and the location/extent in the reference standard. A

mark is considered a True Positive (TP) if it meets the mark-labeling rule, and a False Positive (FP) otherwise. While in the field of object detection the IoU is the overwhelmingly popular choice, the choice is much more nuanced in the CAD literature.

A wide range of matching criteria has been used by CAD researchers. They are usually defined based on a threshold over the area overlap (either fixed or dependent on the sum of the areas, as in the IoU), a threshold on the distance between the centroid of the CAD mark and the annotated reference standard (either fixed or depending on the size of the reference standard), or whether the center of the mark is within the annotated abnormality (or viceversa). Sometimes two or more criteria may be combined. The choice depends on the type of lesions, on the type of graphical marks used to represent the CAD output, and on how the reference standard is recorded.

Compared to visual evaluation by a radiologist, an objective, rule-based matching criterion is generally deemed consistent and reproducible. However, the matching rule should be carefully selected to ensure that computer/reading device markings are consistent with clinical interpretations. Marks labeled as TP which are unlikely to draw the reader's attention to the abnormality, are of particular concern because they may lead to inflated estimates of stand-alone CAD performance [92]. For example, a large bounding box covering a substantial portion of the breast has a higher probability of overlapping with a reference standard patch, but the CAD system may have actually missed the true lesion. In general, the areas of the TP markers are expected to be comparable to the lesion sizes determined by the reference standard. Unfortunately, the effect of the matching criterion on performance estimate is often under-estimated, and much of the CAD literature does not describe the marking rules used [92].

2.3 Materials and methods

This section is dedicated to explaining the training framework and points out the issue with noisy bounding boxes. As shown in Figure 2.1, Faster R-CNN is made up of convolutional layers to provide feature maps, a region proposal network, and a classifier. The images alongside localization and classification bounding boxes (i.e., the annotations) are passed to the network for training. However, noisy annotations (i.e. not tight to the lesion) will result in labeling noise when the proposals are generated as discussed earlier in Section 2.2.4. The red crossed circle shows where the noise is created. As shown with a red arrow, backpropagation with noisy labels will misguide the training and affects the final prediction in practice. Therefore, a framework has been devised for analysing the scale of noisy labels and examining the networks performance given different levels of noise.

Section 2.3.1 is dedicated to explaining the dataset used in the experiments. The

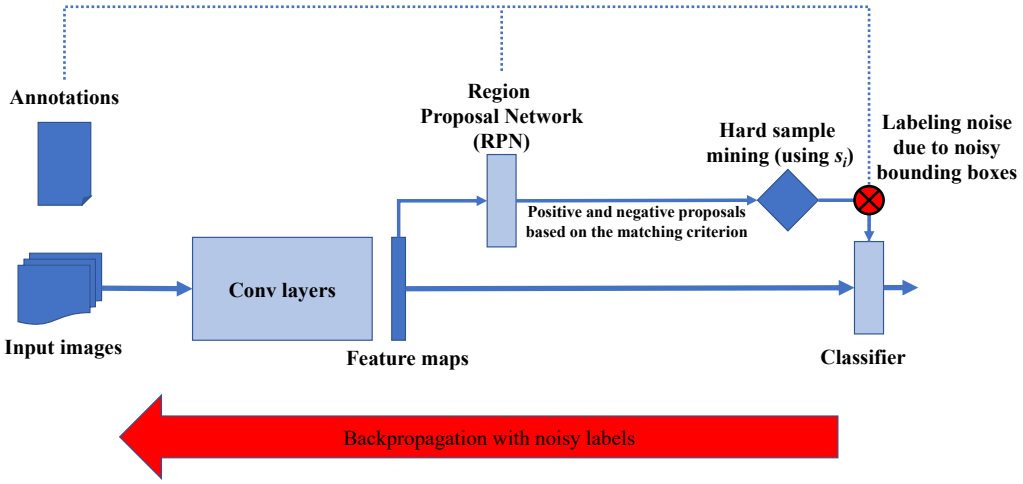


Figure 2.1: The training framework of Faster R-CNN is summarized in this figure. The noise is injected onto the bounding boxes in the annotations and its impact on the training is highlighted by the red elements. Specifically, the red circle shows where the labeling noise is generated. Faster R-CNN is a two-stage object detector: the RPN filters out candidate regions from the background, while the classifier assigns them the most likely class. Since lesions are rare compared to background, a hard-sampling procedure was added to the framework to avoid overfitting and ensure that more informative region proposals are passed to the classifier. If the bounding boxes are not tightly drawn around the lesion borders, the region proposal that are passed to the classifier may be mislabelled during the training procedure. This is because the region proposals are automatically compared to the reference standard using a matching criterion, such as the IoU, which can lead to incorrect results if the bounding boxes do not match the lesion borders exactly. Figure reproduced from [34].

subset that has been used in the experiments is free of noise, therefore, a noise model has been utilized for a systematic comparison between different noise levels. The noise model has been explained in Section 2.3.2. Next, the shortcomings of widely used IoU has been explained and a new criterion for matching bounding boxes with the reference standard has been proposed to counter the effects of labeling noise in section 2.3.3. Then, in Sections 2.3.4, 2.3.5, and 2.3.6 the network structure alongside the use of hard sample mining, hyper-parameters setup, and evaluation metrics are discussed accordingly.

2.3.1 Dataset

The CBIS-DDSM collection [68, 69, 24] is a standardized version of part of the DDSM dataset [48], selected and curated by a trained mammographer. It is publicly available to download from the Cancer Imaging Archive (TCIA). DDSM is composed of 2620 scanned screen film mammography studies, including normal, benign, and malignant findings with verified pathology information. Each study contains up to four images acquired in the cranio-caudal (CC) and medio-lateral oblique (MLO) orientations; however, only images with findings are included in the CBIS-DDSM dataset, for a total of 3,089 images. The CBIS-DDSM database, detailed in [69], was developed and converted to DICOM using standard techniques. Furthermore, for reducing the computation time the breast region was cropped before handling it using an automated algorithm [85]. The present thesis focuses only on the differences in mass discoveries, with a carefully segmentation by the radiologists with experience in the CBIS-DDSM collection. The reference standard bonding boxes are assumed as a strictly tight box that covers the entire mass in the segmentation. So it can be assumed that the first boxes are very accurate and soundless. For microcalcification clusters only a coarse segmentation is provided. Therefore, it was assumed that the initial bounding boxes are clean and free of noise. The standard training/test split (80%/20%) defined by the CBIS-DDSM authors was used. The final training and test set included 550 patients (with 613 masses) and 200 patients (with 222 masses), respectively. For each patient, up to four images are available and, since each finding is in most cases visible in both CC and MLO views, the total number of lesion views is 1,316 and 374 for the training and test set, respectively.

2.3.2 Noise modeling

Analysing the effects of noise on the training procedure of the network requires a comparison between a model that has been trained on the clean and noisy datasets. Here, the bounding boxes are enlarged based on a random noise model. The reference standard bounding box $b_i = (x_{1i}, y_{1i}, x_{2i}, y_{2i})$ for lesion $i \in \{1,2,...,m\}$ which perfectly fits the boundaries of the lesion is enlarged by a random factor as shown below:

$$\begin{aligned} w_i' &= (1 + n_{wi})w_i, \\ h_i' &= (1 + n_{hi})h_i. \end{aligned} \tag{2.2}$$

where (w_i, h_i) are the width and height of b_i, (w_i', h_i') are the width and height of b_i', and (n_{wi}, n_{hi}) are sampled from a normal distribution $n_{wi}, n_{hi} \sim \mathcal{N}(\mu, 1)$ with mean μ. Bounding boxes are defined in pixels in the reference standard, whereas μ is defined as a dimensionless multiplicative noise.

The purpose of the bounding box is to roughly localize the lesion, therefore b'_i is assumed to be equal or larger than the original clean bounding box. It should be noted that extremely large bounding boxes are unlikely to happen, hence, n_{wi} and n_{hi} are clipped in the range [0,6), meaning that the bounding box is at most six times larger than the original one after noise injection.

Despite this limit, since the typical size of mammography masses ranges between 1 and 3 cm, the resulting bounding box may still exceed the breast region, which is unrealistic: based on the size of the largest lesions in CBIS-DDSM, b'_i was truncated to be at most 80% of the total image width. The center of b'_i is as same as b_i, except in those cases where the bounding box has been cropped to fit within these limits.

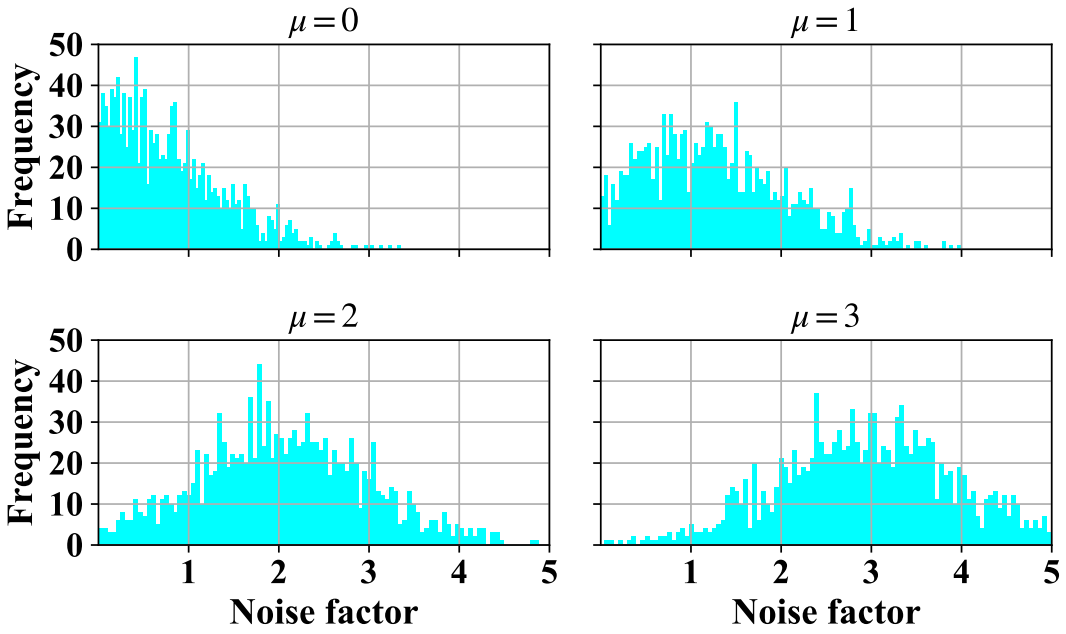


Figure 2.2: Histograms of the noise factor n_{wi} from $\mu = 0$ to $\mu = 3$. All distributions are clipped to the range [0, 5]. Figure reproduced from [34].

Four different levels of noise were generated with $\mu = \{0,1,2,3\}$. Histograms of n_{wi} and n_{hi} drawn from the model are depicted in Figure 2.2, whereas the distributions of the bounding boxes diameters at different levels of noise are compared in Figure 2.3.

2.3.3 Matching criterion

Following section 2.2.4, a matching criterion measures the degree of matching between the reference standard bounding boxes and the proposed boxes to distinguish between TP and FP during training and testing. The reference standard

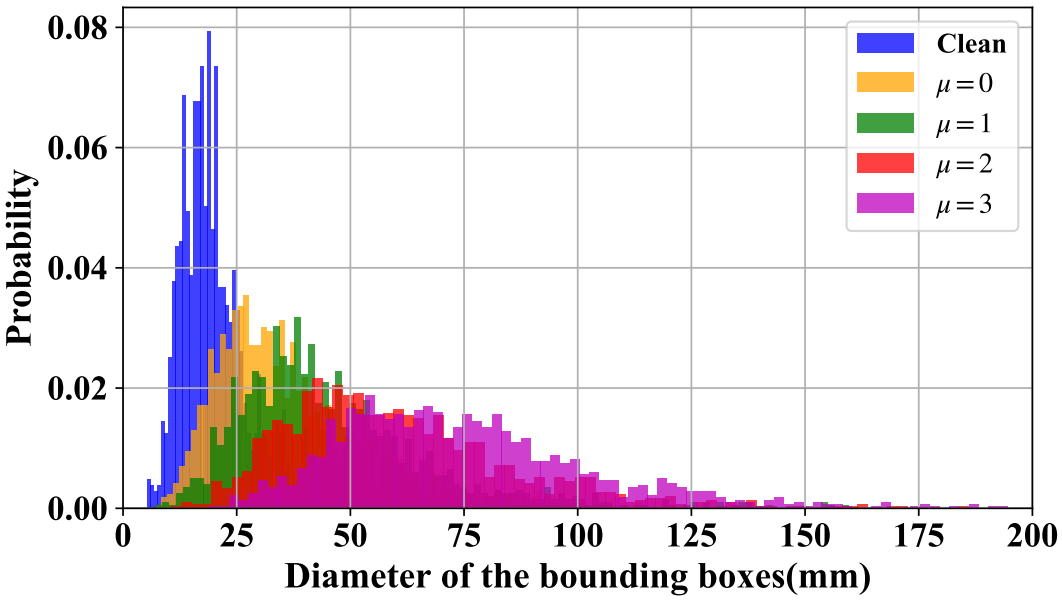


Figure 2.3: Diameter distribution (histogram) of the noisy bounding boxes for the clean dataset and for each level of noise. Figure reproduced from [34].

bounding boxes are provided manually by experts. For example, it helps to recognize the labels the proposed bounding boxes given by the RPN which are later passed to the classifier during training. Three different matching criterion were included in the experiments, extracted from recognized papers in CAD evaluation and object detection: the IoU, which is the *de facto* standard in object detection, a simple Centroid criterion, and a combination of distance and overlap, which is denoted as Exp_IoU in the following.

In some cases such as IoU, a threshold is used to distinguish between positive and negative anchor boxes. Anchor boxes with a higher IoU than threshold T_u are considered as positives, while negative bounding boxes are those with an IoU score lower than a second threshold T_l. Any bounding box with an IoU score in between the aforementioned thresholds is considered as neutral examples which are ignored during training. If no bounding box can be selected as positive, which rarely happens,the anchor box with the highest IoU will be chosen and labeled as positive [99]. Following the works that have been done by Ribli et. al. [102], T_u was decreased from 0.7 to 0.5 in order to allow more positive examples in each batch. Higher thresholds have not been experimented on, since, they may lead to unstable training [102]. T_l is instead equal to 0.3 [99]. T_l is instead equal to 0.3 [99].

The Centroid criterion or "centroid inside the bounding box" simply checks whether the center of the proposed bounding box falls inside the reference standard

bounding box. This is a common criteria for evaluating CAD systems, but has never been used for training [102].

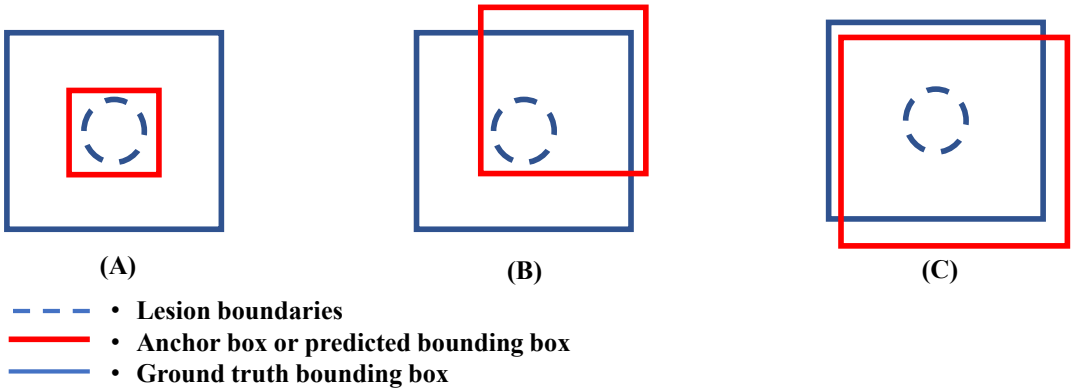


Figure 2.4: The red boxes show three different bounding box proposals for a lesion. If IoU is used as the matching criterion, the proposals would be scored as $C > B > A$, where the highest score indicates higher similarity to the reference standard. However, based on the distance between centroids, the order would be $A > C > B$. Figure reproduced from [34].

Figure 2.4 depicts how overlap and distance-based metrics can provide different rankings for a given bounding box. This difference later affects how the bonding boxes are labeled. In other terms, the supervised labels for the object detector are dependent on the matching criterion that has been used for training the network. Each criterion has its own strengths and shortcomings: IoU favors bounding boxes that match in both size and position, but as shown later, may fail in the presence of noise (e. g. in Figure 2.4-A the bounding box perfectly covers the lesion but the IoU will be lower because the size of the bounding boxes do not match). Utilizing "centroid inside the bounding box" which is confirmed by experimental results in Section 2.4. Another reason may be the fact that this criterion does not provide insight on how closely two bounding boxes match its either positive or negative as apposed to scores IoU which returns the degree of matching.

Hence, in this thesis, a new criterion is proposed, denoted Exp_IoU, which explicitly considers both the size and relative position of the bounding boxes, as follows:

$$S_{exp\_iou}(b'_i, b_j) = \frac{IoU(b'_i, b_j) + e^{-\beta D(b'_i, b_j)}}{2} \tag{2.3}$$

where b'_i is the proposed bounding/anchor box, b_j corresponds to the reference standard, $D(.)$ represents the Euclidean distance between the centers of each bounding box, and β balances the two contributions. The value of β was set to 0.1 based on experimental results. The same thresholds T_u and T_l are used for both IoU and Exp_IoU.

2.3.4 Deep network architecture

In this section, the choice of hyper-parameters selection and the steps that has been taken toward improving the object detector are discussed. As it has already been mentioned, the framework used in this work is based the Faster R-CNN. Further details about the original Faster-RCNN has been explained earlier in Section 2.2.1.

A crucial part of the Faster-RCNN framework is the choice backbone structure, since the backbone is where the feature map is created and the rest of the components are built upon it. Based on prior investigations [102], a better result can be achieved, using the ResNet50 in comparison to VGG16. Layers up to conv4_x are included in the backbone and the top layers (conv5_x) are https://www.overleaf.com/project/60dd6b497a9716d2706c697fincluded in the RPN and the classifier head.

At each location nine anchor boxes will be generated. The anchor box scales are $\{128, 256, 512\}$, aspect ratios are $\{(1,1), (0.7, 1.4), (1.4, 0.7)\}$, and the stride is 16. The boundaries of lesions in mammography is much less defined than usual object detection task such as objects in a natural scene [102]. Therefore, the NMS overlap threshold has been set much lower in comparison to the original Faster-RCNN paper. For training, the threshold is 0.7 and for testing the threshold is 0.1. The generated bounding boxes have been limited by the NMS to 300 for both train and test. The value of λ is set to 8.3 for the RPN loss, and 12.5 for the detector.

Given the specifications of the NMS, the RPN provides 300 bounding boxes per image. The number of mass lesions in a mammogram is usually low (usually one or two), consequently, the portion of positive examples out of 300 in total will be drastically lower in comparison to the negative ones. Even considering images with at least one lesion, the number of positive examples generated by the RPN ranges between 2 to 10. This results in a severely imbalanced examples fed into the classifier. A possible strategy is to randomly balancing the dataset by choosing four random ROIs (half positive, half negative). This still may hinder the training, resulting in non-robust performance that can be caused by poor selection of samples.

On the other hand, more informative samples can be mined after the NMS to keep the diversity based on classification difficulty. One of the main difficulties in mammography is the discrimination of lesions from glandular patterns that mimic their presence. Intuitively, the misclassified samples by the RPN are more difficult for the network to learn, therefore are assumed to be hard samples. This heuristic

can be used to improve the performance of the network. Therefore, a score was used to measure how difficult a sample is for the network, and ranked all the proposed ROIs as followed:

$$s_i = (p_i^{'} - p_i)^2 \tag{2.4}$$

where $p_i^{'}$ is the predicted probability that b_i contains a lesion, and p_i is the reference standard label. As s_i increases, the margin between the probability and the true label grows, meaning that it is a hard sample. In the presence of labeling noise, the margin s_i is also expected to be higher for noisy samples than clean ones [39].

The positive and negative samples are sorted separately based on their score s_i, and the corresponding mean scores S_P and S_N are calculated. The samples can then be split into four categories:

- easy positive: positive samples with $s_i < S_P$;
- hard positive: positive samples with $s_i \geq S_P$;
- easy negative: negative samples with $s_i < S_N$;
- hard negative: negative samples with $s_i \geq S_N$.

From each category, 25 positive samples and 25 negative samples are selected in order to maintain a balance between difficult and easy examples, as well as noisy and clean ones. Finally, four ROIs are randomly sampled from this subset.

2.3.5 Experimental setup

A preliminary experiment has been done to illustrate the advantage of hard negative sampling strategy compared to the baseline model. The matching criterion used in both experiments is the IOU to only study the effect of hard sample mining. Next, after the superiority of the hard sample mining has been shown it has been used as the new baseline in the performed experiments. Based on the fact that there are three different matching criterion and five different versions of the dataset (ranging from clean to extremely noisy reference standard), a total of 15 different possible experiments has been taken into consideration to evaluate and compare the performance of the proposed methodology. All the experiments use the same hyper-parameterization as explained in Section 2.3.3. The backbone is initialized by weights pre-trained on ImageNet.

Each network was trained for 80 epochs (were each epoch is made of 500 iterations) with the choice of Adam [61] as optimizer and the learning rate of 10^{-5}. The learning rate is a scalar value which controls how quickly the model converges to a local or global optima. Experimentally, going over 80 epochs has resulted in overfitting. In order to reduce the computational effort, the images were downsampled

in such a way that the largest dimension was equal to 600 pixels despite the fact that better results could be obtained with higher resolution images [102].

A constant seed was set for TensorFlow and NumPy [1] to minimize the variability between experiments. The dataset has been shuffeled but fixed for all experiments. The network was implemented in Keras 2.2 with Tensorflow 1.13.1. All experiments were conducted on an Nvidia Titan Xp GPU.

2.3.6 Evaluation

For the sake of comparability, the test set is free of noise for all the experiments. The evaluation method used in the experiments is based on the Free-Response ROC (FROC) which plots sensitivity against the average number of False positives (FPs) per image [10, 92]. The FROC is a superior in object detection tasks since it is a location-specific variant of ROC analysis where the number of detections per image is not constrained, and each detection can be assigned a separate score by the CAD algorithm. Note that FROC curve is also a widely accepted methodology in the medical image analysis literature [92].

To evaluate whether or not a bounding box matches with the reference standard while testing, the centroid inside the bounding box criterion was used. This method has been commonly used for evaluating CAD systems and also employed in similar works [102]. The fact that lesion boundaries are not as well-defined as boundaries in objects in traditional images alongside the sparsity of lesions existing in an image makes tightly matching bounding boxes with reference standard less relevant as long as the lesion is clearly shown to the radiologist. Moreover, this is a fair choice for comparing the performance at different noise levels. The reason is that comparing the network predictions with noisy bounding boxes against the clean reference standard with IoU would significantly drop the performance, given that the bounding box is larger. Hence, the focus here is on whether each network can correctly locate the true lesion or not.

The FROC curves were computed on 1000 bootstrap samples, each containing 200 cases (which is the size of the validation set). All FROC curves were cut at 2 FPs/image, as higher false positive rates are less useful from a clinical view point. The area under the FROC (AFROC) was used as a summary measure to compare experiments [10]. Previous research demonstrated that the AFROC penalizes the number of erroneous marks, rewarded for the fraction of detected abnormalities, and adjusted for the effect of the target size [10]. Geometrically, it can be interpreted as a measure of average performance superiority over an artificial "guessing" free-response process and it represents an analogy to the area between the ROC curve

[1] Open source libraries for developing and training deep neural networks and tensor manipulation.

and the diagonal line. The AFROC was proved to be correlated with the AROC, as the empirical FROC curve can be interpreted as a scaled ROC curve under the assumption of independence of the rated marks within a subject [140]. Confidence intervals were also calculated by bootstrapping.

2.4 Results

In this section, the results obtained will be introduced and discussed in details. Before delving into the effect of noise, the proposed training procedure, and specifically the hard negative mining, was validated on the clean dataset. As depicted in Figure 2.5, the proposed hard sample mining strategies improve the FROC curve, with an +0.14 increase in AFROC (from 1.03 to 1.17) and a +0.15 increase in sensitivity at a fixed false positive rate of 0.5 FPs/images. Cha et al [16] reported a sensitivity of 80% at 2 FPs/image, which is similar to results obtained in Figure 2.5. In both [16] and the experiments performed in the present research, significant overfitting was observed, as will be detailed further in Chapter 3.

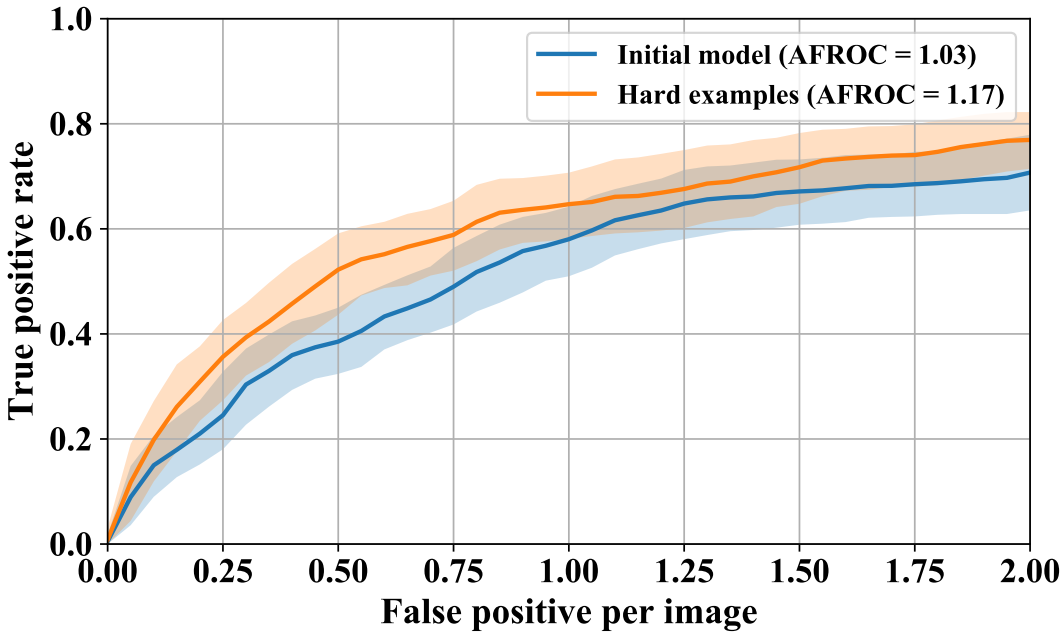


Figure 2.5: Comparison of the FROC curves with and without hard sample mining. Figure reproduced from [34].

To gain a deeper understanding of each matching criteria and their tolerance to noise, the number of anchors per lesion that were labelled as positive during the first iteration of RPN training is shown in Figure 2.6. Since the actual number

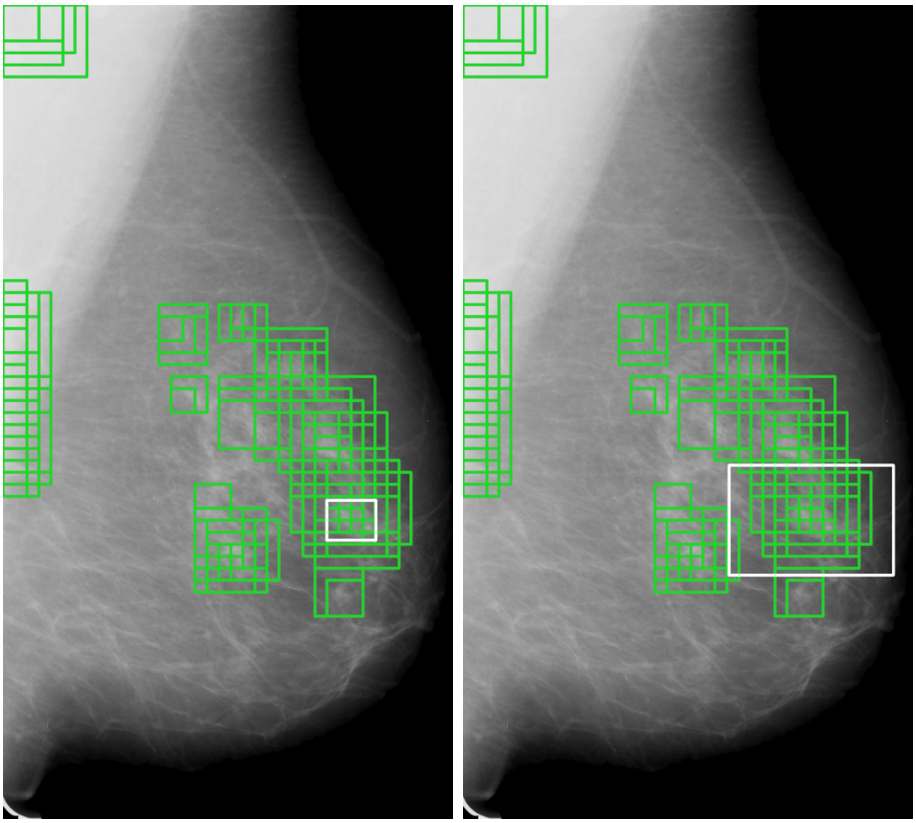


Figure 2.6: Clean (left) and noisy (right) reference standard box compared with ROI proposals generated by the RPN (in green). Many FPs overlap with the noisy reference standard box and, hence, may be mislabelled as TPs by a suboptimal matching criterion. Figure reproduced from [34].

of lesions is constant, an increase in the number of positive anchors should given an indication of the amount of labeling noise induced by an imperfect reference standard and/or the lack of a robust matching criterion. Indeed, many FP ROIs overlap with noisy reference standard boxes and may be mislabeled. In Figure 2.6, a difference can be seen between the number of positive anchors for the different matching criteria, especially for the centroid criterion, which is the loosest. It is also worth noting that the amount of box coordinate noise leads to a higher amount of labeling noise when training the RPN and the detector. As shown in Figure 2.7, the number of positive anchors increased eightfold for the IoU criterion and tenfold for the centroid criterion. In contrast, for the proposed Exp_IoU criterion, the increase is only twofold. Therefore, it is expected that the final performance will be less sensitive to noise.

Moving on the network performance, the FROC curves for different noise levels and matching criteria are shown in Figure 2.8. For easier comparison, the mean

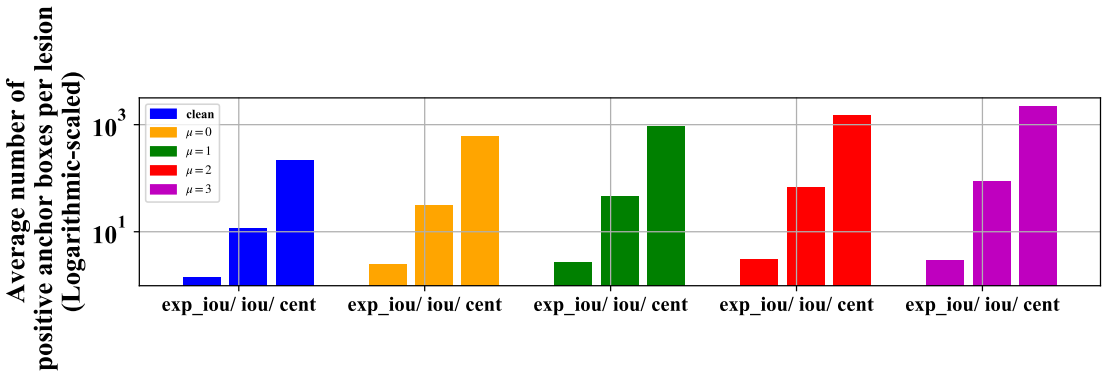


Figure 2.7: Average number of anchors per lesion labeled as positive examples in the first iteration of RPN training. Results are compared for Exp_IoU, IoU, and Centroid criteria for the clean dataset (blue) and for increasing noise (from yellow to purple). The number of positive anchors (and thus noise) increases with more relaxed matching criteria and increases more than linearly with the amount of noise. All scales are logarithmic. Figure reproduced from [34].

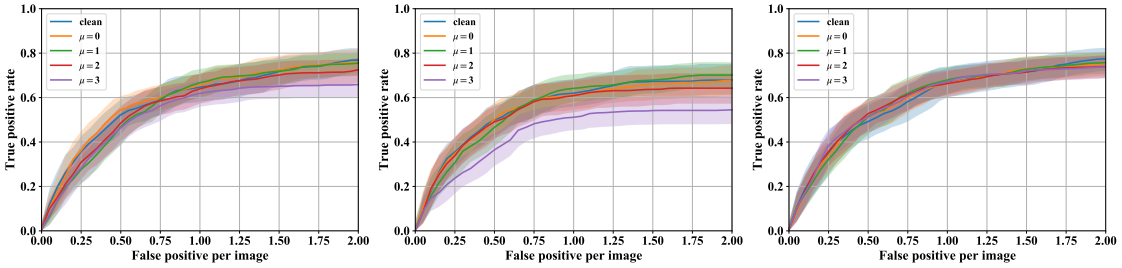


Figure 2.8: FROC curves with 95% confidence interval (calculated by bootstrapping). From left to right the IoU, Centroid, and Exp_IoU criterion were used. The latter is more tolerant towards noise with comparable performance across all levels of noise. Figure reproduced from [34].

AFROC as a function of noise level is reported in in Figure 2.9, and the corresponding confidence intervals in Table 2.1. The results confirm that the Centroid criterion is a poor choice for training because the AFROC is always lower. Both the IoU and Exp_IoU criteria perform best on the clean dataset. However, the performance of the IoU criterion deteriorates almost linearly with increasing noise from 1.17 to 1.06. Exp_IoU is the most robust criterion with respect to noise, as no drop in performance can be observed.

Comparing Figure 2.7 and Figure 2.9, the number of positive anchors and performance are negatively correlated, establishing a link between the noisy reference standard and the degraded network performance. Similar trends were observed in classification networks where the number of clean labels is relatively low and the

ratio of noisy to clean labels exceeds 10:1 [105].

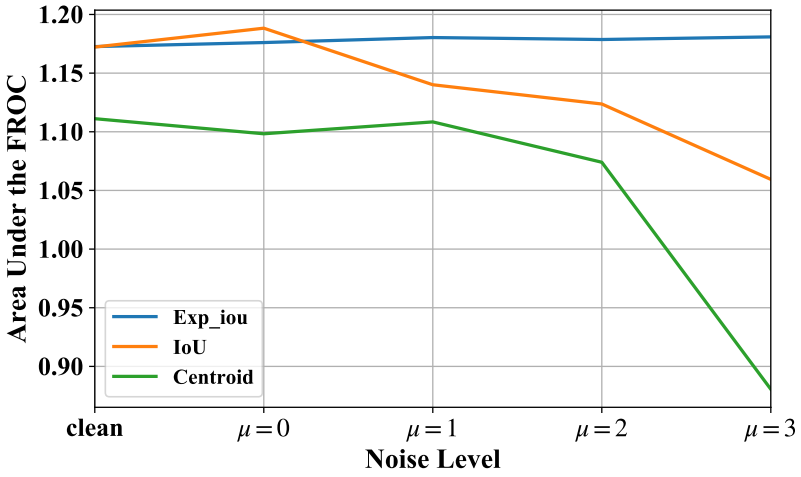


Figure 2.9: Area under the FROC curves for the IoU (orange), the Centroid inside the bounding box (green) and the Exp_IoU criteria as a function of the noise level. Figure reproduced from [34].

Table 2.1: Confidence intervals for the AFROC.

Criterion	**Clean**	$\mu = 0$	$\mu = 1$	$\mu = 2$	$\mu = 3$
IoU	1.29-1.05	1.29-1.09	1.23-1.039	1.23-1.01	1.16-0.95
Centroid	1.23-0.9	1.22-0.98	1.23-1.00	1.20-0.94	1.00-0.76
Exp_IoU	1.30-1.04	1.27-1.07	1.29-1.07	1.28-1.08	1.28-1.08

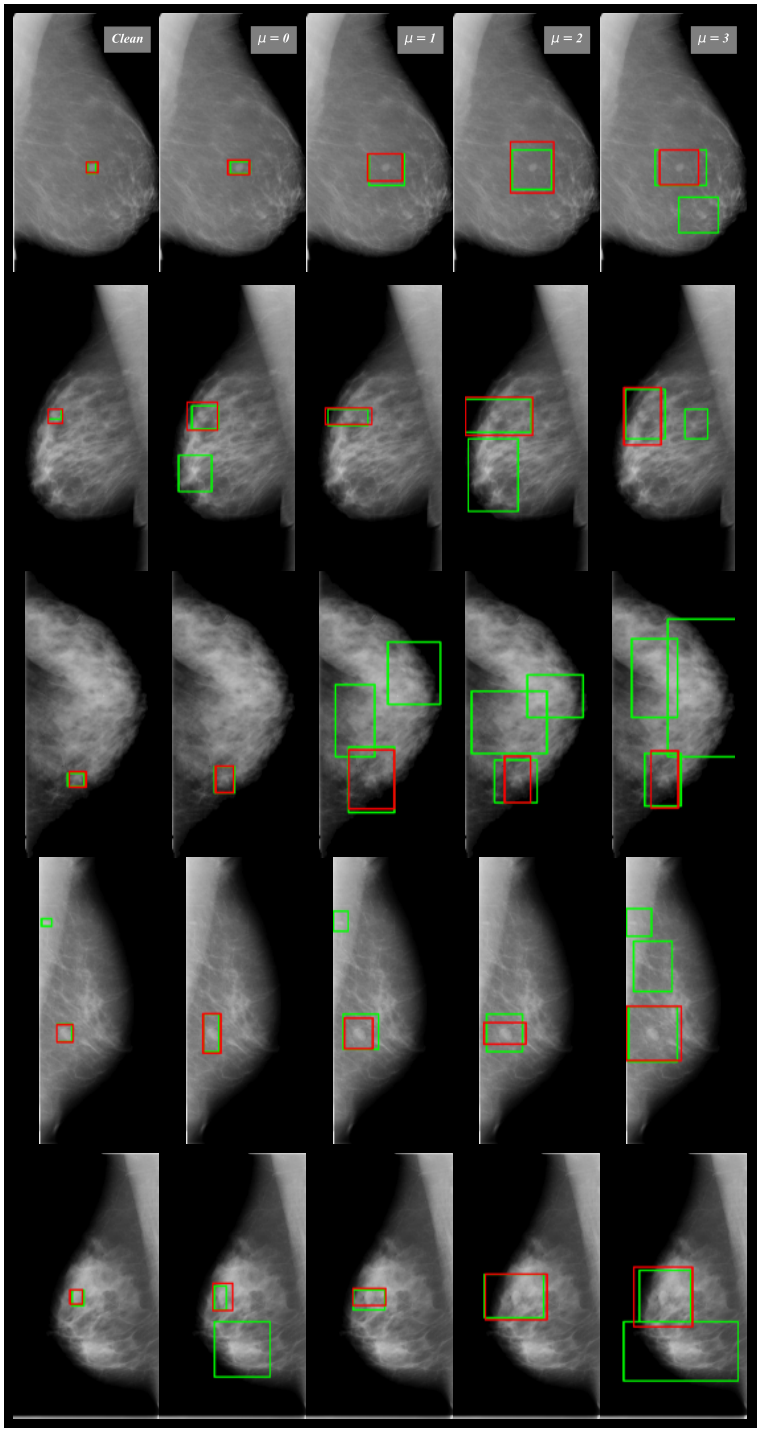


Figure 2.10: Examples showing the clean and noisy reference standard annotations (red) vs. Faster R-CNN predictions (green), with IoU matching criterion. From left to right the level of noise increases: it can be seen how the number of false positives detected by Faster R-CNN increases as well. Figure reproduced from [34].

2.5 Discussion

The present research work focused on a specific type of noise (bounding box coordinate noise) that is extremely important for training object detectors against an imperfect reference standard and has not been addressed in previous literature. The experimental results presented in this chapter show that coordinate noise, especially in the form of enlarged bounding boxes of the reference standard, leads to labeling noise when training the classifier heads, ultimately reducing the performance of the Faster R-CNN object detector.

In Object detectors, the tolerance of the network to noise is determined by the matching criterion used to label anchor boxes and RPN proposals during training. The default criterion used in object detection, the IoU, is surprisingly robust to moderate noise ($\mu = 1$) and only degrades with very high noise: at 2 FPs/frame, for example, the average sensitivity drops from 77% to 65%. Other criteria, such as the Exp_IoU in particular, are able to keep the performance stable across all noise levels. The main reason for this is that it explicitly considers the position of the anchor box with respect to the center of the lesion, not just the overlapping region. Therefore, misleading anchor boxes that are not well positioned are discarded during training. However, this advantage may be lost if the lesion centers are also affected by random noise. For example, cases in which the ground truth annotations contains lesions located closer to the borders of the bounding box rather than its center would be in conflict with this underlying assumption. This may happen in cases that the radiologist may want to include some additional context or tissue which is relevant to the diagnosis, or in case of lesions and abnormalities that have irregular asymmetric shape such as microcalcification clusters, as opposed to breast mass. However, this violation would hinder the performance only if such cases are dominant in the dataset.

Previous work had already shown that training object detectors for mammography and for medical images in general requires specific selection of hyperparameters, such as appropriate IoU thresholds, because lesions occur less frequently and with fuzzier boundaries compared to objects in natural scenes [102]. Here, these results were confirmed, and a hard-negative mining strategy was used to select informative and balanced samples for training the detector.

The present study has potential limitations. First, the dataset is relatively small and insufficient to achieve peak performance as in [102]. This is partly due to overfitting, which has also been observed in previous work [17]. In these experiments, overfitting was partially mitigated by early stopping and hard sample mining, but larger datasets (real or synthetic) are needed to solve the problem [16]. The impact of noise may also depend on the size of the data set. Noise tolerance can be expected to improve further with larger data sets, provided more examples of lesions are available [105].

Second, in the current training procedure, the effect of noise on the regression

parameters is unavoidable, i.e., the predicted bounding boxes are increased, as shown in Figure 2.10. Further work is required to improve the robustness of the regression task. For example, knowing the level of noise in a given dataset, the regression parameters could be adjusted during or after training. Strategies to mitigate the effects of bounding-box noise on regression have also been proposed by Gao et al. [39].

In this work, noise was simulated by explicitly manipulating annotations in a well-curated dataset. This is a common approach to study the effects of noise in machine learning, and allows to carefully control the experimental conditions. However, conclusions should be validated in a real dataset. Similarly, the proposed approach could be extended to other types of lesions or datasets.

2.6 Conclusion

In this research it was quantitatively investigated the effect of bounding-box coordinate noise while training object detection networks for mammography. It was shown how state-of-the-art object detectors are robust to varying degrees of labelling noise, and proposed strategies to mitigate its effect at extreme noise levels. This study has important implications for dataset collection and annotation, since it shows that the bounding boxes do not need to be very precise for training to be effective. In the case of extreme noise levels, small changes in the training procedure, such as introducing a different matching criterion, can improve performance without increasing the complexity of the model, and can be easily incorporated in the training procedure of any object detector. These findings open new opportunities to train lesion detection model by using bookmarks and annotations routinely recorded by radiologists in their clinical practice.

Chapter 3

Multi-view lesion detection and registration in mammography

Work described in this chapter was partially presented in [33].

3.1 Introduction

Population screening using digital mammography was shown to reduce mortality associated with breast cancer [14, 88]. However, the 2D projective nature of mammography results in strong tissue superposition [114, 85]. On the one hand, superimposed tissue may mimic the presence of lesions, increasing the false positive rate of both radiologists and computer-aided diagnosis systems alike. On the other hand, tissue superposition can mask the presence of lesions. This is particularly true when the breast tissue is very dense [108], since the fibrous and glandular components have a greater attenuation than the adipose tissue. Dense breast are thus associated with both increased cancer risk and reduced screening sensitivity. In the United States, the proper management of dense breasts has fueled a massive debate, especially since recent legislation was passed that requires radiologists in many states to notify women regarding their breast density [119, 50]. While 3D techniques such as digital breast tomosynthesis can alleviate this problem [114, 85], digital mammography is still the de facto standard for mammography screening, especially in Europe. Hence, techniques are needed to increase sensitivity and specificity of screening mammography.

In a standard screening exam, two views are acquired for each breast, referred to as CC and MLO [30]. The breast is positioned between two compression plates; in the MLO view, the compression plates are rotated 45 ° -50 °, towards the armpit or axilla. The radiologist is thus able to locate suspicious areas on both views by roughly triangulating from these projections. This increases diagnostic confidence as false positives due to tissue overlap are likely to disappear in the contralateral

view.

Likewise, traditional CAD algorithms also showed reduced false positive rates when taking into account the two views [91, 112, 30]. Conventional CAD algorithms achieved this goal by processing each view independently, and then integrating the results by matching detection on both sides based on their positions and visual features. Visual similarity can be estimated based on handcrafted characteristics such as texture, size, intensity, etc. [112] or, with the advent of deep learning, by training a Siamese CNN, in which features are learned by comparing patches, pushing corresponding patches closer in feature space, while seeking to separate patches extracted from different patients or images [91]. More recently, deep learning techniques that can simultaneously process two or more projections at the time have become available [137, 132, 76]. However, few techniques are available that can match corresponding areas in the CC and MLO view as the radiologist would do [76, 137]. The goal of this research is to translate the aforementioned aspect into a multi-view object detector that processes two mammographic views simultaneously. The proposed framework should be able to combine information as soon as possible during lesion detection.

A registration phase is then added that precedes the multiview network with the aim of geometrically transforming one view (CC) to align it with the other (MLO). Unfortunately, breast registration is considerably more challenging than other imaging modalities as the soft tissues of the breast are compressed and distorted during the acquisition [42]. In the current literature, few authors have explored CC and MLO view registration and there is no established deep learning approach for this activity [42, 46]. Compared to approaches that match lesion candidates, the proposed registration technique works directly on the input image and can be applied before, after, or independently of other lesion detection or classification networks. At the same time, it is a flexible and versatile module that can be incorporated and jointly trained in more complex pipelines. Successfully training a registration CNN requires defining a robust loss while reducing the cost of annotation [46]. To this end, in this thesis work the standard loss of mean square error (MSE) is extended by exploiting the lesion annotations available in the form of bounding boxes. GIoU forces registration to match actual lesions in both views.

In summary, the contributions presented in this chapter are as follows:

- architectures for both affine and non-affine CC-MLO registrations are presented. Experimental results on the CBIS-DDSM dataset show that the proposed networks are able to successfully align corresponding lesion views in 75% without requiring additional annotations;
- a novel semi-supervised loss is introduced to exploit lesion bounding boxes as landmarks to complement standard similarity losses such as the MSE. This choice compensates the lack of fixed structures in the breast that could be used as landmarks;

- a novel multi-view architecture for breast lesion detection is proposed building on the Faster R-CNN framework. The detector takes as input co-registered CC and MLO frames and fuses information from both views to classify each region of interest as either true positive (lesion) or background.

The rest of the chapter is organized as follows. Section 3.2 discusses the relevant background and related work on medical image registration and multi-view lesion detection in mammography. The proposed architecture is introduced in Section 3.3. The experimental setup is introduced in Section 3.6. Finally, results are presented in Section 3.7, and conclusions are draw in Section 3.8, in which possible extensions are also discussed.

3.2 Background and related work

3.2.1 Deep learning for medical image registration

Image registration is the process of geometrically aligning two images of the same organ that present intrinsic differences that are introduced by different imaging conditions such as the images being taken at different times, from different point of views or by different sensors. Medical image registration plays a practical role in a wide range of clinical applications. The general goal of image registration is to align images into one coordinate system. A given image is designated as reference – called the *fixed* or target image – and by applying geometric transformations or local displacements another image – called the *moving* or source image – is aligned to the fixed image in such a way that anatomical or functional locations correspond with the reference image [42, 84]. Registration allows clinicians to compare and cross reference multiple images of subjects that can be captured at different time points (serial image registration), from different view points, or from different image modalities (multi-modal image registration). It may also be used as a preprocessing step in computer-aided diagnosis pipelines [84].

Generally, registration methods are based on the following steps:

- feature space computation from input images;
- feature matching with a similarity measure to quantify the alignment;
- estimation of the mapping function between the moving and fixed image;
- image resampling by means of the mapping function.

Conventional registration methods are based on on iterative optimization techniques, that do not require a training phase, but rather iteratively fit the model for each registration, making them computationally slow and not ideal for practical

clinical operations at inference time. Registration methods differ based on the domain of the transformation (global, local), its nature (rigid, affine, or elastic) and the optimization procedure [128, 42].

Recently, CNN-based techniques have been proposed to learn the registration transformation from unregistered image pairs [46]. Compared to traditional optimization approaches, CNN-based approaches are bound to have a substantial advantage: even though the training process is slower and requires hundreds or thousands of image pairs, at the time of inference it is usually much faster than computing the transformation for each pair of images.

Initially, different groups investigated the application of reinforcement learning to image registration [81, 72], but the demand for faster registration and the hurdle associated with acquiring registration ground-truth have motivated the exploration of unsupervised frameworks. The most common approaches include fully convolutional networks or encoder-decoder architectures for elastic transformations [52, 71, 93, 9] and Spatial Transformer Networks to encode affine transformations [129].

One of the main obstacles to efficient CNN-based training is the definition of an adequate loss. In principle, registration can be trained from image pairs, without additional annotations, by defining a similarity metric, such as the MSE, and a regularization term (registration is a generally ill-posed inverse problem). This approach forms the basis of unsupervised approaches, such as Voxelmorph [9], which has been applied to different imaging modalities, such as brain, breast and cardiac magnetic resonance imaging [2]. However, defining a solid measurement of image similarity is notoriously challenging, especially in the presence of different modalities, anatomical deformations or temporal changes [52, 46]. Unlike common registration tasks in brain, heart or abdominal imaging, mammography images are characterized by strong changes in point of view and high tissue deformation induced by organ compression; this fact makes the task more complex and, to the best of my knowledge, the feasibility of registering mammography images has yet to be established.

Alternatively, the registration network could be trained in a supervised fashion, which however requires to define an adequate number of manually paired points. This type of reference standard is usually difficult and expensive to obtain in the medical field. In this case, the breast is highly compressible and devoid of rigid structures, and therefore very few anatomical landmarks can be accurately matched. Large calcifications have been used as benchmarks for validating registration algorithms as they can be matched relatively easy based on their shape, and they are sufficiently small to act as localized landmarks [127]. However, collecting large numbers of such annotations would take a long time and such benign structures are generally ignored in radiological reports.

Within the specific context of mammography, many approaches for mammogram registration focused on temporal pairs of mammograms [43, 111, 131], using both conventional and deep learning approaches, in order to facilitate the detection

of changes across different screening rounds. This setting is slight easier as the registration can be independently applied to different images of the same view. In this work, the feasibility of registring the CC and MLO view of the same breast is evaluated, which introduces additional challenges due to compression and tissue distortion. The proposed methodology falls into the semi-supervised domain, exploiting existing partial annotations. A similar strategy was successfully applied to train prostate MR registration from organ segmentation maps [52].

Finally, the proposed work shares similarities with multi-task learning settings in which the registration task is learned in conjunction with another task. For example, Qin *et al.* combined cardiac motion and segmentation estimation for cardiac MRI into a single network with shared weights [93]. The devised approach is complementary in that the bounding boxes, which act as an approximate reference standard, are used to directly supervise the recording activity.

3.2.2 Multi-view lesion detection in mammography

Multi-view architectures analyse multiple views simultaneously, thus emulating radiologists' reading practice. Due to their combined structure and their ability to put together information from multiple views, they usually achieve better performance compared to single-view architectures.

Many multi-view architectures are based on convolutional neural networks that compute view-specific high level representations, which are then concatenated and input to the final classifier stage [132, 117, 113]. These type of architectures are typically trained from case-level labels, and do not directly perform lesion localization. More importantly, they work on unregistered views and do not attempt to correlate local structures across the two views, as a radiologist would do.

One of the first methods to fully aggregate the information from all views at both local and global level was MommiNet [137], a tri-view mass identification approach, simultaneously performing bilateral and ipsilateral analysis of mammogram images. In summary, the MommiNet combines two branches: the ipsilateral branch that combines the CC and MLO views of the same breast, and the bilateral branch, that combines the same view across both breasts. Affine registration is performed in the bilateral branch, but not between CC and MLO views.

MommiNet employs a Faster-RCNN Network with Siamese input module [90] and a DeepLab Network [18] with Siamese input module in parallel, to perform the ipsilateral and bilateral analysis simultaneously. Specifically, the Siamese input module consists of a Siamese neural network which tries to match corresponding bounding boxes on the two views, by computing a similarity score. It is important to note that in MommiNet, the Siamese network is trained to match corresponding lesion views: in other words, two bounding boxes are labelled as similar if they correspond to the same lesion (True Positive-True Positive pair), and dissimilar if they correspond to a lesion and false positive (True Positive-False Positive pair).

This entails that the matching can only be defined for cases which contain at least one lesion, and can only be trained on a dataset with annotated lesions.

A more recent framework is the Anatomy-aware Graph convolutional Network (AGN) for mammogram mass detection proposed by Liu et al. [76], endowing existing detection methods with multi-view reasoning ability. Differently from MommiNet, the AGN considers the point-to-point correspondence among different mammographic views, regardless of the presence of lesions, which is important for the success of multi-view reasoning. Specifically, this is achieved by dividing the breast into regular patches, each associated to a specific pseudo-landmarks, and matching pseudo-landmarks across views using a bipartite weighted graph based on geometrical and visual similarity. A graph convolutional network is then used to calculate attention maps that enhance the features calculated by the convolutional backbone. The main disadvantage of the proposed approach, that the present research seeks to overcome, is that the pseudo-landmarks do not correspond to specific anatomic structures. In the framework proposed in this work, the input images are pre-aligned so that RoI proposals across multiple views can be put in direct correspondence, without relying on an arbitrary tessellation of the breast that may not correspond to anatomical and clinically meaningful boundaries.

3.3 Deep learning methods for lesion detection in co-registered multiple mammography views

3.3.1 Overview

The proposed framework consists of two main stages: registration of the CC-MLO views and multi-view lesion detection. The first stage takes care of registering the CC view, as moving image, on the fixed MLO view; while the multi-view network performs lesion detection. This section intends to discuss and describe the details of all parts of the proposed method. A preliminary version of the proposed methodology, and specifically of the affine registration network, was published in [33].

3.3.2 Affine registration

The proposed affine registration network is an end-to-end architecture that accepts as input a pair of unregistered CC and MLO images and outputs the resampled CC image. The MLO was chosen as the fixed image and the CC as the moving image because the former also includes the pectoral muscle, which is outside the CC field of view. Registering the MLO to the CC would push the pectoral muscle out of the pixel grid of the image, and it would be impossible to estimate the correct deformation for the pixels belonging to the pectoral muscle.

The overall architecture, shown in Figure 3.1, is divided into two parts: the feature extraction block and a spatial transformer block. Feature maps are extracted for each view separately, before being concatenated and passed to the Spatial Transformer network. The architecture is trained end-to-end leveraging ground truth lesion bounding boxes as additional oversight. This provides cues for higher quality registration compared to regular MSE.

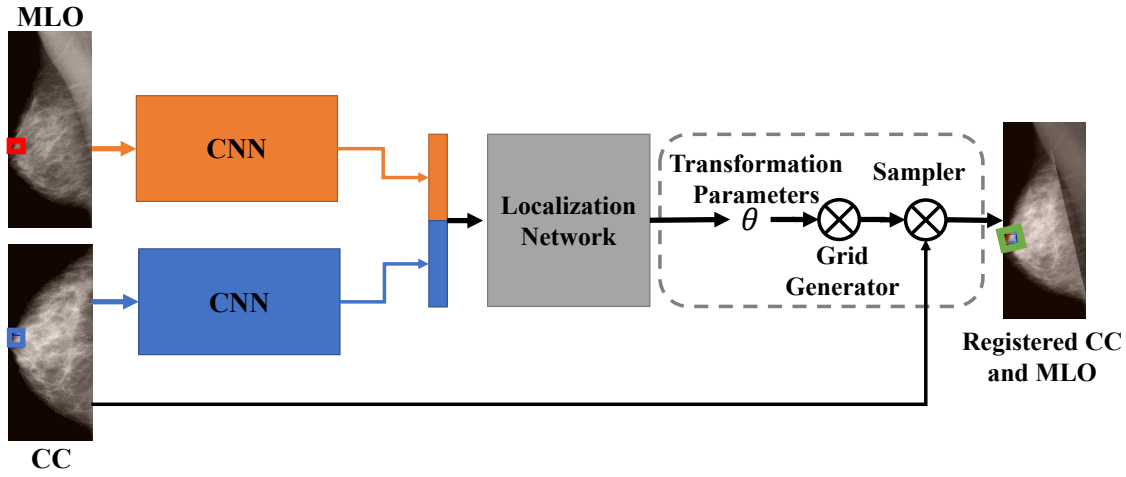


Figure 3.1: Architecture of the proposed affine registration network. The feature extraction backbone is the ResNet50 network up to the Conv4_x blocks [47]. Weight sharing between the CC and MLO views reduces the parameters count and prevents overfitting. Figure reproduced from [33].

Spatial Transformer

The affine network is based on a spatial transformer network, i.e., a lightweight block that predicts and applies a spatial transformation to an input feature map in a single forward pass. It was originally proposed as a way to improve an image classification network by allowing to align feature maps to an expected canonical pose to simplify inference in subsequent layers [56]. The spatial transformer is composed of a localization network, which predicts the parameters of an affine transformation, thus requiring in principle only six output parameters. Next, a sampling grid is created, i.e., a set of points where the input map should be sampled to produce the transformed output. Finally, the input feature map is resampled and interpolated to produce the output image. Spatial transformers include a differentiable implementation of the sampling grid and resampling layer, allowing end-to-end training, with standard backpropagation, of the models they are injected into. The network learns to actively transform feature maps to minimize the overall cost function of the network during training.

Localization network

The localization network is made of a residual block (corresponding to the Conv5_x block of the ResNet50) followed by a dense layer to predict the parameters of the affine transformation:

$$\theta = \begin{bmatrix} a_{1,1} & a_{1,2} & t_1 \\ a_{2,1} & a_{2,2} & t_2 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.1}$$

On the first tests carried out it was noted that some instances were transformed with a rotation opposite to the ideal one. During training, the loss function therefore entered a local minimum from which it was unable to escape. To guide the training of the network towards a correct solution, it was decided to encode differently the affine transformation, in order to enforce additional constraints on its output.

The translation, scaling and shear parameters were then separated from the rotation ones, breaking down the affine matrix as follows:

$$\theta = \begin{bmatrix} s_x(cos(r) + h_y sin(r)) & s_x(h_x cos(r) + h_y sin(r)) & t_x \\ s_y(h_y cos(r) - sin(r)) & s_x(cos(r) - h_x sin(r)) & t_y \\ 0 & 0 & 1 \end{bmatrix} \tag{3.2}$$

where $t_x, t_y, s_x, s_y, h_x, h_y$ represent the translation, scale, and shearing parameters, respectively, along the two x and y axes, and r represents the rotation parameter. To impose a positive constraint on the rotation parameter the Rectified Linear Unit (ReLU) activation function was used, which assumes only values in the interval $[0, + \ inf]$. A linear activation function was set for the other parameters. Finally, since the outputs of the localization network no longer represent an affine matrix, they must be interpreted and combined to obtain the equivalent affine matrix according to Eq. 3.2.

Resampler

In the case of image registration, the sampling grid is simply the pixel grid of the fixed image, which greatly simplifies the implementation of the *grid generator*[56]. The output warped CC image is obtained by applying the affine transformation to this sampling grid using a bi-linear interpolation scheme.

The above resampling scheme can be applied indifferently to the original images (as done here), as well as channel-wise to the feature maps (which could be useful if the feature maps were used for other tasks) and the bounding boxes coordinates. Points that after the registration fall outside of the field of view are assigned a zero intensity value. Bounding boxes are converted by applying the inverse affine transformation and then rectifying the results. All layers including the bounding box resampling are differentiable and, hence, can be trained end-to-end.

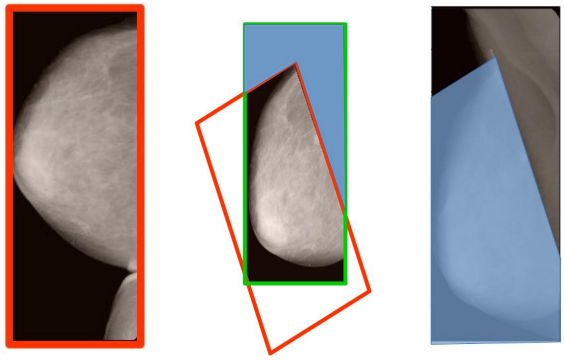
3.3.3 Loss: affine registration

The MSE cannot by itself achieve successful registration. One of the underlying reasons is that the pectoral muscle is visible only in the MLO view. It was experimentally observed that the network may overstretch the CC to cover the pectoral muscle in order to achieve a lower loss. When the images are correctly aligned, the CC should align to the border of the pectoral muscle in the MLO, as exemplified in Figure 3.2(a).

To avoid this pitfall, only the region in which the moving CC image and the fixed MLO overlap is included in the loss computation, as shown in Figure 3.2(c). The effect of the pectoral muscle, as well as of external air, is thus minimized. The resulting loss is defined as:

$$\mathcal{L}_{MSE}(X^{mlo}, X^{cc}) = \left\| (X^{mlo} - X^{cc_{reg}})\mathcal{M} \right\|^2 \tag{3.3}$$

where X^{mlo} is the MLO image, $X^{cc_{reg}}$ is the warped CC image and $\mathcal{M}$ is a binary overlap mask.



(a) (b) (c)

Figure 3.2: Calculation of the overlap mask for the MSE loss. Unregistered (red box) and registered (green box) CC views are shown in (a) and (b). The shaded blue area is included in the calculation of the loss (b). In (c) the registered CC, fixed MLO and overlap mask are shown superimposed. It can be noticed how the margin of the CC view aligns with the pectoral muscle, outside of the overlap area..

Besides the nipple and the pectoral muscle, the breast does not contain many useful anatomical landmarks that can be exploited for registration. When present, lesions can be exploited as anatomical landmarks. However, precise pixel-level segmentation may not always be available, and the compression may change the shape of the lesion in the two views. For this reason, lesion bounding boxes were exploited to complement MSE and further guide the registration process. At the same time, compared to frameworks that exploit Siamese networks (such as MommiNet [137]),

the MSE loss can be calculated also in the case of negative exams, which are the vast majority of breast examinations.

To exploit the lesion bounding boxes, a loss is needed to compare whether the two bounding boxes are being aligned by the registration process. The IoU is a widely used measure to compare bounding boxes, but it is undefined when two bounding boxes do not overlap. On the other hand, the GIoU also allows computing the relative distance between two non-overlapping bounding boxes, and thus is defined in both cases [101].

Given a pair of bounding boxes, the GIoU is defined as:

$$GIoU(B_i^{mlo}, B_i^{cc_{reg}}) = IoU(B_i^{mlo}, B_i^{cc_{reg}}) - \frac{A_c - U}{A_c} \tag{3.4}$$

where B_i^{mlo} and $B_i^{cc_{reg}}$ are the two bounding boxes, A_c is the area of the smallest enclosing box that includes them both and U is their union. When two bounding boxes do not overlap ($IoU = 0$), the GIoU loss simplifies to $\mathcal{L}_{GIoU} = 2 - \frac{U}{A_c} \geq 1$ [101]. In order to minimize $\frac{U}{A_c}$, the distance between the two bounding boxes must be reduced to the point where they eventually overlap. The GIoU loss ($\mathcal{L}_{GIoU} = 1 - GIoU$) was initially proposed as a regression loss to train object detection networks. To the best of my knowledge, this is the first time it has been used for registration purposes.

To conclude, for each pair of mammographic views the total loss is calculated as:

$$\mathcal{L}_{affine} = \mathcal{L}_{MSE} + \lambda \mathcal{L}_{GIoU} \tag{3.5}$$

where λ is a rescaling parameter.

3.4 Elastic registration

The elastic registration module is built upon the learning-based registration framework called Voxelmorph [9], which achieves state-of-the-art results while reducing the overall computation time of a deformation field. An overview of the overall architecture is presented in Figure 3.3. Deformable registration techniques often assume an existing global alignment between moving and fixed images. For this assumption, as in most registration pipelines, the non-affine module is placed after the affine one. The non-affine module inputs are essentially the resulting outputs of the affine registration.

The MLO view will be used as fixed (target) image and the CC view as moving (source) image. Both will be successively denoted with f and m, respectively. The deformable registration optimization problem is formulated as follows:

$$\hat{\phi} = \arg\min_{\phi} \mathcal{L}(f, m, \phi) = \arg\min_{\phi} \mathcal{L}_{sim}(f, m \circ \phi) + \gamma \mathcal{L}_{Lsmooth} \tag{3.6}$$

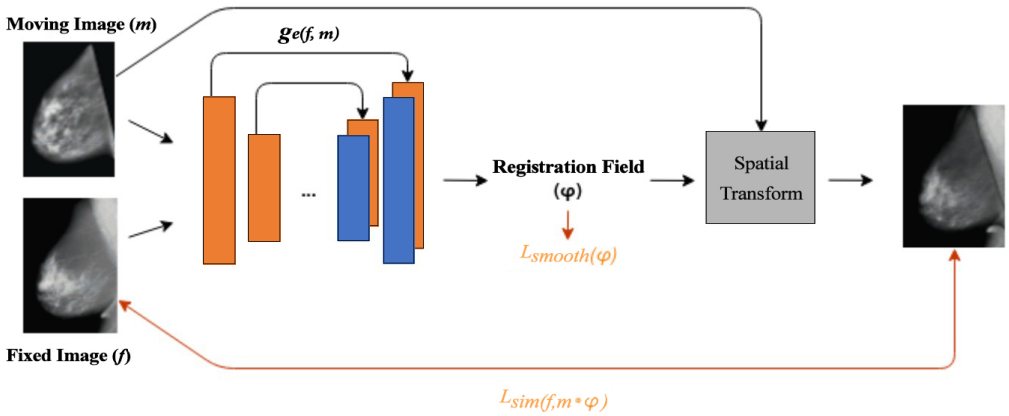


Figure 3.3: Deformable registration framework.

The registration field is denoted with ϕ and it is in the space of f. It maps the coordinates of the moving CC view m to the fixed MLO view f. $\mathcal{L}_{sim}$ denotes a similarity metric between the fixed image and the warped image. A common formulation of ϕ is characterized with a displacement vector field:

$$\phi = I + u \tag{3.7}$$

where I is the identity transform. A convolutional neural network models the displacement vector through a function $g_\theta(f, m) = u$, where θ represents the network parameters. In [9], the U-Net is chosen for this task. In order to exploit previous training iterations of the affine module, the U-Net default encoder path is substituted with ResNet-50; enabling successful transfer learning between the two registration modules.

3.4.1 Loss: elastic registration

The original Voxelmorph loss is composed of two terms: $\mathcal{L}_{sim}$that penalizes visual differences between the images and $\mathcal{L}_{smooth}$ that penalizes local variations in ϕ. The latter is essentially an L2 Loss function. In the original setting, the mean squared error (MSE) is chosen for $\mathcal{L}_{sim}$ and is defined as in Eq. 3.3. As the pectoral muscle issue cannot be perfectly mitigated, using the sole MSE would yield a non-realistic registration of the CC view by trying to over-stretch towards the pectoral muscle region of the MLO view. To exclude those parts from the loss calculation, as done in the affine registration a mask is introduced so that only overlapping regions are included in the MSE.

The elastic registration is further weakly supervised by aligning the lesion bounding boxes by means of the GIoU. Moreover, as the resulting registration will be non-affine, each pixel may be subject to a different deformation and this

could cause lesions to deform in any direction. For this reason, a simple constraint is added to enforce the warped bounding box to have a similar aspect ratio to the ground truth bounding box.

$$BB_{ar} = \sum_{i=w,h} (|B^i_{warped} - B^i_{gt}|)^2 \tag{3.8}$$

The final loss is thus calculated as follows:

$$\mathcal{L}_{elastic} = \mathcal{L}_{MSE} + \lambda \mathcal{L}_{GIoU} + BB_{ar} + \gamma \mathcal{L}_{Lsmooth} \tag{3.9}$$

3.5 Multi-stream object detection

The registered images are fed to a multi-view object detection network, illustrated in Figure 3.4, which extends the Faster R-CNN architecture introduced in Chapter 3.

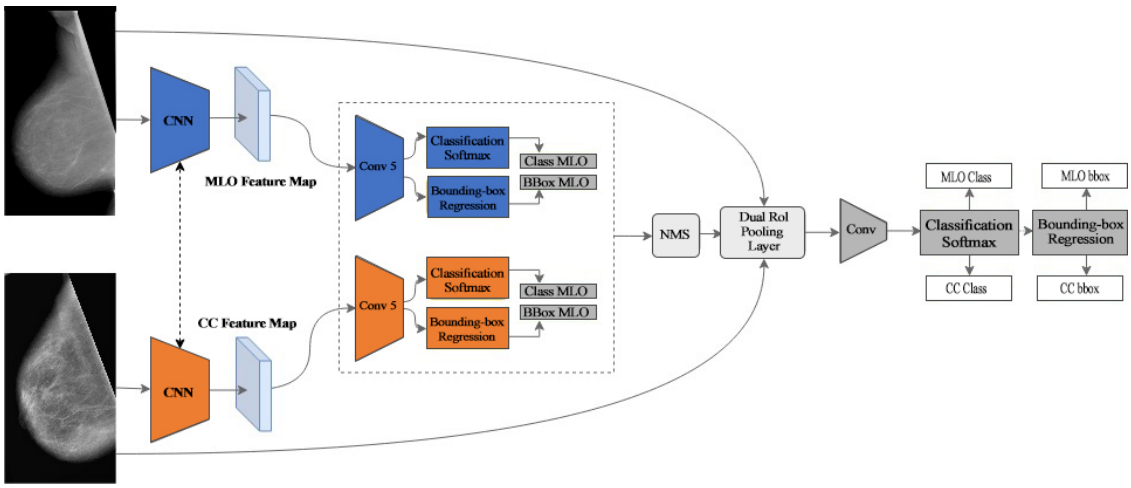


Figure 3.4: Multi-view Faster R-CNN architecture. The network takes as input the MLO and the co-registered CC image. A backbone with shared weights computes the feature maps, which are then fed to the RPN and classifier heads. The region proposals output by the RPN are fed to a Non-Maximums Suppression layer and then to a Dual-view Region Pooling layer which combines features from both views. The region proposals are then classified by a dual-output classifier head which outputs separate classification and regression parameters for each view.

3.5.1 Architecture

The multi-view network receives as input the two views simultaneously: the CC view and the MLO view. The registered CC view is the result of the previous registration modules. Feature maps of the two views are extracted with ResNet-50 and

passed to the RPN in a dual-stream fashion which will have a dedicated classifier and regressor for the CC and MLO views, respectively. The RoI Pooling layer is re-engineered to concatenate the proposals of both views by performing max-pooling for all RoIs generated for the CC and MLO views. The region proposals are then fed to the classifier head, which comprises a series of convolutional and fully connected layers and has two outputs: the classification vector and the regressed bounding box parameters. In principle, one classification vector, shared among the two views, could be sufficient assuming that the two input views are always successfully registered. However, since a successful registration, both shape-wise and bounding box-wise, cannot be achieved in 100% of the cases, the network performance would be drastically affected by the cases in which ground-truth bounding boxes of the CC and MLO are not aligned, while the predicted class labels would be symmetrical instead. To avoid this problem, two independent class vectors are dedicated to the two views.

3.5.2 Loss: multi-view object detection

As its single-view counterpart, the RPN and classifier heads of the multi-view Faster R-CNN are trained in an alternating fashion. For each training batch, the sum of image pair is simply the sum over the CC and MLO views:

$$\mathcal{L} = \sum_{w \in \{CC, MLO\}} \mathcal{L}(\{p_i^w\}, \{b_i^w\}) \tag{3.10}$$

The view-wise loss is defined for both modules as a combination of a regression and classification loss [99]:

$$\mathcal{L}(\{p_i^w\}, \{b_i^w\}) = \frac{1}{n_r^w} \sum_i \mathcal{L}_{cls}(p_i^w, y_i^w) + \lambda \frac{1}{n_c^w} \sum_i \mathcal{L}_{reg}(b_i, b_i') \tag{3.11}$$

where $\mathcal{L}_{reg}$ is the smooth L1 loss for regression, $\mathcal{L}_{cls}$ is the categorical cross entropy, b_i is the ground truth bounding box, b_i' are the output coordinates, y_i^w is the predicted probability that b_i^w contains a lesion, and p_i^w is the reference standard probability.

A more in-depth description of the training procedure is provided in Chapter 3.

3.6 Experimental setup

3.6.1 Dataset

Experiments were conducted on the publicly available CBIS-DDSM [48] for the task of mask detection. However, since the goal of these experiments is to investigate the method's ability to detect a lesion by combining information from

the CC and MLO views, only masses visible on both views were included in the experiments. For this reason, the dataset used is slightly different than Chapter 3, and thus all networks were retrained for the purpose of comparing single- and multi-view lesion detection. Specifically, there are a total of 1230 lesions visible on both sides (930 in the training set, 274 in the validation set), for a total of 1164 (CC or MLO). The aforementioned numbers take into account the exclusion of a certain number of incorrect samples that are present in CBIS-DDSM.

During visual inspection of the CBIS-DDSM samples with their relative ground truth bounding boxes, it was observed that in specific samples the bounding box coordinates were shifted from the right lesion position on the CC or MLO view. The issue has been investigated further and tracked down to the following cause: CBIS-DDSM is provided for each view with corresponding mask files for the lesions and since bounding box coordinates are extracted from those mask files - which should have the same size of their relative view image file - it is not the case for all samples; unfortunately, some mask file are of different size, thus, relative bounding box coordinates will not be correctly centered on the respective lesion location. Care has been taken in order to collect invalid samples from the data set and exclude them from all experiments.

Images were downsampled so that the largest dimension was equal to 600 pixels. Although in digital mammography patient positioning and other useful information would be available in the image headers, the DDSM collection comprises only scannerized screen-film mammography, therefore it was not possible to use any information available from DICOM headers. Images were converted to grayscale by replicating the intensity values across the RGB channels and normalized by subtracting the ImageNet mean. No other pixel normalization was applied.

3.6.2 Transfer learning

As the registration and lesion detection tasks rely on similar features, it was found beneficial to share information across different tasks. In early experiments on affine registration [33], the ResNet50 backbone was pre-trained on the ImageNet dataset and then on the task of single-view object detection, before training the affine registration. Specifically, the backbone was pre-trained for 80 epochs using Faster R-CNN as detailed in Chapter 2. This allowed for faster and better convergence on the affine registration task than transferring directly from ImageNet. In the rest of this chapter, the opposite direction is explored, in that the affine registration is pre-trained on ImageNet and then the resulting weights are transferred from the affine registration to initialize the elastic registration module, following the logical flow of the proposed worfklow.

3.6.3 Pectoral muscle removal

The pectoral muscle appears as a triangular and white region in the upper part of the MLO and as abnormalities in a mammogram will show up as focused white area, it may degrade the quality of the feature extractor used in the devised method. The extension of the pectoral muscle region depends on the ability of the radiologist to position during the exam the breast in such a way to limit its presence; extend and angulation of the pectoral muscle among mammograms appear therefore with some variability. As a preprocessing step, a simple method based on Hough transform [125, 29] was used in order to tackle the said variability and estimate the pectoral muscle boundary on the MLO view. Firstly, the image is oriented on a common side (left), then the Canny Edge Detector is applied for contour detection followed by Sobel filter and finally Hough transform.

3.6.4 Hyperparameter setup

Affine registration

Hyperparameters were experimentally finetuned on a small subset of the trianing set. The Adam optimizer was used with learning rate 10^{-4} and batch size 1. The network was trained for 300 epochs, each comprising 500 batches. The λ parameter (see Eq. 3.9) was set to 0.3. The output dense layer of the Spatial Transformer was randomly initialized using Glorot initialization. The affine transformation parameters bias parameters were initialized to a 45 degree counterclockwise rotation, which is based on prior knowledge of the acquisition process.

After the registration, the bounding-box coordinates of the ground-truth were no longer rectangular. In order to maintain valid bounding-box coordinates throughout the network, a simple solution would be to consider the enclosing box of the polygon. Likely, the newly registered bounding-box would be greater in size than its original rectangle resulting in less accurate annotations. To counterbalance this effect, the size of the enclosing bounding box was reducted by 10%.

Elastic registration

The non-affine registration module receives as inputs the affinely registered CC views from the previous affine registration module. As in the case of affine transformation, the network has a tendency to over-stretch the CC view to match the overall shape of the ML view, producing unrealistic registrations. This phenomenon is counterbalanced by removing the pectoral muscle, masking the loss to include only the overlapping region between the MLO and registered CC image, and tuning the γ regularizing parameter to enforce a smooth displacement field.

To select the best hyper-parameters, a grid search was performed on the learning rate and loss coefficients: $l_r \in \{10^{-5}, 10^{-4}, 10^{-3}\}$, $\lambda \in \{0.1, 0.2, 0.3, 0.4\}$, $\gamma \in$

$\{0.4, 0.7\}$. These experiments were conducted on a smaller set of 100 images, which were blurred in the first epochs to further reduce the computational effort, with Adam optimizer and batch size equal to 1. The final network was trained for 300 epochs, with learning rate set to 10^{-4}, λ is set to 0.3 and γ to 0.6. Images were not blurred.

On the other hand, training the network on blurred images for a few epochs, before switching to the original images, did not bring any advantage.

Single-view Faster R-CNN training

The hyper-parameter setting and training methodology, including hard negative sampling, is the same as in Chapter 3.

The network was trained for 120 epochs (were each epoch is made of 500 iterations) with the choice of Adam [61] as optimizer and a learning rate of 10^{-5}. Images were downsampled in such a way that the largest dimension was equal to 600 pixels to reduce the computational effort, although it should be noticed that better performance could be obtained with higher resolution images [102]. The anchor box scales and aspect ratios were $\{58, 256, 256\}$ and {(1,1), (0.7, 1.4), (1.4, 0.7)}, and the stride was 16. The NMS threshold was set to 0.7 at training time and 0.1 at testing time. The value of λ was set to 8.3 for the RPN loss, and 12.5 for the detector (see Chapter 4 for a detailed description of the Faster R-CNN loss).

Multi-view Faster R-CNN training

For the multi-view network parameters were inherited from the single-view network with the exception of the learning rate which was separately tuned to compare the two architectures on fair grounds. Experimental runs on a smaller dataset have been conducted in order to coarse select good candidate learning rates and a finer search was performed afterwards. Maintaining the same λ values for the RPN and detector head losses, as the single-view network, yields poor learning curves and have been therefore changed to the default configuration of Faster R-CNN and fine-tuned for balancing the performance between the CC and MLO branches. The hard-mining strategy illustrated for the single-view network was also used for the multi-view network.

The network was trained for 120 epochs (each comprising 500 iterations) with the choice of Adam [61] as optimizer and a learning rate of 10^{-5}. Image size, anchor box scales and aspect ratios, and NMS settings were set as for the single-view network. The value of λ was set to 8.3 for the RPN loss, and 15 for the detector (see Chapter 3 for a detailed description of the Faster R-CNN loss).

3.6.5 Hardware and software setup

The network was implemented in Keras 2.2 with Tensorflow 1.13.1. In order to reduce the variability between different experiments, a constant seed was set for all libraries (Numpy and Tensorflow). The order of the images was randomized, but fixed for all experiments. All experiments were conducted on an 1080Ti GPU with 12Gb of memory.

3.6.6 Evaluation

A major problem of registration methods is related to the evaluation of the registration results. The crucial question is how can a registered image be quantitatively evaluated? This issue has been identified early on in registration surveys, and has been considered a challenging issue for the past 20 years [77, 128]. In the context of mammography, since the breast does not contain many anatomical landmarks, one possibility is to exploit lesion annotation and verify to what extent the lesions are aligned before and after registration. Since the GIoU takes into account both the intersection and the distance of each pair of bounding boxes, it was determined to be a viable evaluation metric. In addition, registration results on the test set were visually inspected. Lesion detection was evaluated using the FROC curve and Area under the FROC curve (AFROC), as detailed in Chapter 3.

3.7 Results

3.7.1 Affine registration

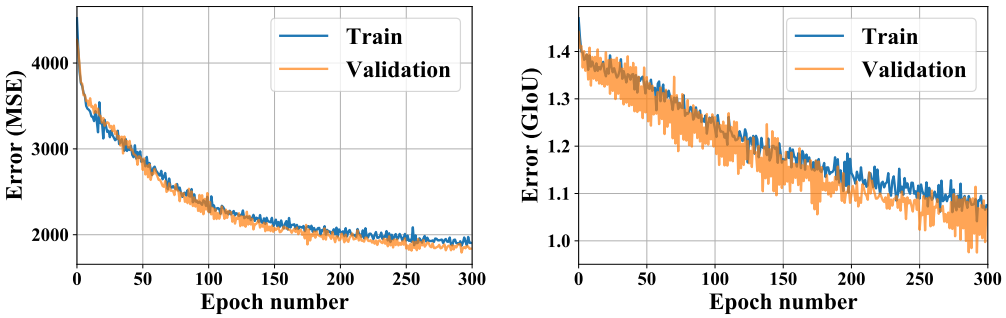
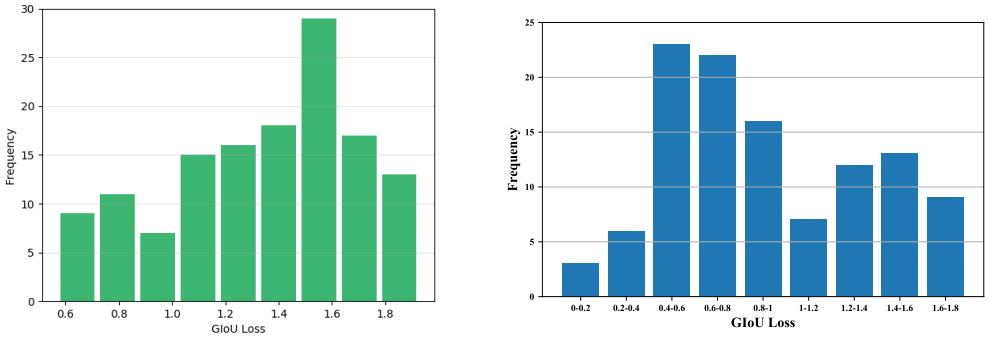


Figure 3.5: Evolution of the loss during training: MSE (a) and GIoU (b)

The affine registration was evaluated in two settings: when the backbone is pre-trained on ImageNet, and when the backbone is first fine-tuned for the task of single-view lesion detection for 80 epochs - in other words, when using a starting point the network described in Chapter 3. In both cases, the network was trained

for 300 epochs without showing signs of overfitting, as shown in Figure 3.5. Both the MSE and GIoU losses decreased indicating a synergistic behaviour.

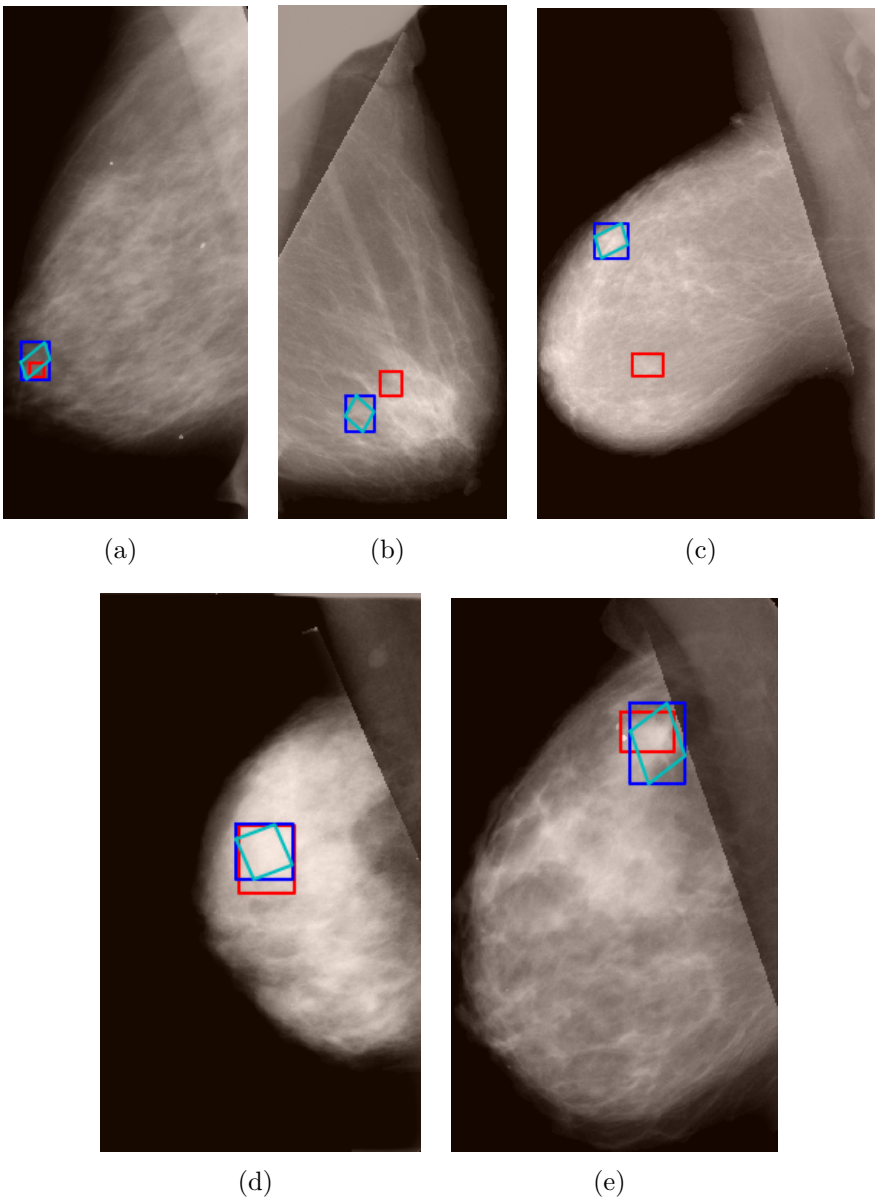
The distribution of $\mathcal{L}_{GIoU}$ on the test set is shown in Figure 3.6. The bounding boxes for the registered CC and MLO overlap in 69% of the cases, when pre-training on ImageNet, and 66.7% of the cases, when pre-training on single-view object detection. However, when the backbone is pre-trained on object detection, a higher number of lesions reaches a very small value for the GIoU (≤ 0.6). In practice, due to the rectification process, the bounding boxes are unlikely to achieve perfect overlap, and lower IoU values are to be expected. Visually, in the large majority of cases the registration was successful in aligning the two views in terms of shape and global features, although the clinical significance of the results should be confirmed by a trained radiologist.



(a) (b)

Figure 3.6: Histogram of the GIoU loss for the test set when incorporating a backbone pre-trained on ImageNet (a) and on single-view object detection (b).

Examples of successful and unsuccessful registration results are shown in Figure 3.7. In roughly 10% of the cases, the CC is still slightly overstretched to cover the pectoral muscle (Figure 3.7a). It can be shown that in two cases, even if global alignment is successful, the bounding boxes do not overlap, sometimes by a large amount (Figure 3.7c): this indicates that certain deformations cannot be recovered with the proposed affine transformation.



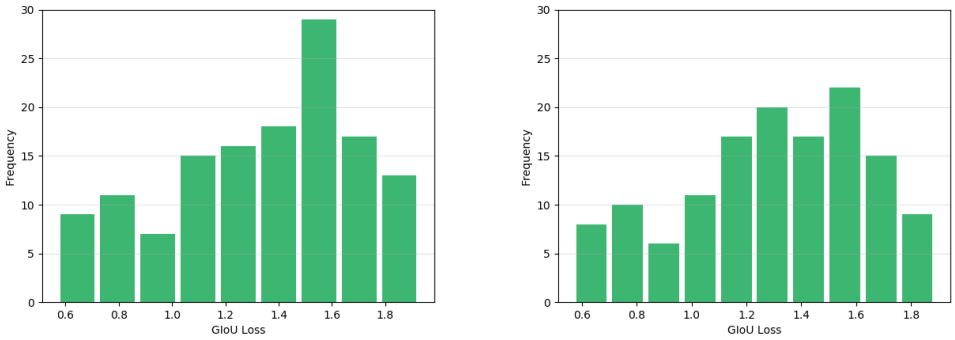
(a) (b) (c)

(d) (e)

Figure 3.7: Affine registration examples: the MLO and registered CC views are shown overlapped. The MLO bounding box is shown in red, the CC in blue, before (cyan blue) and after rectification.

3.7.2 Elastic registration

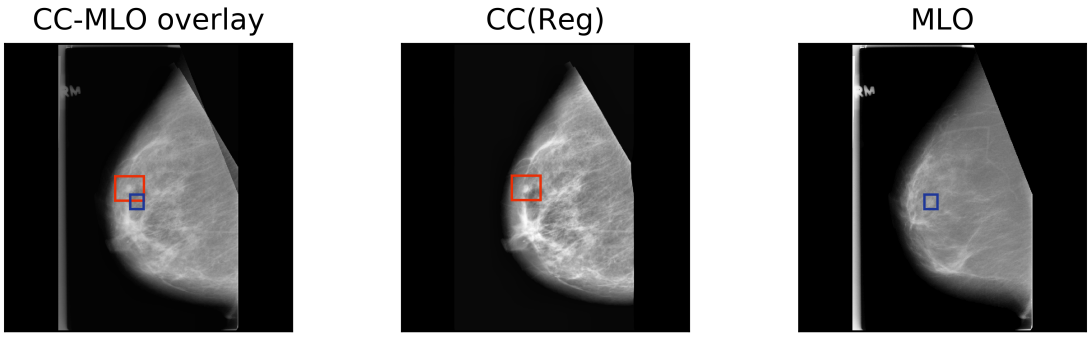
The distribution of $\mathcal{L}_{GIoU}$ on the training and test set is shown in Figure 3.8. After training, bounding boxes between the MLO and registered CC view overlap in 70% of cases. A notable improvement is observed for $\mathcal{L}_{GIoU} \geq 1.4$ after introducing elastic registration given the drop in frequencies of overlapping bounding boxes for each bin at the right tail of the distribution histogram..



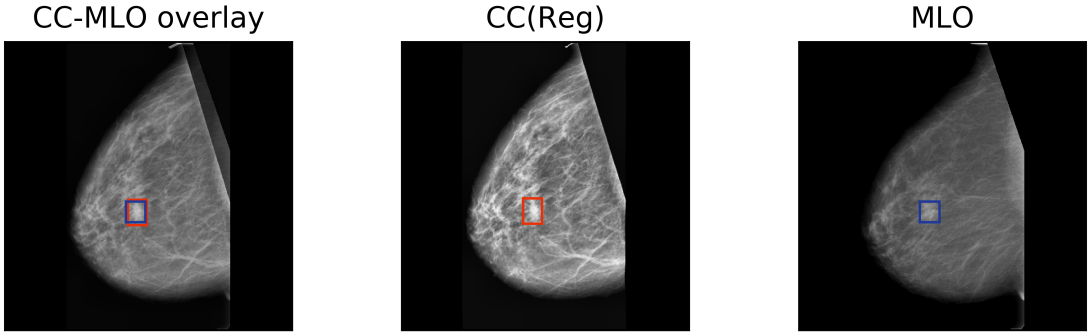
(a) (b)

Figure 3.8: Histogram of the GIoU loss for the test set for affine (a) and elastic (b) registration.

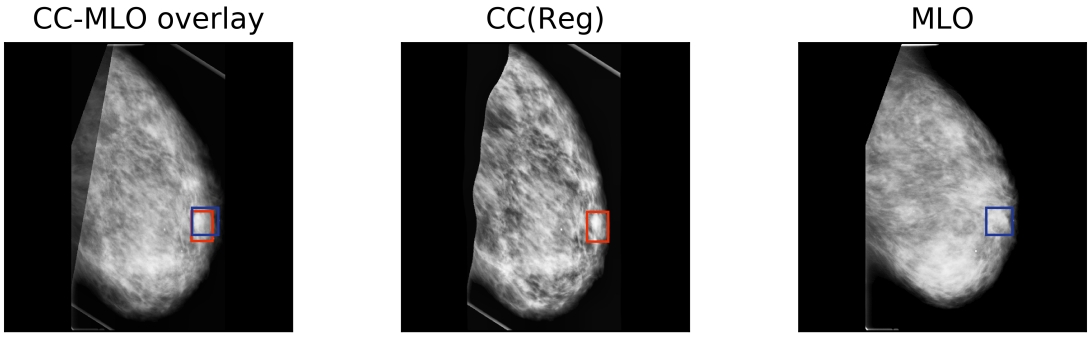
Examples of elastic registration outputs are shown in Figure 3.9 and Figure 3.10. As it can be noticed in Figure 3.10, there are still cases in which, even after the registration, the lesions do not perfectly overlap. This is due to the highly non-rigid and compressible nature of the breast, combined with the effect of tissue compression and breast manipulation by the radiographer.



(a)

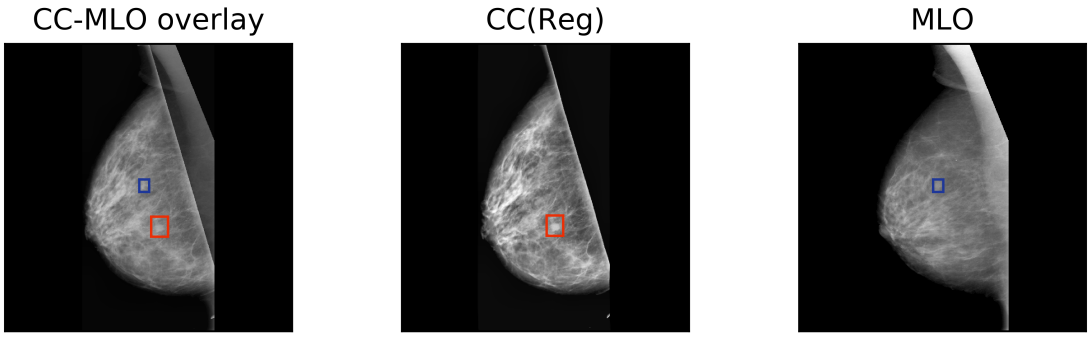


(b)

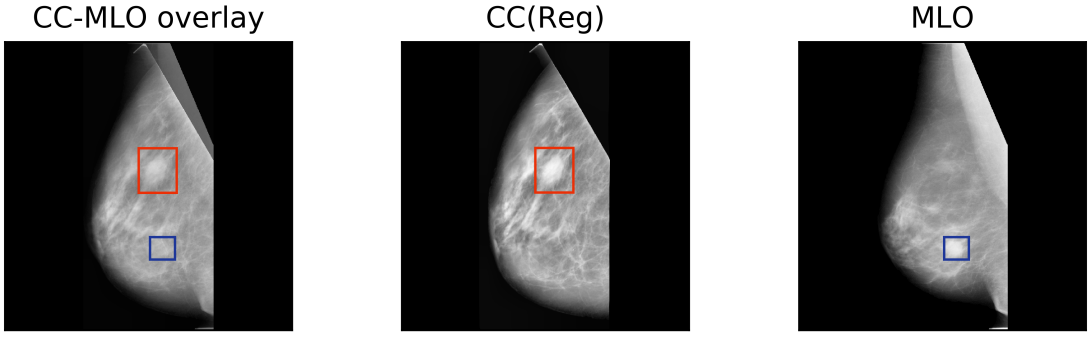


(c)

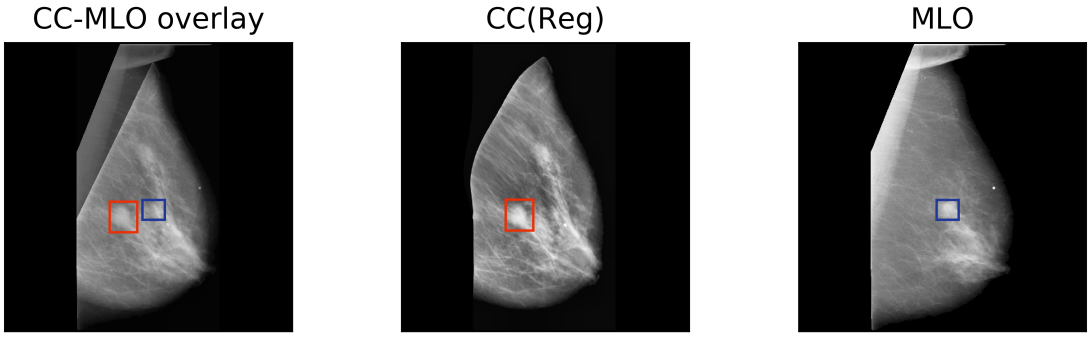
Figure 3.9: Examples of successful elastic registration (test set). The MLO and registered CC views are shown separately and then overlapped. The MLO bounding box is shown in red, the CC in blue, after rectification.



(a)



(b)



(c)

Figure 3.10: Examples of unsuccessful elastic registration (test set). The MLO and registered CC views are shown separately and then overlapped. The MLO bounding box is shown in red, the CC in blue, after rectification.

3.7.3 Multi-view vs. single-view lesion detection: convergence analysis

The multi-view detector classifier and regression loss plots, together with the mean overlapping bounding boxes over the training period are shown in Fig. 3.11.

During the training of the single-view network, as previously mentioned in Chapter 3, a potential overfitting problem emerged. FROC curves of the train and test set are plotted for epochs 20, 40, 60, 80 and 100 to understand when the model starts degrading performance on unseen data. As highlighted in Figure 3.12, the single-view model performance on newly seen data starts degrading around epoch 60 while performance on training data continues to improve with less false positives. This confirms previous literature that showed how object detectors, trained on CBIS-DDSM, have a tendency to overfit [17].

Therefore, the first test conducted was aimed at understanding if the multi-view network would also be affected by a similar overfitting problem. FROC curves were plotted for the multi-view model at epochs 20, 40, 60, 80 and 100, respectively, and are shown in Figure 3.13. Overfitting occurs in epoch 60 when the model performance on unseen data starts to decline. Compared to the single-view architecture, the gap between train and test FROC curves indicates that the overfitting problem is even stronger in a multi-view setting.

FROC curves of training set have been plotted for both the multi-view and the single-view network for epochs 10, 20, 30, 40, 50 and 60, to understand the convergence rate of the two models. As shown in Figure 3.14, the multi-view model has similar FROC curves on the training set as the single-view network for the first 30 epochs; specifically, the sensitivity is 71%, 84% and 86% at 2 FP per image, respectively. Subsequently, the FROC curves of the multi-view model are distinctively higher, achieving 96% at 2 FP/image compared to a sensitivity of 93% at 2 FP/image of the singleview network.

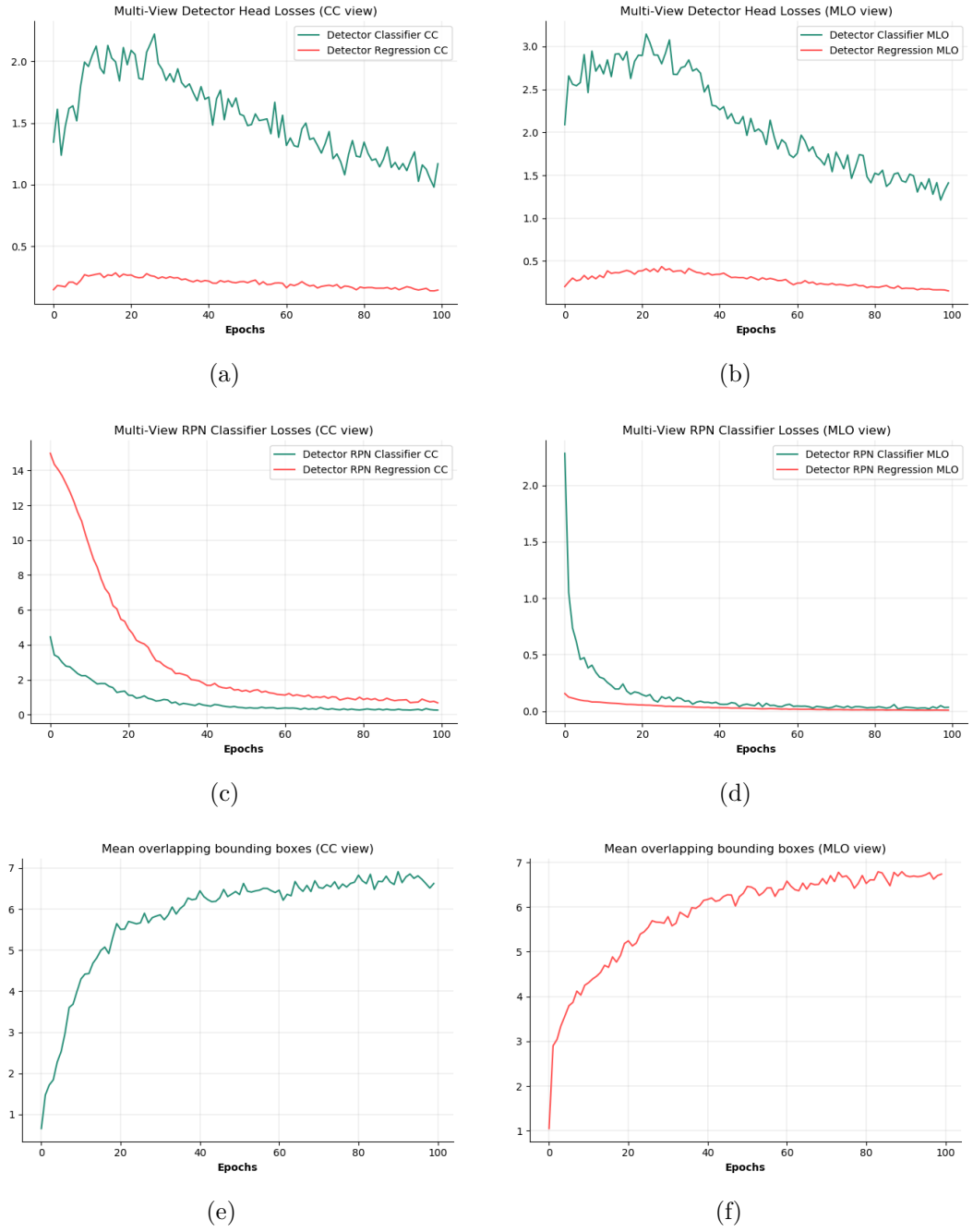
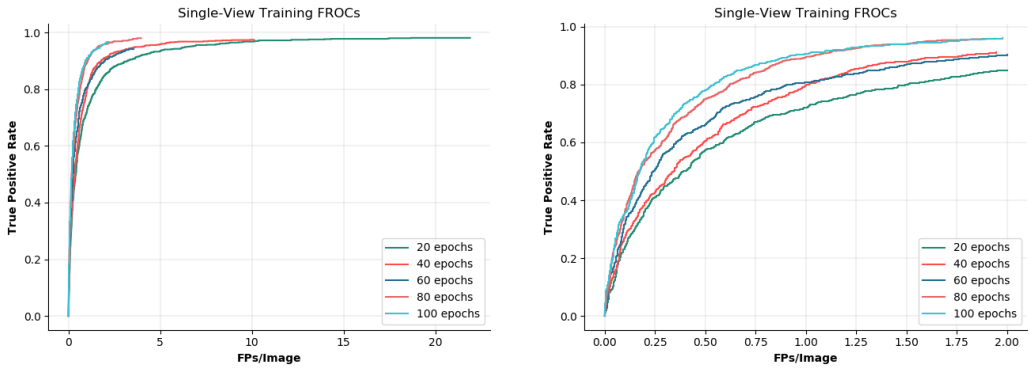
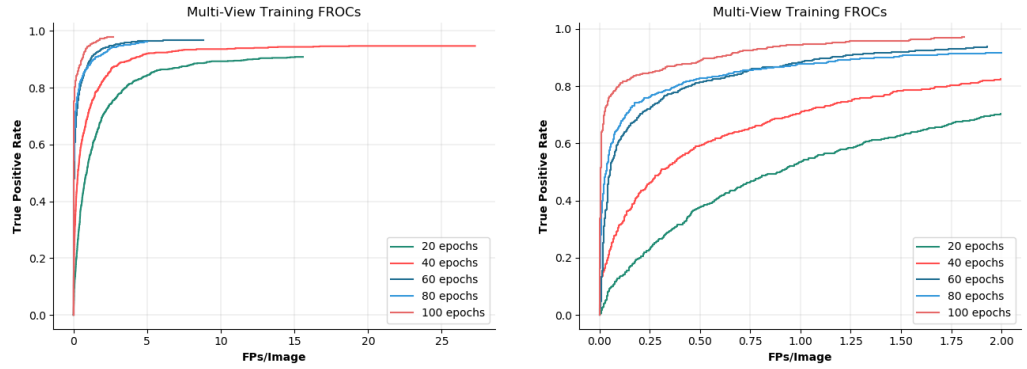


Figure 3.11: Monitoring of Multi-View Training (CC view on the left column and MLO view on the right column): Detector Losses (a,b), RPN Losses (c,d) and Mean Overlapping Bounding Boxes with Ground Truth (e).



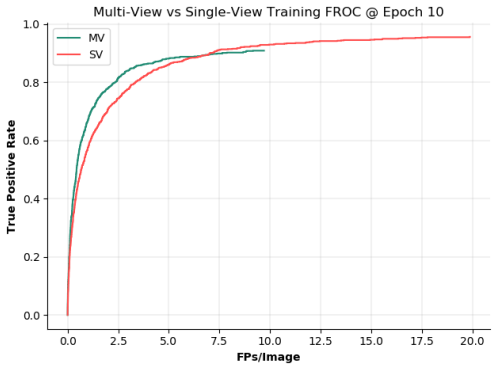
(a) (b)

Figure 3.12: Single-View Network Training FROC @ 20, 40, 60, 80, 100 epochs: full curve (a) and truncated at 2FPs/image (b).

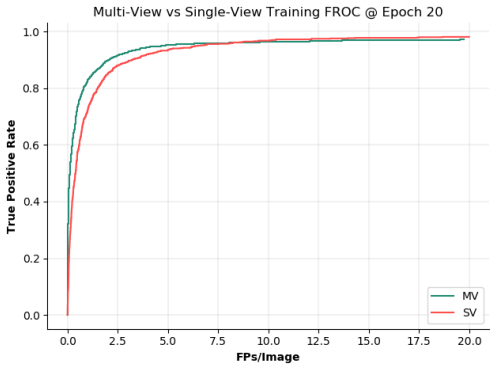


(a) (b)

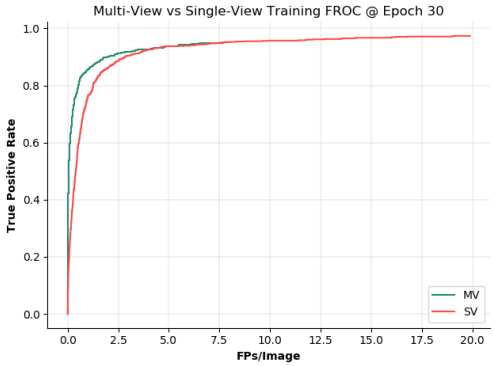
Figure 3.13: Multi-View Network Training FROC @ 20, 40, 60, 80, 100 epochs: full curve (a) and truncated at 2FPs/image (b).



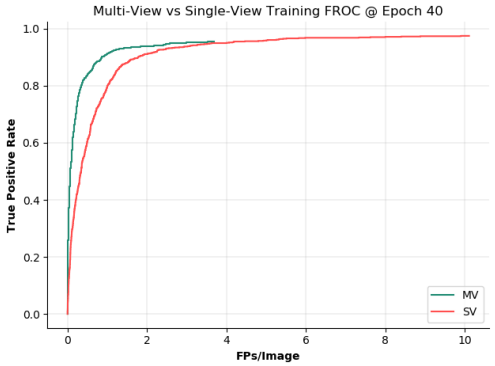
(a) FROC@epoch 10



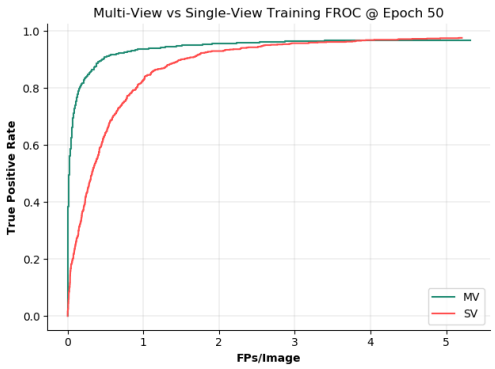
(b) FROC@epoch 20



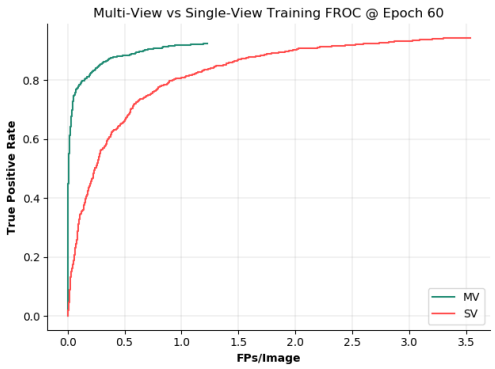
(c) FROC@epoch 30



(d) FROC@epoch 40



(e) FROC@epoch 50



(f) FROC@epoch 60

Figure 3.14: Model convergence analysis: train set FROC curves of Multi-View network and Single-View network at epochs 10, 20, 30, 40, 50, 60.

3.7.4 Multi-view vs. single-view lesion detection: performance analysis

Performance of the single-view and multi-view network on the test set is shown in Figure 3.15. The multi-view model achieves a sensitivity of 56% at 1 FP/image and 70% at 2 FP/image with an AFROC value of 1.03. Whereas, the single-view achieves sensitivity of 63% at 1 FP/image and 73% at 2 FP/image with 1.17 AUFROC value. It is interesting to note that the multi-view model has lower sensitivity scores but still manages to detect a greater number of lesions of the test set compared to the baseline single-view network. Performance of the respective best models of the multi-view and single-view network with their respective AFROC values, number of detected lesions and true positive rates are reported in Table 3.1.

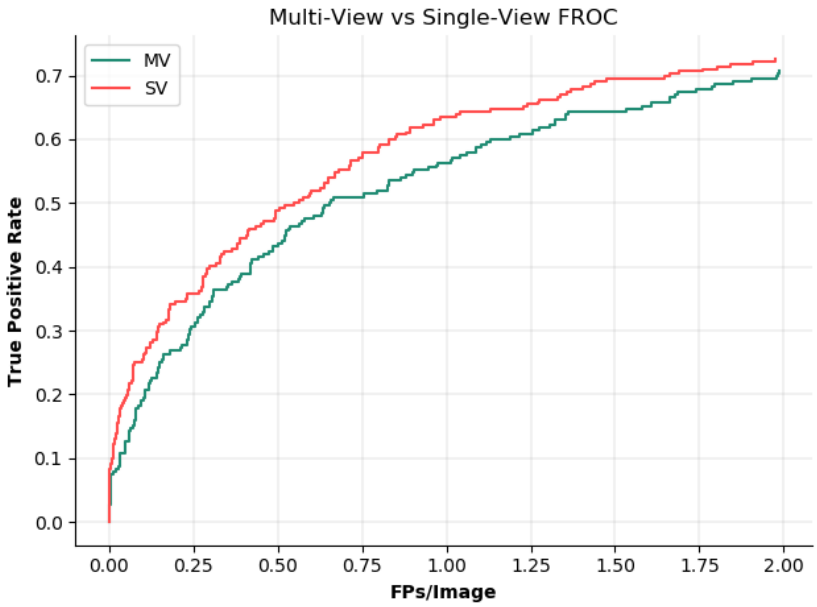


Figure 3.15: FROC curves on the CBIS-DDSM test set.

Network	TPR	AFROC	Detected/Tot. lesions
Single-View	0.73@2FPI	1.17	236/252
Multi-View	0.70@2FPI	1.03	241/252

Table 3.1: FROC curve comparison between Multi-View and Single-View network.

As mentioned earlier, when designing the multi-view network, its detector head could potentially have just one class vector that is shared between both views. This

choice was initially motivated by the strong assumption that the CC and MLO views would be perfectly registered. This led to an interesting behaviour of the multi-view network that can be observed in the next figure; the predictions on the CC and MLO view become more dependant, indicating that the network's purpose of detecting matching lesions on both views is effectively working. However, the network Achille's heel is represented by those cases of CC-MLO couples in which the registration fails to align the CC and MLO bounding boxes. As illustrated in Figure 3.16, the symmetric behavior of the network is maintained but in one of the views, the predictions could be associated with the right bounding box and label, while in the other, the prediction could be associated with a different bounding box that also has a different label.

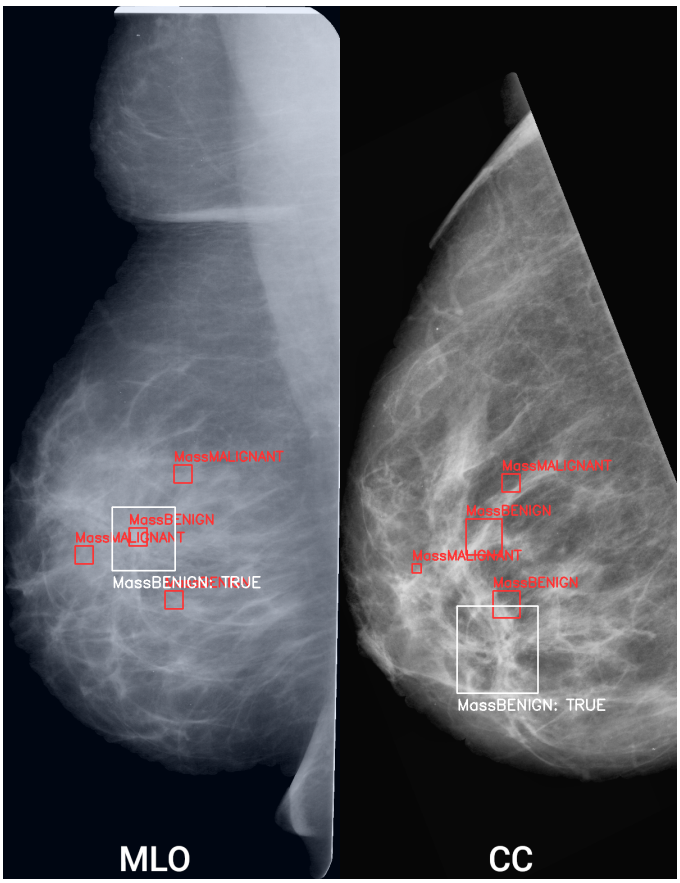


Figure 3.16: Predictions of the multi-view network with a single class vector for both views on an example in which the registration was not able to align the lesion bounding boxes.

Examples of mass detection results of the single-view and the multi-view networks are shown in Figures 3.17 and 3.18, respectively. For the multi-view network,

the co-registered CC views (right) are shown alongside the MLO views (left). Reference standard boxes are shown in white, and predictions in red. It can be noticed how, at the selected operating point, the multi-view network appears to be producing less false positives compared to the single-view counterpart.

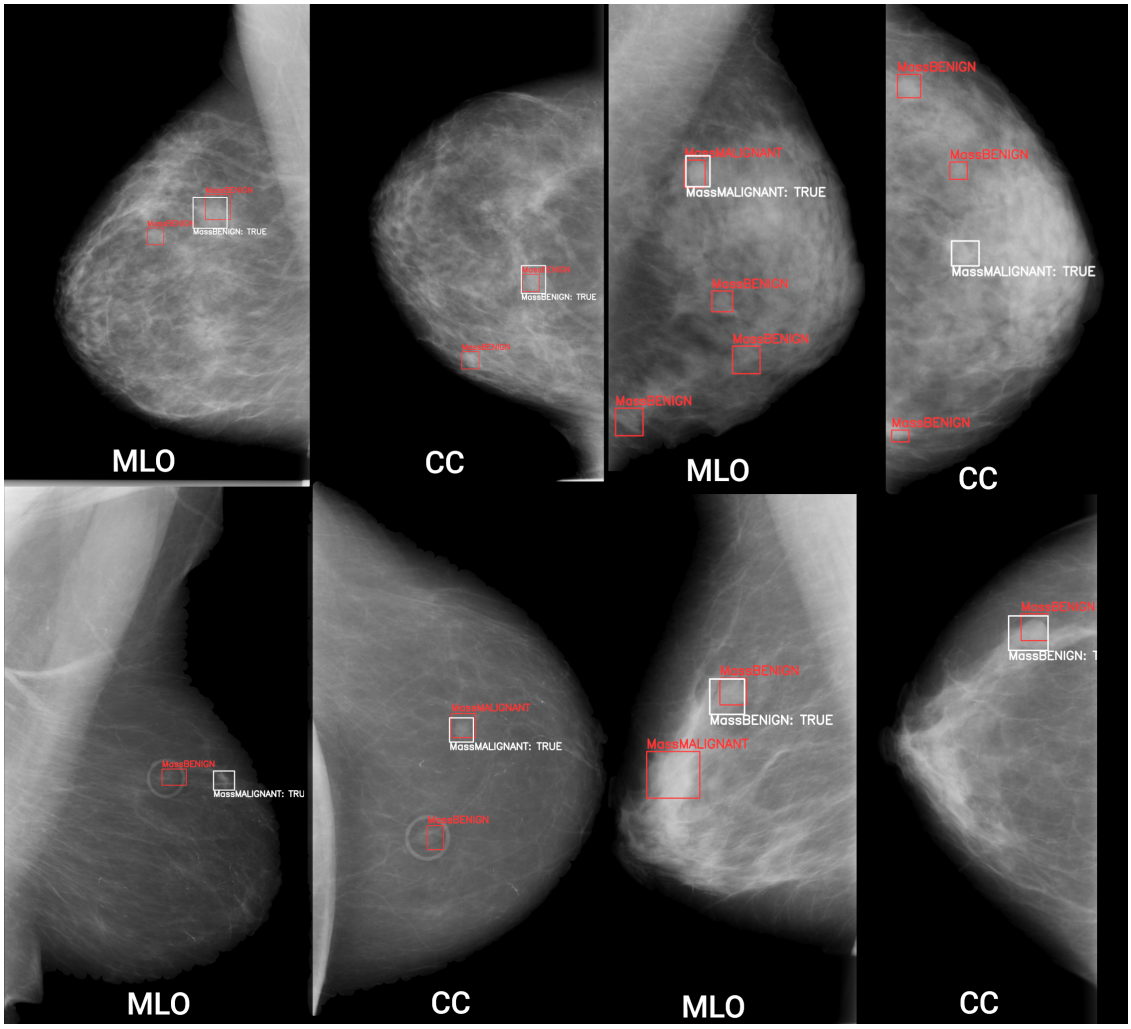


Figure 3.17: Single-View Network: Mass detection Results on CBIS-DDSM.

3.8 Conclusion

The presented research tackles two challenges: how to design a framework for mammographic image registration based on deep learning, and how to exploit co-registered CC and MLO views to design a multi-view lesion detector.

The first challenge is tackled by designing a two-steps fully trainable registration procedure, which includes an affine and non-affine registration step. The main

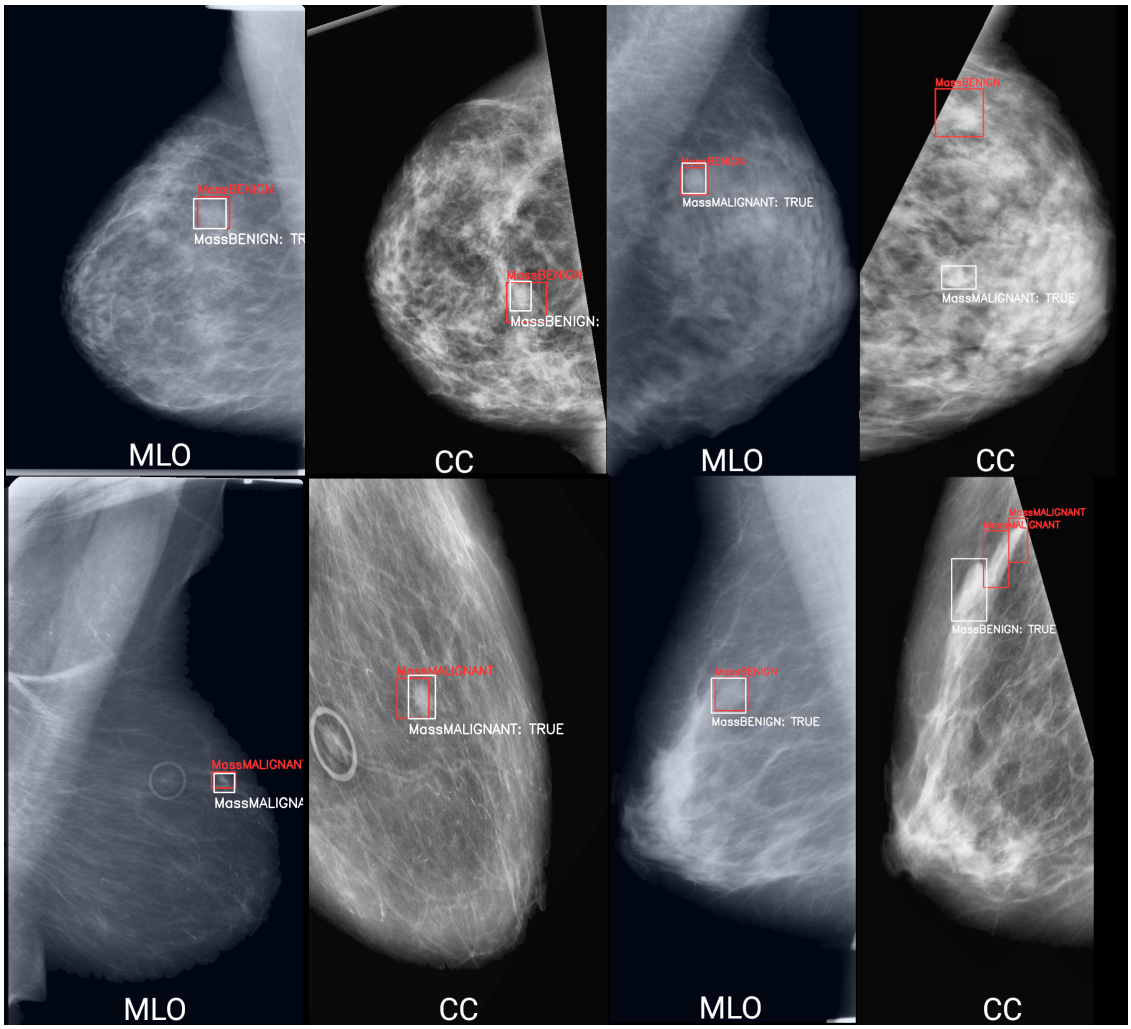


Figure 3.18: Multi-View Network: Mass detection results in CBIS-DDSM.

novelty of the proposed approach is a form of weak supervision, based on the generalized IoU loss, that exploits available lesion annotations achieves promising results in terms of visual alignment and lesion registration. Transferring the backbone weights from the affine to the elastic registration model allows to reduce convergence time. The elastic registration is thus able to recover local deformations that are not successfully tackled with the affine registration.

In the absence of other clear anatomical landmarks, the performance of registration was defined based on whether the lesion bounding boxes were aligned before and after registration. The bounding boxes for the registered CC and the MLO overlap in 70% of the cases. Overall, numerical results show that, although the images are globally and locally aligned, the registration cannot fully capture

the complex deformations occurring due to breast compression. Hence, it is not always possible to guarantee that anatomical structures, such as lesions, are perfectly aligned after registration.

The elastic registration network loss could be further improved by exploring different similarity metrics from the classic MSE, such as mutual information [115]. Improvement of the pectoral muscle segmentation could also play a role in obtaining better results with the registration modules. For example, a more accurate non-linear approximation of the pectoral muscle region could be used instead of always approximating the pectoral muscle with a straight line. The proposed technique could also be adapted to related tasks, such as temporal registration of images from subsequent screening rounds, which may be easier to solve than CC-MLO matching.

Finally, the registered images were input to a relatively straightforward multi-view extension of the Faster R-CNN. The multi-view Faster R-CNN takes as input a pair of co-registered CC and MLO views, and assumes that each anchor in one view is aligned to the corresponding anchor in the second view. Hence, each anchor is associated to a pair of bounding boxes and feature vectors, which are combined to yield the final classification. Under the assumption that the two views achieve a good alignment, the network should be facilitated in discriminating false positive detections from true lesions.

The results, although promising, do not show an improvement over the single-view architecture in terms of performance. At visual inspection, the multi-view lesion detection appears to yield less false positives. However, the proposed multi-view Faster R-CNN is based on the assumption of perfect co-registration. When this hypothesis is not verified, the network performance is reduced, thus leading to stronger overfitting and offsetting the increased specificity.

The main drawback of the proposed network is that it enforces a strict matching of the CC and MLO view based on an imperfect registration. Hence, in order to outperform the single-view network, the registration must become more accurate and/or the multi-view object detector must be more robust with respect to registration errors.

For instance, the elastic registration and the object detection could be integrated into a single network, jointly trained in order to leverage the complementarity of the two tasks. Variants which do not require warping the entire CC image could be investigated in this setting, as it would be "only" necessary to establish a matching between the anchors. The resulting architecture would be more similar to Retina-Match [100], which enforces a one-to-one mapping between the CC and MLO views by considering the similarity between the detected RoIs. However, Retina-Match is applied after the RPN stage is separately applied to the two networks, whereas the proposed multi-view network incorporates information from both views at all stages of the lesion detection process. Alternatively, attention-based mechanisms, such as the one proposed in [76], could be leveraged to account for imperfect matching

between the two views, and allow each anchor to be associated to multiple anchors on the other view. A joint network would also solve the chicken-and-egg problem of which task (object detection or registration) should be trained first to achieve optimal performance, with the added benefit of shorter inference time as only one backbone would be computed.

Further improvements could be obtained by exploring other object detector architectures, such as RetinaNet [73], and by enlarging the dataset size, also by employing more aggressive data augmentation strategies. Synthetic lesion insertion, which has shown to be beneficial in single-view architecture [17], could be also beneficial if consistency between the CC and MLO views is retained.

Chapter 4

Self-supervised pre-training for robust representation learning: the MedNet framework

4.1 Introduction

Nowadays, the amount of information being generated every day is fascinatingly large and data hungry deep learning methods have become a crucial resource for processing the data and building intelligent systems. Healthcare and medicine can largely benefit from the advances in deep learning. The main reason is that such methods provide the possibility of automatically learning representations suitable for the task at hand as opposed to traditional machine learning methods which require domain knowledge and manual feature engineering. Even though we are living in an era where the scarcity of information seems to have been vanished, data starvation is still the primary challenge for building supervised deep learning-based systems [118].

The primary limitation to training supervised deep learning models for a new imaging task is the lack of sufficiently large, labeled dataset. Small and labeled datasets that are publicly available are easier to collect, but networks trained on smaller datasets may not generalize well to unseen data [32, 109, 84]. One of the main reasons for the lack of publicly available data in medicine is that patient health data is protected by privacy laws which requires patient consent and thorough data anonymization [62]. Even when patient consent is available, the process of labeling and annotating medical images is expensive and time-consuming in comparison to natural images, since medical images may require the consensus of several experts for annotation [92, 32, 7, 84]. These challenges make it difficult and extensive to provide collect training datasets, and few public datasets are available as their collection requires extensive industrial and governmental support. Even though some datasets have been made publicly available and anonymized, they only reach

tens to several thousands of images associated with annotations for each task and modality [55, 95], which represents a significant gap with respect to datasets such as ImageNet. Furthermore, there is a lot of fragmentation in the medical domain, since it has been a long-standing practice to develop optimized pipelines targeting different organs, lesions types, imaging modalities and tasks. Hence, a plethora of specialized datasets are available, and such specialization makes it difficult to apply transfer learning within the medical domain, as learned representations do not transfer well from one organ to another [21, 86, 106]. Some examples are available in Section 4.3.

Given the cost factor of building medical datasets, there are some common approaches which may help to improve the model's generalization. One is heavy data augmentation, which has proven to be effective [11]. Another approach is transfer learning from supervised pre-training on large datasets such as ImageNet, which is commonly used through the literature [21, 51, 86] even though some practitioners have shown that this methodology does not always improve the performance and rather speed up convergence [94].

As it has been established earlier, collecting annotated datasets makes deep learning highly expensive and economically infeasible, however, collecting similarly large unlabeled datasets, is much easier [32, 62]. Therefore unsupervised, semi-supervised, and self-supervised techniques have gained the attention of researchers. Among those, self-supervision refers to a broad set of techniques in which pre-training is based on synthetic labels or tasks which does not require experts and can be generated automatically with the aim of learning context, texture, and shapes of objects present in an image. Self-supervision has gained a lot of traction for it has many attractive properties: first, it provides all the benefits of unsupervised training (since labels are self-generated by the model), without the added complexity of adversarial training. Second, experimental evidence is accumulating that models trained using the most recent self-supervised techniques may transfer to new tasks better than those trained using traditional supervised learning, at least in the general computer vision domain [31].

The main goal of this research is to leverage self-supervised techniques to pre-train deep neural networks on unlabeled medical images, and then transfer the learned representations to diagnostic tasks, thus avoiding the domain mismatch with natural images [94]. The present thesis aims to develop a multi-task, cross-domain training methodology that can be used to simultaneously train deep neural models on a variety of tasks (e.g., lesion types) and domains (e.g., imaging modalities), and test whether it is possible to achieve better generalizability on new tasks for which training data is scarce, by exploiting perceptual similarities that are exhibited by a wide variety of organs, lesions and anatomical structures in general. The aim is to demonstrate the existence of a set of "universal" descriptors, based on Convolutional Neural Networks, that are tailored to medical images. It is reasonable to assume that low- and mid-level features in medical images are sufficiently

different from natural scene images that a dedicated approach has a higher chance of long term success. This is not only due to the different nature of the acquisition modality (e.g., ionizing radiation), but also the fact that large intra-subject variability can easily obscure subtle, yet crucial in prognostic values, differences. For this reason, the proposed approach will be experimentally compared with the popular approach of pre-training on ImageNet and fine-tuning on medical datasets, that despite some controversy is still adopted by many practitioners.

Furthermore, the medical community places a greater emphasis on extracting quantitative numerical features, that are robust and highly reproducible: that is why there is a general trend to switch towards deep learning for detection tasks, but textural features are still widely used for other applications such as radiomics. Hence, it would have a great impact community at large to have a set of pre-trained descriptors that can be used for rapid initialization and subsequent fine-tuning.

The contributions of this chapter are:

- acquiring a comprehensive multi-modal dataset based on publicly available datasets which covers a variety of body parts,
- training a fully convolutional self-supervised network for representation learning,
- a comprehensive study of the representations provided by the aforementioned network,
- a study of transfer learning for classification task with unseen data and comparison with the well-known ImageNet pretrained weights.

The rest of the chapter is organized as follows. In Section 4.2, the most relevant work to this research has been studied. Section 4.3, the pretraining dataset and the datasets for transfer learning has been discussed. Section 4.4, focuses on methodologies, for pretraining and transfer learning, as well as, describing evaluation methods for studying the representations. The experimental results are presented in Section 4.5. The results have been discussed in Section 4.6. Sections 4.7 and 4.8 are dedicated to conclusion and future work.

4.2 Related work

4.2.1 Introduction to self-supervised learning

Self-supervised Learning (SSL) is a branch of unsupervised learning in which a learner models the characteristics of the target data by learning to solve one or more pretext tasks. It can thus be viewed as an autonomous form of supervised learning which combines the advantages of both unsupervised learning, as the learner can

exploit vast amounts of unlabelled data, with those of supervised learning, as the pre-text task performance can be captured, and optimized, by an objective function [28, 118]. In addition, the pretext task often relies on the same knowledge and features required to solve the target task, thus the representation (model) learnt by the network can be transferred (i.e., used to initialize) the target learner [31]. Compared to generative models (such as GANs), SSL algorithms are mostly based on discriminative approaches which are easier to train and less computationally expensive.

Different classes of SSL algorithms have been proposed, which can be broadly characterized as *task-oriented*, *generative* or *reconstruction-based*, *contrastive* and *clustering-based* techniques [118, 31]. It should be noticed that multiple techniques can be combined to achieve more robust self-supervised frameworks [118]. Task-oriented methods require solving one or more supervised tasks such as rotation recognition, colorization, inpainting, outpainting, solving jigsaw puzzles, and so forth [28, 144]. Generative or reconstruction-based methods are based on compressing and reconstructing the original image using autoregressive or autoencoder networks, possibly after distorting the original image by introducing noise or other distortions [143, 44]. More recently, contrastive-based techniques, starting from the seminal work by Chen and colleagues [19], have been proposed [44]. Their premise is to learn a pretext-invariant representation: briefly, each image instance is modified by composing multiple random transformations, and the network learns an embedding (representation) in which each instance is close to its modified versions, whereas unrelated instances are farther apart. Compared to other SSL techniques, contrastive learning is more challenging to implement, as it requires comparing multiple instances at once. However, recent advances have made significant progress towards understanding effective training techniques and tricks (such as very large batch sizes), and novel methodologies are continuously emerging. Clustering-based methods, such SwAV and DeepCluster-v2 [15] learn a representation by partitioning the data into clusters, enforcing consistency between cluster assignments produced for different transformations (or views) of the same image; these techniques, however, should not be confused with shallow clustering algorithms, as the whole deep network is optimized towards the clustering task.

Given these premises, one of the most investigated applications of SSL is as a pre-training strategy, with the goal of substituting or surpassing fully supervised pre-training. In fact, recent exciting results have shown that transferring from a self-supervised pre-trained model, particularly using the most recent contrastive and clustering-based techniques, outperforms fully supervised pre-training on ImageNet on a variety of image recognition datasets [31]. However, the same study highlighted how there is not, at the state-of-the-art, a single SSL technique that works universally well for all downstream tasks. For instance, clustering-based methods excel on image classification on datasets similar to ImageNet, whereas contrastive techniques are particularly apt at spatially sensitive tasks such as dense

prediction/regression (e.g., semantic segmentation).

Also the network architecture may play a role in facilitating transfer. The authors of [63] have compared several pretext tasks by considering the underlying network structure. Their finding shows that the underlying architecture, which has negligible effects on the performance in the supervised setting, can highly impact performance in the self-supervised setting. They have shown that the quality of the representations does not drop towards the end of the models with CNN architectures that have skip-connections such as ResNet. Increasing the number of CNN filters will increase the quality of learned representations. They have also used a linear model to assess the quality of the representations for classification tasks which shows the model is highly sensitive to the learning rate, which also persists in the transfer experiments.

Finally, no single strategy emerges when the target domain substantially departs from ImageNet. This is probably because contrastive methods employ a set of transformations which were optimized for this kind of data, and thus the broad research question as to what extent domain-specific SSL techniques are needed is still open [31].

4.2.2 Transfer and self-supervised learning in the medical domain

Many previous works in the medical domain have exploited ImageNet pre-trained models, which were shown to transfer reasonably well to the medical domain. However, this has been a subject of discussion among researchers. It has been argued that using ImageNet as pre-trained weights does not significantly help performance with respect to starting from scratch, that ImageNet pre-training improves performance only in the small data regime, and that it is not predictive of medical performance [94].

Self-supervision has also found its way through the medical domain [118]. This becomes more important since smaller simpler models have shown to work on par in comparison to more complicated ImageNet models when the dataset is not large enough [94]. Self-supervised learning methods applied to medical tasks in the literature can generally fall in various categories, as in the computer vision domain [118]. Task-oriented methods, such as relative position prediction in 3D images, rotation, jigsaw puzzle, and rubik cube, have been adapted to the medical domain and in particular on 3D images.

Self-supervised pre-training has been exploited in the medical imaging community either as a way to close the domain gap between natural and medical images [51], or as an alternative to pre-train the network on large-scale medical datasets, such as CheXPert [118]. Several works have shown that, within the same imaging modality and organ, self-supervised features (i.e., feature representations obtain through self-supervised learning) provide better initialization than ImageNet for

classification [124, 143] and segmentation tasks [143]. Even when the data is labelled, self-supervised pretraining may increase performance as the pretext task can enforce desirable properties on the learned feature space: for instance, in mammography contrastive self-supervised pretraining was shown to better generalize to images from different scanners [82]. Avoiding ImageNet pre-training also allows the use of dedicated deep architectures specific for the medical domain: for instance, self-supervised learning allow to pre-train 3D convolutional models, that can significantly outperform their 2D counterparts on task such as MR and CT segmentation [143, 144].

Most of the available literature, however, focuses on transferring self-supervised learning across the same modality and organ, which still requires the availability of suitable large datasets for each modality and organ that one wishes to support. In the rest of this chapter, an attempt will be made to learn representations that are general enough to transfer to different modalities and organs. Experiments will be conducted using a reconstruction-based technique, called Model Genesis [143], that incorporates several distortions to promote learning of textural and shape properties of organs in medical images.

4.3 The MedNet dataset

This section is dedicated to explaining how the dataset used for self-supervised pre-training has been collected, sampled, and harmonized.

The MedNet training dataset was constructed by leveraging existing publicly available datasets, accessible free of charge for research purposes. The datasets were selected in order to cover three widely used imaging modalities (CT, MRI, and X-Ray) and a wide selection of body parts. Notably, both 3D and 2D modalities were included. Herewith, the focus is on 2D images, and 3D volumes are converted to 2D images by sub-sampling slices within the volumes; in the future, the same methodology could be applied to 3D volumes by excluding X-Ray datasets. Whenever an official training/validation split was provided, images were sampled only from the training sets, to prevent feature leakage in the case the trained models are fine-tuned on the individual datasets.

In total, 17 datasets were identified, for which a description is provided in Section 4.3.1. Images were sampled, according to the strategy described in Section 4.3.2, to achieve a balanced distribution across modalities and body parts.

4.3.1 Datasets description

The MedNet training set includes images from 14 different datasets, of which 7 contain CT scans, 5 MRI scans and 2 X-ray images. The main characteristics of the selected datasets are provided in Table 4.1.

Dataset name	Modality	Body part	Number of slices
Brain Tumor	MRI	Brain	3064
Cardiac MRI	MRI	Heart	7980
CHAOS	MRI	Abdomen	1917
IBSR	MRI	Brain	12521
MRNEt	MRI	knee	118109
OASIS	MRI	Brain	216064
Prostate	MRI	Prostate	602
Chest X-Ray 14	X-Ray	Chest	112120
MURA	X-Ray	Elbow, Finger, Forearm, Hand, Humerus, Shoulder, Wrist	36808
Colon	CT	Abdomen	13486
CQ500	CT	Brain	169037
CT Lymph Nodes	CT	Abdomen	110003
Hepatic Vessel	CT	Liver	21120
LiTS	CT	Liver	85679
Pancreas	CT	Abdoment	26719
Spleen	CT	Abdomen	3650
Deep Lesion	CT	Lung, Breast, Liver, Renal, Abdominal, Posterior thigh, Perirectal, Pelvic, Omental, Peripancreatic, Splenic, Subcutaneous/skin, Axillary, Vertebral body, Thyroid, Neck	22919

Table 4.1: List of the datasets used for pre-training

1) Deep lesion: This dataset is provided by the National Institutes of health (NIH) Clinical Center [136] and consists of 10,594 scans from 4,427 unique patients. It is the largest CT scan dataset included in this study and covers a variety of different body parts and findings which were obtained through large scale mining of clinical measurements recorded in PACS archives. This dataset was originally built to advance the state of the art in lesion detection and provides annotations in the form of bounding boxes. However, unlike other datasets, it does not provide the entire datasets in DICOM format; instead, a few selected slices, centered around the lesions, are provided after conversion to PNG format.
2) CT Lymphnode: This dataset is provided by the National Institutes of Health,

labeled by radiologists at Imaging Biomarkers and Computer-Aided Diagnosis Laboratory, and contains a collection of mediastinum and abdomen images belonging to 176 patients [107].

3-9) Medical Decathlon Challenge (MDC): This is a collection of 10 different datasets, including various modalities and organs, curated for a competition launched in 2018 to promote research in generalisable 3D semantic segmentation. In this study, the following CT-based datasets from the MDC were included: **Liver Tumor Segmentation (LiTS)**, **Pancreas**, **Spleen**, **Colon**, **and Hepatic Vessel**. The LiTS dataset consists of 201 contrast-enhanced scans, including different types of primary liver cancers. The Pancreas tumor dataset contains 420 3D images taken from patients undergoing resection of pancreatic masses. The Spleen dataset contains 61 images of patients undergoing chemotherapy treatment for liver metastases, and finally, the Colon dataset, provided by the Memorial Sloan Kettering Cancer Center, includes images from 190 patients undergoing resection of primary colon cancer. Overall, these datasets cover a wide variety of acquisition settings, and include both normal and abnormal anatomy. From the MRI-based datasets in MDC, the **Prostate** and **Brain tumors** datasets were selected. The first dataset contains 48 multi-modal MR images of the prostate gland. The second dataset contains 750 multi-modal MR images of brain tumors (gliomas), sourced from the BRATS 2016 and 2017 datasets [4]. Since the brain dataset has been preprocessed to remove the skull (stripping), additional MR brain datasets which include the whole head anatomy were further included.

9) CQ500: This dataset is provided by qure.ai and was designed for the detection of abnormalities in emergency and intensive care. The dataset contains 491 CT scans with several findings such as intracranial hemorrhage, cranial fractures, and mass effects. It was included since the brain is an organ with a relatively standard anatomy, thus providing strong cues for the self-supervised framework. Additionally, it is also an organ which is normally imaged both in MRI and CT, and thus it provides the opportunity to investigate to which extent the learnt features reflect the underlying anatomy, the imaging modality or a combination of both [22].

10) CHAOS: This dataset was released for the CHAOS, or Combined (CT-MR) Healthy Abdominal Organ Segmentation, challenge [59]. This dataset contains 40 CT scans from patients who are potential liver donors, as 120 abdominal MR. This dataset does not contain lesions or tumors, but it was selected because it contained both MR and CT images of the same anatomical district (abdomen).

11) Internet Brain Segmentation Repository (IBSR): The IBSR dataset contains 18 MR brain scans of healthy subjects. It was selected to complement other MR brain datasets. This dataset includes the original MR images, without skull stripping, were included in this study [54].

12) OASIS: This dataset is provided by Knight ADRC and contains MRI scans of 416 subjects with the diagnosis of Alzheimer's disease [78].

13) Cardiac MRI: This dataset contains 7980 images from 33 subject. A sequence for each subject contains 20 frames and 8-15 slices along the long axis of the heart. The dataset is provided by the Department of Diagnostic Imaging of the Hospital for Sick Children in Toronto, Canada [3].
14) MRNet: This dataset has been created by Stanford University Medical Center for the purpose of developing automatic interpretation of knee images [12]. It contains 1104 abnormal MR scans indicating ACL (anterior cruciate ligament) tears and meniscal tears.
15) MURA (MUsculoskeletal RAdiographs): This dataset is provided by Stanford University Medical Center [95]. It contains 40561 bone X-ray images from 14863 studies, including 7 anatomical districts. Images are categorized as either normal or abnormal.
16) ChestX-Ray14: This dataset is provided by the NIH Clinical Center [130]. It contains 30,805 patients for a total of 112,120 images. The average number of images per patient is 3.6 (range [1-184]); overall 20% of patients in this dataset have more than 40 images. The datsaset comes with disease labels which are extracted via natural language processing applied on radiology reports with ove 90% accuracy and suitable for weakly supervised learning.
17) Hepatic Vessel: This dataset contains CT scans of the abdonomial part from patients with a variety of primary and metastatic liver tumors which is provided by Memorial Sloan Kettering Cancer Center (New Yor, NY, USA). It contains 443 images in which the liver vessels were semi-automatically segmented.

4.3.2 Sampling and distribution

As said, 3D volumes were converted to 2D images. The combined datasets include 961,798 2D images belonging to 48,926 patients; however, the resulting dataset was imbalanced and dominated by CT and MRI scans, given that each X-ray image is represented by a single 2D image, whereas for CT and MRI scans each case is represented by hundreds of slices. Therefore, images were subsampled in order to achieve a more balanced distribution with the goal of maximizing the number of patients / subjects included in the collection and reduce correlation between different samples. Therefore, in each dataset, at least one image for each individual samples was sampled, as follows:

- for the ChestX-Ray14 dataset, up to 20 different images per patient were selected, for a total of 88149 images;
- all images from the MURA, CHAOS, Spleen, Prostate, BrainTumor, ISBR and DeepLesion datasets were included without downsampling;
- for the remaining datasets, a fixed number of slices were sampled from each scan, up to 100 for the LiTS, CT Lymph Nodes, CQ500 and Oasis datasets,

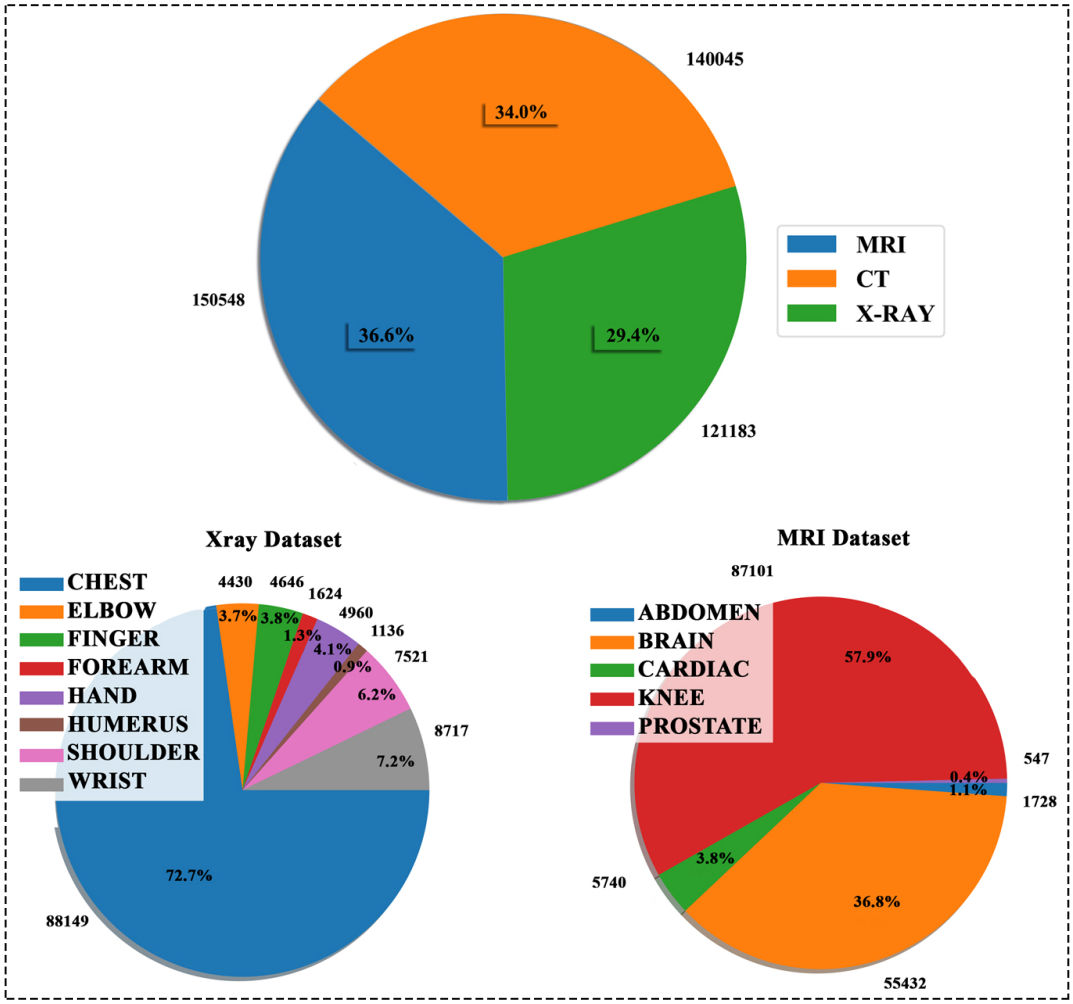


Figure 4.1: Portions of the dataset occupied by each modality is balanced.

up to 80 for the Pancreas, COLON and MRNet datasets, and up to 200 for the Cardiac MRI dataset.

Figure 4.1 depicts the portion of the resulting dataset with respect to modality and body parts. The training (90%), validation (5%), and test (5%) sets are created based on patients to avoid images of the same patient in the split. Validation and Test sets are also balanced based on modality.

4.3.3 Preprocessing

The dataset includes data from various modalities and spans a variety of acquisition devices, preprocessing methods and encoding. All images were harmonized

in terms of intensity range and approximate pixel spacing.

The pixel intensity was always rescaled to the [0, 1] range prior to training. All CT scans were encoded (or converted, in the case of DeepLesion) in Hounsfield units. They were truncated between [-1000, 1400], and then normalized between 0 and 1. X-ray and MRI scans were normalized to the [0, 1] range. X-ray datasets included in were released after converting from DICOM to png format, hence all values were rescaled from [0, 255] to [0, 1]. MRI datasets are encoded in arbitrary values, and hence the specific range depends on the scanner, for this reason, each image was individually rescaled to [0, 1].

All images were rescaled to 224×224. Since the resolution and image size of X-ray and CT is typically larger than MRI, the former were randomly cropped to 448×448 prior to rescaling.

4.3.4 Task specific datasets for transfer learning

The datasets that have been discussed up to this point have been utilized for the self-supervised pre-training. Afterwards, the eligibility of the pre-trained network needs to be assessed. One way of assessing the weights is to use them in more practical scenarios for diagnostic tasks. In other words, the self-supervised pre-trained weights will be evaluated by using the representations in a different context. This will provide a more structured and objective approach to compare the efficiency of the representations. In this section, the characteristics of the datasets are as reported in the following.

1)LUng Nudule Analysis (LUNA16): This dataset is a subset of the publicly available Lung Image Database Consortium image collection (LIDC-IDRI) aimed for a challenge in 2016 [116]. It contains 888 chest CT scans. Here, the focus is on false positive reduction rather than a full CAD system. The annotations were collected by four experienced radiologists. The target task is classification of 2D candidate patches extracted from the 3D scan. Out of 754,975 candidates, only 1182 examples are labeled as positive. The annotations contain candidate patches labeled as positive or negative. For the false positive reduction candidate patches are extracted and passed to the network. The preprocessing procedure is the same as CT scans discussed in Section 4.3.3. 80% of the data has been used for training and 20% has been used for testing. The split was done at patient level not patch level.

2)Chexpert: This dataset is a large collection of chest X-Ray images containing 224,316 radiographs from 65,240 patients [55]. The dataset was collected by Stanford Hospital and is available on Stanford ML Group's website as a challenge. This is a multi-label classification problem were each image has 14 different labels denoting different pathologies and each label can be classified as positive, negative, or uncertain. The labels for training has been extracted automatically from radiology reports. However, the test set consists of 500 unseen patients annotated by

a consensus of eight human experts.

Dataset name	Modality	Body Part	task	Seen
Luna16	CT	Chest	classification	No
Chexpert	X-Ray	Chest	multi-label classification	No

Table 4.2: Datasets used for transfer learning

4.4 Methods

After preparing the data, first it has been passed to a self-supervised framework based on a previous work called Model Genesis [143] for pre-training. After fine tuning the pre-training structure, the aim is to transfer to domain specific classification task. The overall framework is depicted in Figure 4.2. This section is dedicated to explaining the pre-training structure and the transfer procedure.

4.4.1 Pre-training

For pre-training, Model Genesis randomly applies four different types of distortions on the image, also known as pretext task, and tries to reconstruct them. The underlying structure is called Unet, which is a fully convolutional encoder-decoder network [143]. The pretext tasks are meant to extract important features from several perspectives. The framework learns appearance via non-linear transformation, texture via local pixel shuffling and context via inner and outer cutouts. Non-linear transformation refers to using a monotonic function called Bezier Curve which assigns new distinct value to different pixel values. Local pixel shuffling refers to shuffling the pixels in randomly selected small windows. Inner cutouts refer to masking an area within the inner parts of the images as opposed to outer cutouts, which refer to masking the outer parts of the image. Given the pretext task, Model Genesis requires no manual labelling. This characteristic makes this framework suitable for this work since the goal is to pre-train the encoder for a variety of domains and tasks in which some datasets are not manually labelled. Given the large amount of data that is being fed to the network, scalability is another important factor, since the framework unifies all the tasks into a single image restoration task. The overall framework has been depicted in Figure 4.3.

In the original work, the main focus is on 3D medical image analysis; in this work, image modalities contain both 3D and 2D images. Therefore, 2D convolutions have been used throughout the whole process. The base network for the encoder is ResNet-50 instead of ResNet-18 to have higher learning capacity on the model.

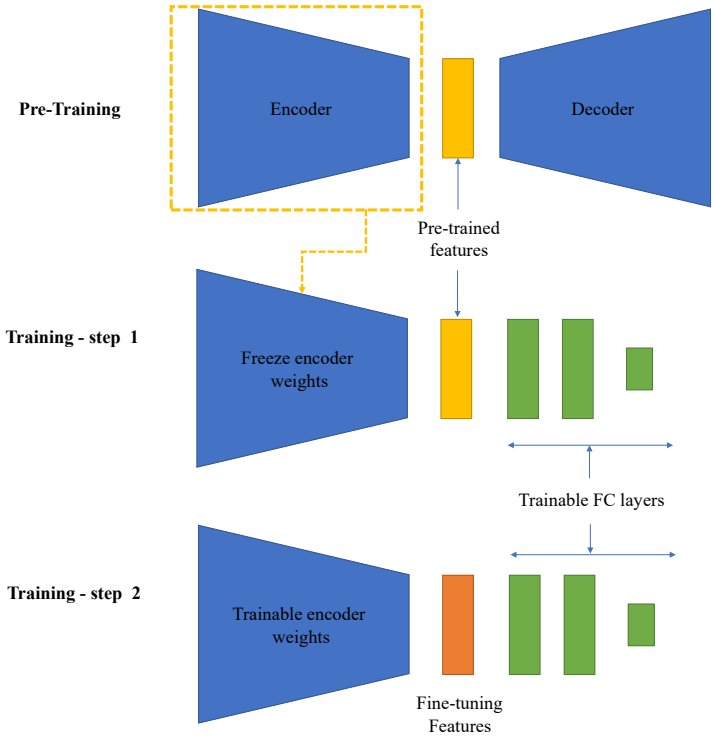


Figure 4.2: The organization of the methodology and experiments for evaluation of the pretrained weights is illustrated. The pretrained weights features shown in yellow have been used for visualization via TSNE and also transfer learning at the first step.

L1-Norm distance loss function has been the same way as the original paper. Fine-tuning the model is based on the loss on evaluation set which has not been seen during training. The model with the best loss has been selected for visualization and transfer learning.

Since the main purpose for transfer learning is to compare with ImageNet the model input size is 224x224. A SGD optimizer was used with a batch size of 128 and a learning rate of 0.1. The Loss for the model is the MSE between the original image and the reconstructed image. Another version of the model with a 128x128 input has also been trained to see the effect of scale and size of the input on the training.

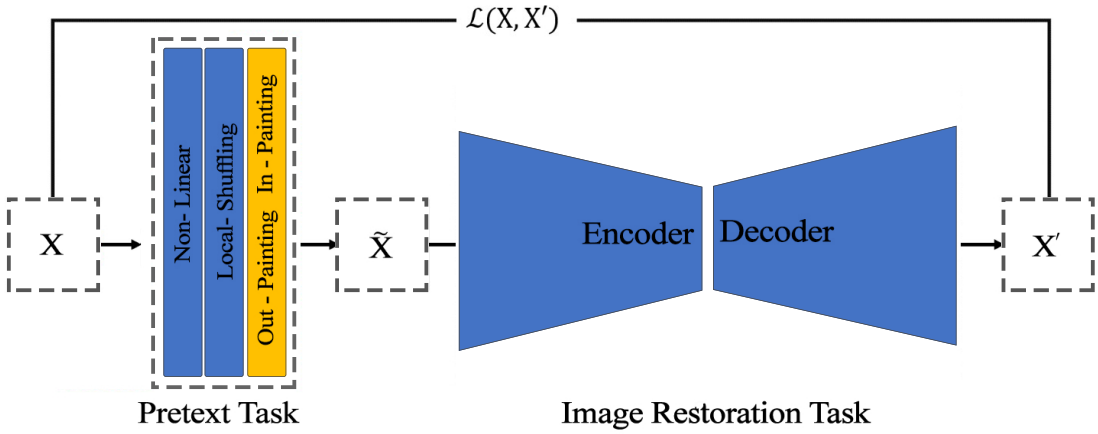


Figure 4.3: Illustration of Model Genesis framework. The images will randomly go through at most three transformations. $\tilde{X}$ is the transformed version of X. during training the model tries to minimize the reconstruction error between X' and X.

4.4.2 Transfer learning

After pre-training, only the encoder weights have been used in a classification structure for transfer learning. The network structure's base structure is ResNet50, the same as the encoder, followed by two fully connected (FC) layers for transfering to LUNA16 and one FC layer for Chexper plus a classification layer. Transfer learning has been done in two consecutive steps. In the first step, the encoder weights have been frozen, and only weights of the FC layers were being updated during training. In the second step, the previous training has been continued, however, all the layers were trainable [23].

As mentioned, the training consists of two steps which makes fine-tuning more complicated. The complication comes from the fact that during the training of the first step – which initiates the second step – several variations of weights in different epochs will be generated. Finding the best weights to initiate the training besides finding the best hyperparameters for convergence, takes a lot of time and computational resources. Hence, it was decided to narrow down the experiment to finding the best hyper parameters which gives the best performance on the validation set based on the evaluation method with respect to the task. Next, using the weights with the best results to initiate the second step, and fine-tune independently. This decision has been made with the intuition that the weights with the best performance would lead to positive transfer with respect to the rest. However, this does not hold true always.

Luna16 is considered as a binary classification task and the classification layer is composed of softmax activation function. Positive and negative patches have been extracted from the dataset. The size of each patch is 64×64. The labels are highly

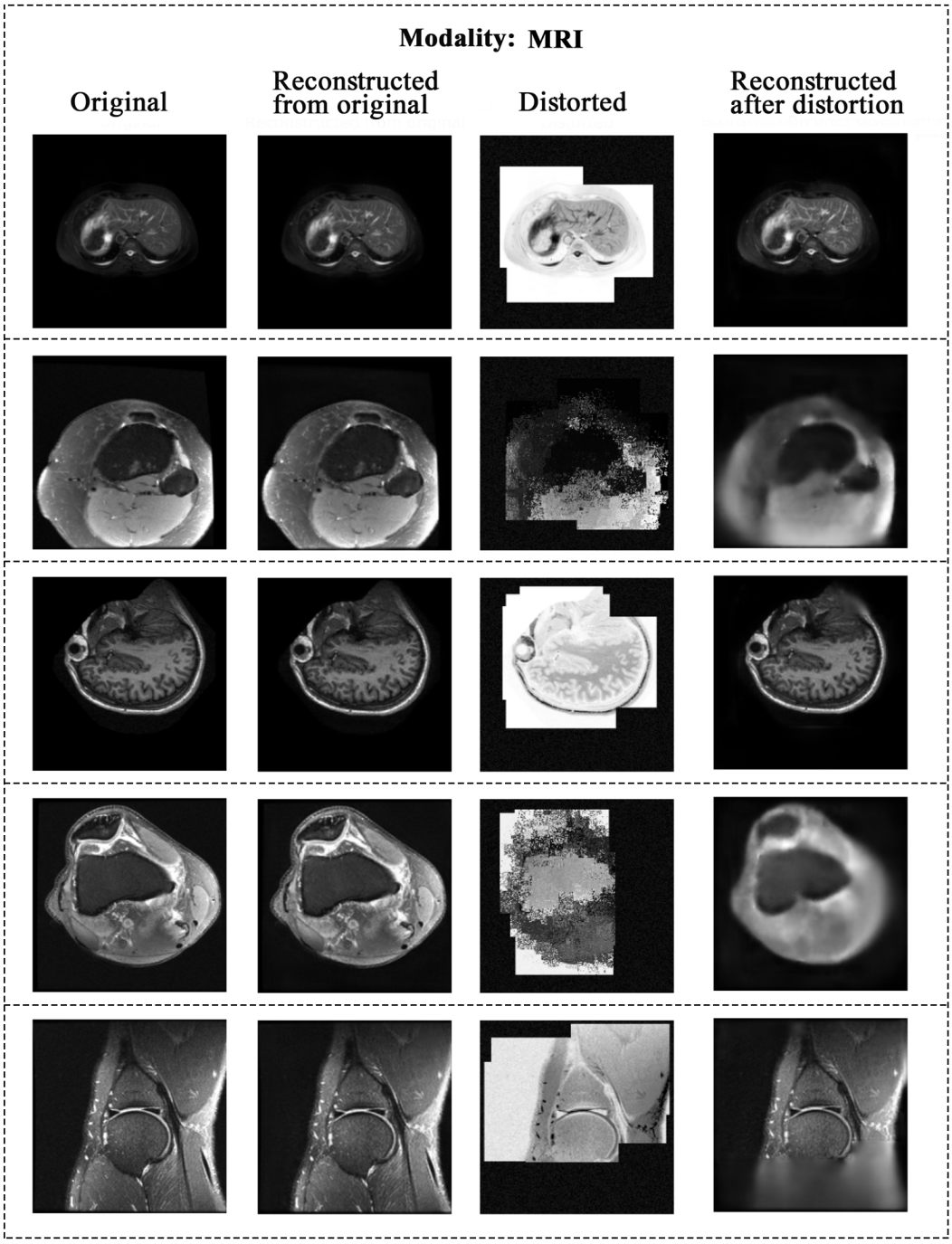


Figure 4.4: Examples of MRI images in the test set.

imbalanced, therefore the classifier 10:1 was weighted in favor of the minority class (positives). Cyclical learning rates have been used to estimate the range of suitable

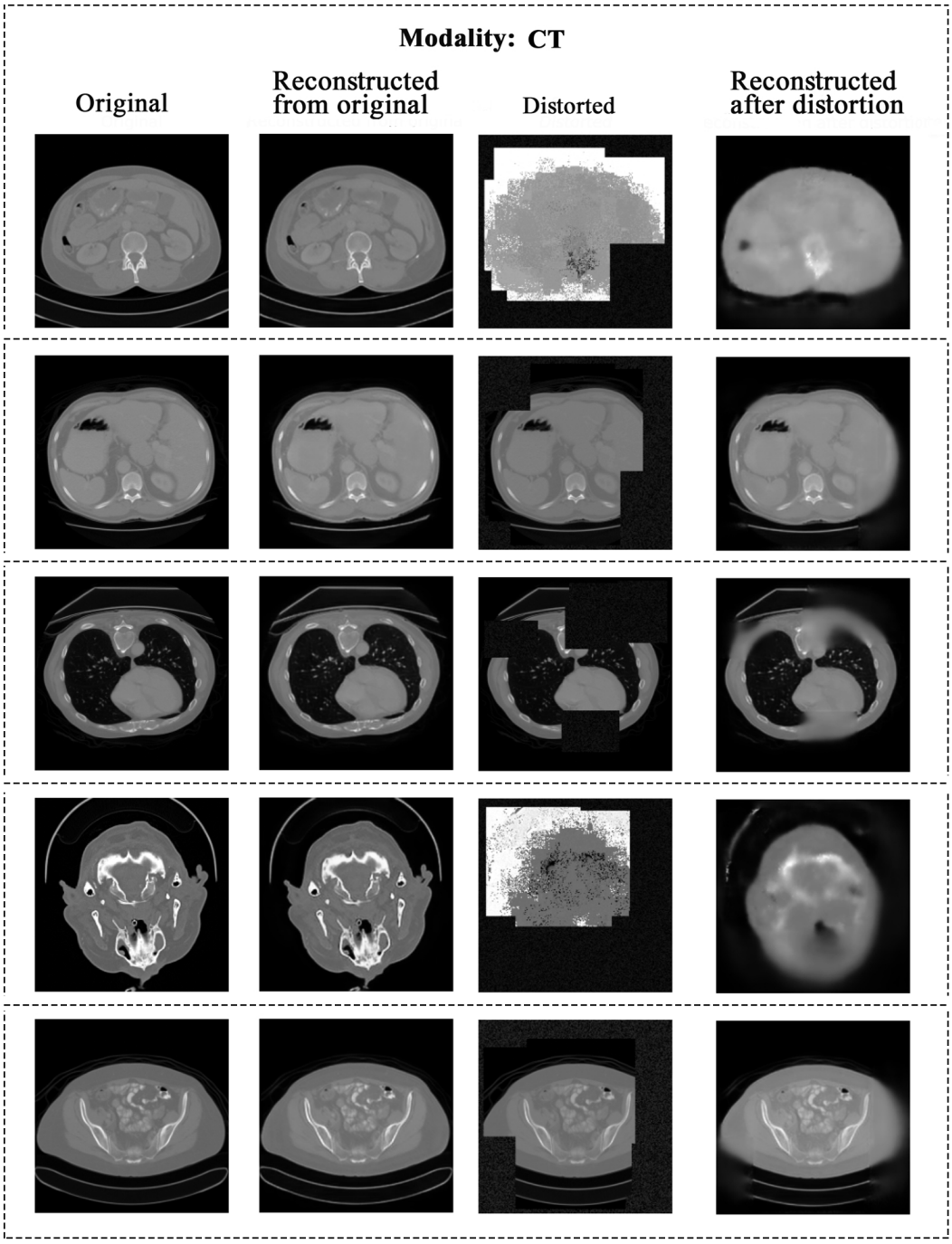


Figure 4.5: Examples of CT images from the test set.

learning rate [120]. For the first step of the training it has been used 0.01 and for

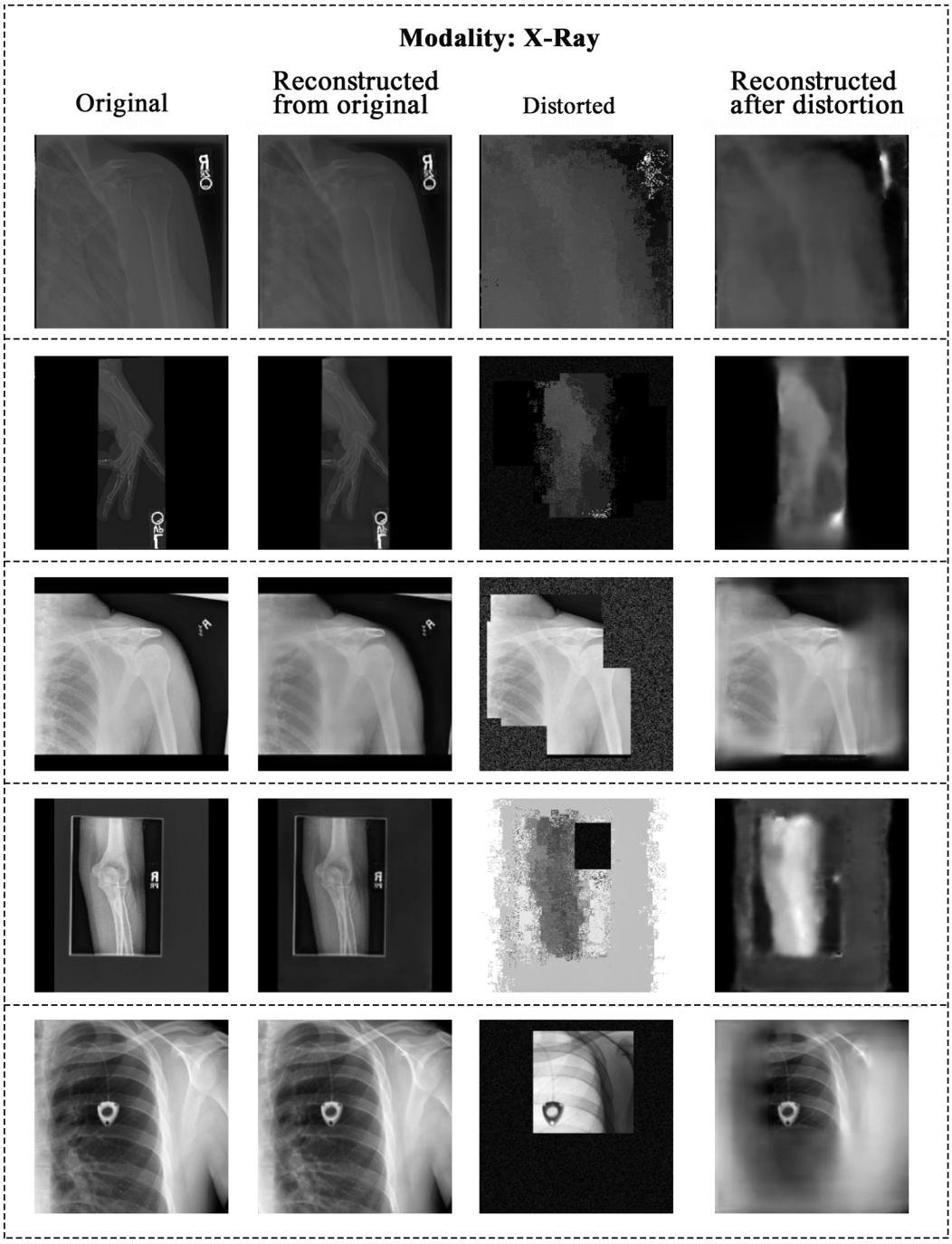


Figure 4.6: Examples of X-Ray images from the test set.

the second step 10^{-4}. It should be noted that experiments showed medical pre-trained weights are much more sensitive to change in learning rate with respect to

ImageNet at the second step. For all variations of experiments with ImageNet the learning rate of 0.01 worked well and the results were coherent with the literature. The batch size in all the experiments was set to 64.

The Chexpert dataset is quite different than Luna16. The target task has multiple labels each referring to a specific pathology that might have been found in the image. The original dataset has 14 different labels, however, the evaluation is done on five selected classes: Atelectasis, Cardiomegaly, Consolidation, Edema, and Pleural Effusio [86]. Since it is a multi-label problem, sigmoid activation have been used instead of softmax in the classification layer. The images were resized and cropped to 320×320. For data augmentation random rotation in range (-0.1, 0.1), random zoom in range (0, 0.1) and random crop have been applied. The learning rate for medical pre-trained weights was set to 0.01, 0.001, and 0.001 for first step, second step and a direct transfer respectively. The learning rate was set to 0.01 for the transfers from ImageNet. Same as Luna16, the transfers from ImageNet are inline with the results in the literature and the pre-trained weights from medical training is sensitive to learning rate. The batch size for all experiments was set to 32.

Moreover, a direct one step transfer with fully trainable layers has also been applied. The only difference in hypermarameters mentioned above is the learning rate. The learning rate selected for Chexpert and LUNA16 for both ImageNet and MedNet is 0.01 which gave the best results. It should be also mentioned that, experimentally, the one step transfer is not as sensitive to the value of the learning rate as the two step transfer.

4.5 Results

4.5.1 Representations

The main purpose of pre-training is to provide the model with a prior understanding of medical images regardless of modality. In this section, the goal is to analyse this hypothesis. The feature space for the pre-training is taken out from the encoder layer which is followed by a pooling layer. Consequently, the number of features after the pooling layer would have a dimension of 2048 which requires finding the right visualization method to analyse how well the representations can describe the dataset. A most common approach is to project the test set into a 2D space. Appropriate projection for analysis needs to preserve the distribution and relative distance between neighbour set of points. Therefore, the representations are projected into a 2D space using the T-SNE method [126]. In short, the T-SNE method works by taking into account the pairwise distance of the samples in hypothesis space based on t-student distribution centered round each point and minimizes the KL divergence between distribution of the data in high dimensional

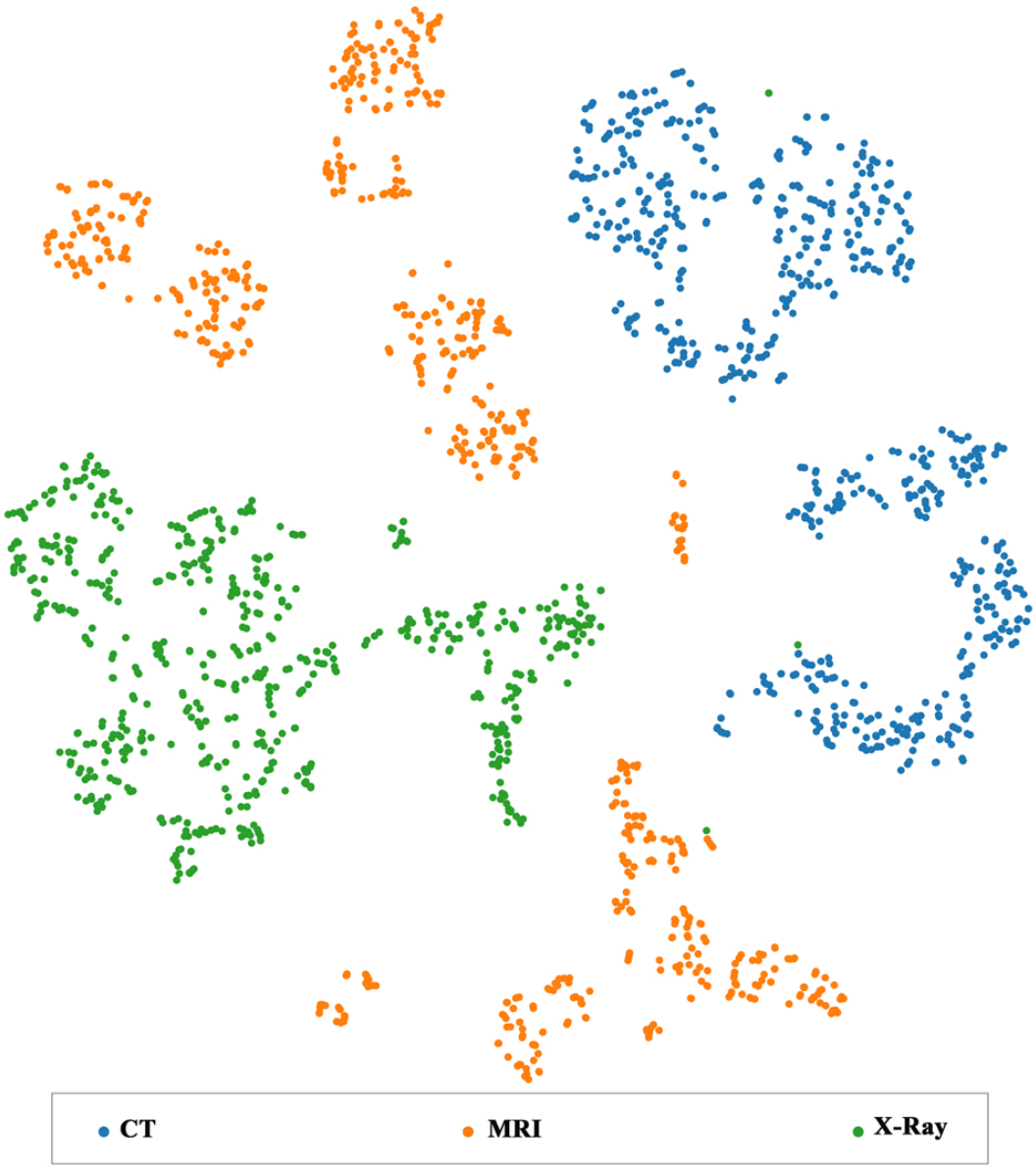


Figure 4.7: T-SNE visualization of the representations colored by modality

space and low dimensional space enabling us to project the test set into a 2D space for visualization.

A random sub-sample of 2000 datapoints were selected from the dataset for visualization. As for the parameters of T-SNE the number of dimensions is set to two and preplexity is set to 18, learning rate of 10, with early stopping active, and z maximum of 1000 iterations. Figures 4.7 and 4.8 shows the results of T-SNE on the

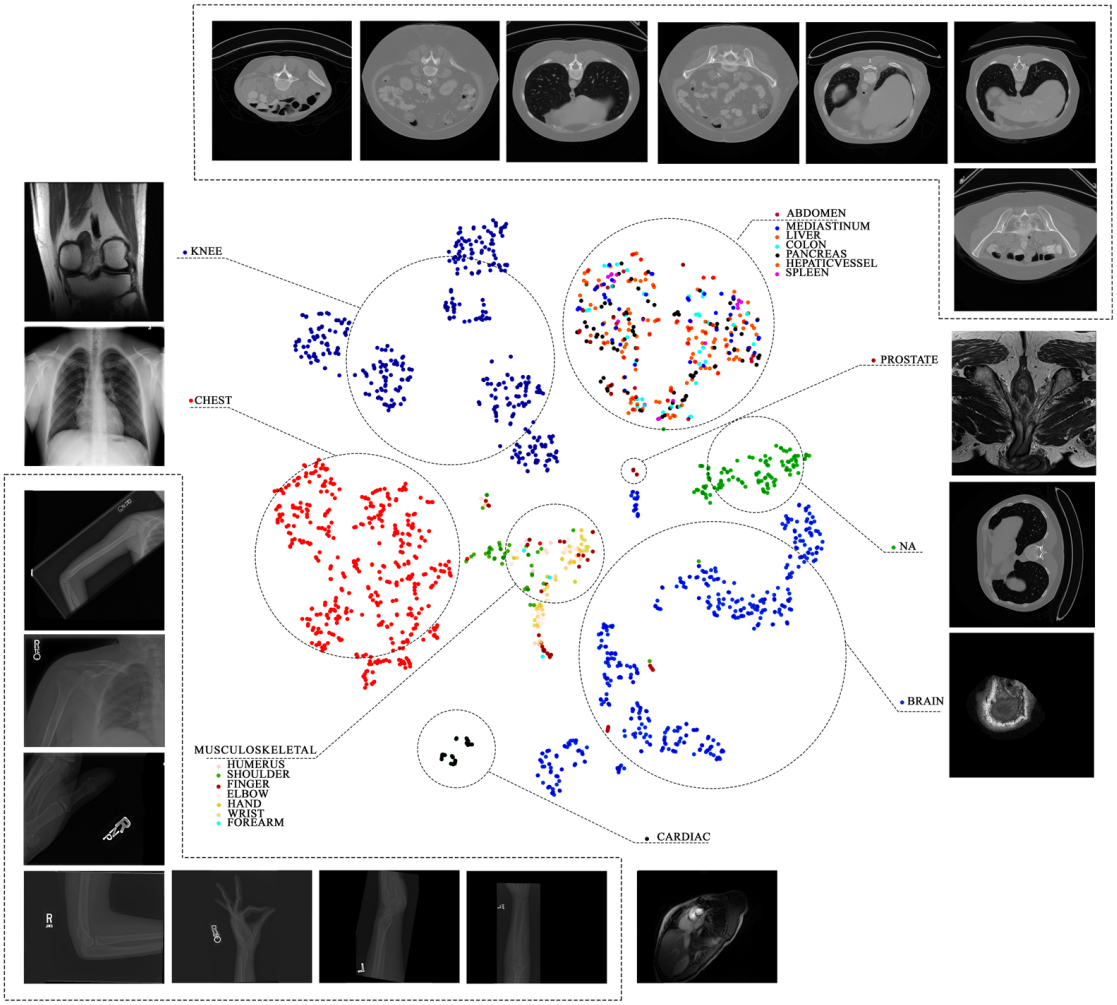


Figure 4.8: The T-SNE embeddings shows that the model is able to cluster similar images together as well as clustering them by modality. Each cluster is marked and on the border of the image example images from each cluster is included.

test set. As shown in Figure 4.7, the modalities are perfectly clustered. In addition to that, an analysis has also been performed on different body parts available in the dataset. It has been shown that besides understanding the modality of the images, the model is also able to cluster body parts based on their relative position in body. For example one cluster contains different datasets on abdominal area. The cluster that is marked as NA comes from the Deep Lesion dataset which contains CT scans that cover from neck to the abdomen area, and as seen in the figure it is placed in the vicinity of abdomen where in both it is possible to observe lungs of the patient. This observation confirms that the similarities in clusters in the latter space is based on overall shape, context, and texture.

Besides the representations, it can also be observed that the model is able to reconstruct the original images from the embedding clearly (Figures 4.5, 4.4, and 4.6). The reconstruction from distorted images, even though blurry, reconstructs the overall structure of the image quite well. It should be noted that in some cases the images have gone through a lot of distortions and the image is still reconstructed relatively well. The reconstructions alongside the T-SNE embeddings show that the model is able to extract local and global features which distinguish body parts, modalities, and textures. In this regard the representation itself is quite superior than pre-trained models from natural images. The next step is to test how well these representations help diagnostic tasks or in other words classification.

Besides the representations provided by the pre-trained network, the aim is also to study how the weights can contribute to a diagnostic problem such as classification. Since the self-supervised framework task is aimed at reconstruction of images, the network has been trained three times with three different image crops. Image crops at different sizes resembles changing the focus of attention from the whole organ to small patches with more details and local structures.

For each dataset, three sets of experiments have been devised, which are shown in Tables 5.1-5.3. The same transfer learning procedure has been applied to three datasets. In each experiment, the objective is to assess transferability of the features learnt at the pre-training stage for classification tasks. Step 1 refers to experiments in which all the convolutional layers are frozen followed by fully connected classification layers with random initialization. The purpose of this step is to evaluate solely the representations. Step 2 refers to experiments in which the training from Step 1 has continued and all convolutional layers have been unfrozen. Finally, Table 5.3, refers to experiments where the transfer learning only contains one step. By comparing these results it is possible to acquire additional information into the features extracted at pre-training step, convergence rate, and the quality of classification for each task. It should be noted that the policy for selecting the initial weights for Step 2 is taking the weights from the last epoch. Since using the one with the best AUC on test could result in biased training.

Dataset	ImageNet	MedNet224	MedNet128
Luna16	$0.548 \pm 1.8e-07$	$0.826 \pm 1.7e-05$	$0.743 \pm 1.4e-03$
Chexpert	$0.568 \pm 1.6e-04$	$0.761 \pm 5.9e-06$	$0.479 \pm 1.0e-03$

Table 4.3: AUC @ step 1

Since all the tasks share the common attributes existing in medical images such as low number of samples, different pixel ranges, and imbalanced datasets, plus the fact that the task in all cases is classification, it has been decided to evaluate each experiment by using the AUC metric. It should also be mentioned that each experiment was executed three times, in order to acquire the variance of the score

Dataset	ImageNet	MedNet224	MedNet128
Luna16	$0.986 \pm 5.2e-06$	$0.983 \pm 3.56e-08$	$0.979 \pm 4.3e-05$
Chexpert	$0.881 \pm 2.36e-05$	$0.879 \pm 1.7-e07$	$0.860 \pm 2.2e-05$

Table 4.4: AUC @ step 2

in different execution which shows the robustness of the training.

Dataset	ImageNet	MedNet224	MedNet128
Luna16	$0.983 \pm 1.0e-05$	$0.980 \pm 1.1e-05$	$0.980 \pm 7.6e-06$
Chexpert	$0.880 \pm 1.5e-05$	$0.878 \pm 1.3e-05$	$0.863 \pm 1.4e-05$

Table 4.5: AUC for one step transfer

The results of transfers from MedNet at Step 1 shows complete superiority of the performance on classification over the transfers from ImageNet across all modalities and input sizes. At Step 2, however, the situation is different where all the different pre-trained settings have comparable performance and it can be observed that ImageNet is even marginally better. As the input size for pre-training decreases the results on transfer learning at second step also decreases.

Dataset	ImageNet	MedNet224	MedNet128
Luna16	(0.973, 0.993)	(0.972, 0.988)	(0.975, 0.984)
Chexpert	(0.868, 0.892)	(0.866, 0.886)	(0.851, 0.874)

Table 4.6: Confidence interval

4.6 Discussion

Although the representations describe the dataset well and close the gap between the pre-trained network and the target task, ImageNet still outperforms the pre-trained networks on medical images. Therefore, a step forward is taken by analyzing the convolutional filters used in the network.

Starting from the earlier layers, we can see that the filters are quite similar. Moving to the middle layers, one can see that both ImageNet and MedNet extract features that represent textures. Although many filters still look similar, some filters appear to be different. This may be because MedNet extracts textures that are better suited to describe medical images.

In the last levels, it can be observed that the filters are quite different. It seems that most filters in ImageNet mainly look for patterns or objects in the center of the

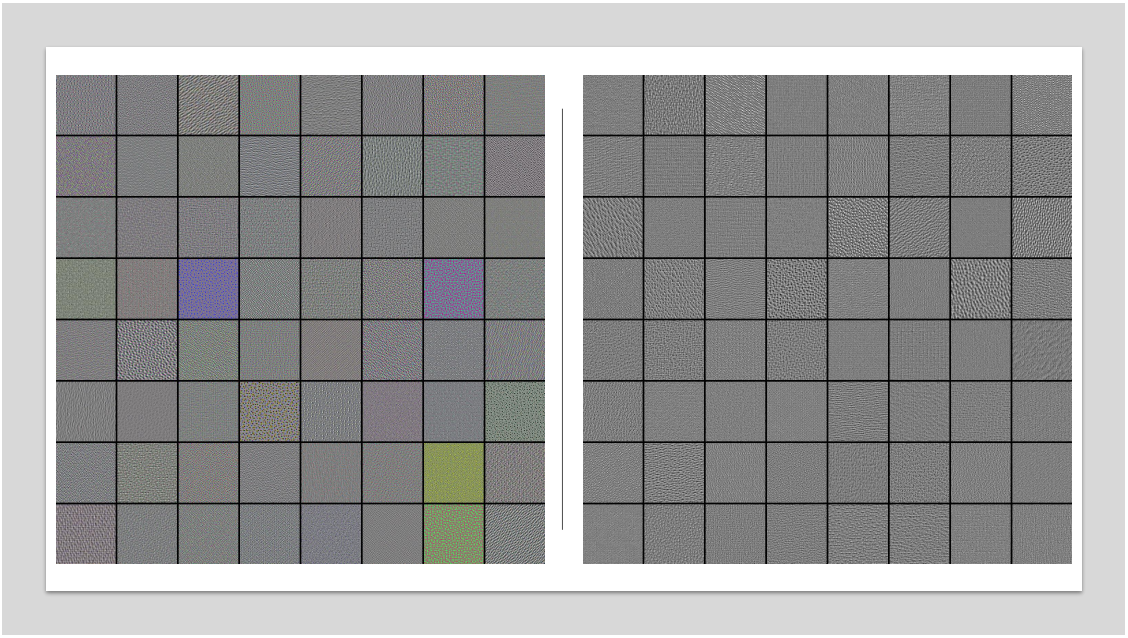


Figure 4.9: Filters of the first convolution layers of the encoder

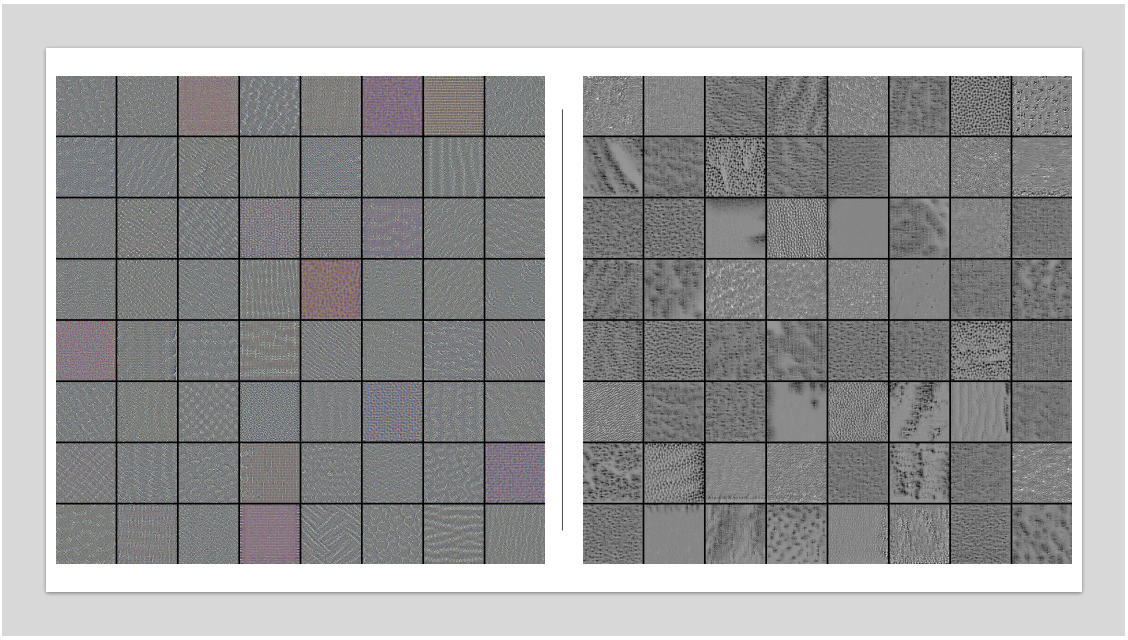


Figure 4.10: Filters of the middle convolution layers of the encoder

image, which help the model to classify images that may contain objects of interest. In the case of transfer to medical diagnosis, the objects may be abnormalities that are present in the image. On the other hand, the MedNet filters have different

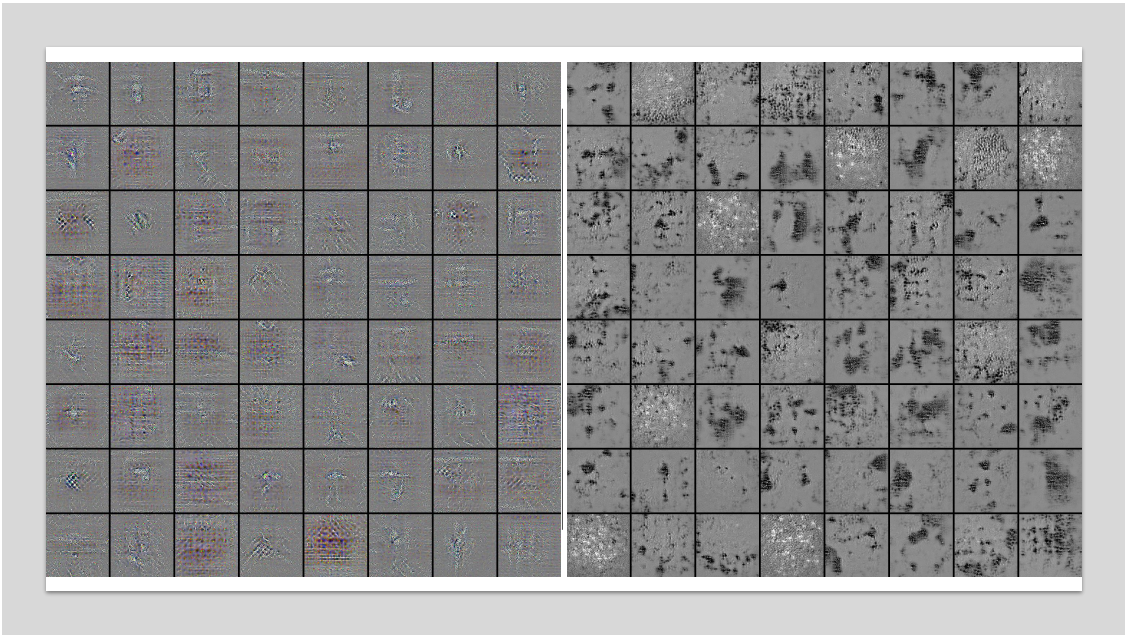


Figure 4.11: Filters of the final convolution layers of the encoder

patterns scattered across each filter that contribute to shape and textures in the dataset needed to reconstruct the images. This indicates that regardless of the distance between the training datasets, the task for which the model was trained has a greater influence.

As mentioned earlier, the main difference between the filters of both models is in the last levels and the rest of the levels extract common features that are mostly the same. Therefore, one can see great improvements in representation but not in classification. As seen in previous literature, such self-supervised methods work best in segmentation and 3D environments where there are no pre-trained models such as ImageNet [144, 143, 8]. For example, in [143], the authors achieve an AUC value of 0.974 when referring to LUNA16 itself and 0.978 when using pre-trained ImageNet weights. On the other hand, by using 3D patches, they achieve an AUC value of 0.983, which outperforms both 2D transfers.

Further study of Luna's abnormal stains contributes to the above theory. The reconstructed spots on the entire image look good. However, when zooming in on the area of the nodule, it can be seen that in some cases the nodule was not clearly reconstructed because it may have been treated as a distortion that the network was trying to correct. This behavior of the model can be seen in Figure 4.12, where we randomly selected 8 data points from the Luna16 test set. In the third row, for example, the node was reconstructed as if the model was trying to reconstruct a non-linear transformation. This suggests that such models are not the best choice for diagnosis and classification in the medical field, as the focus of the model is on

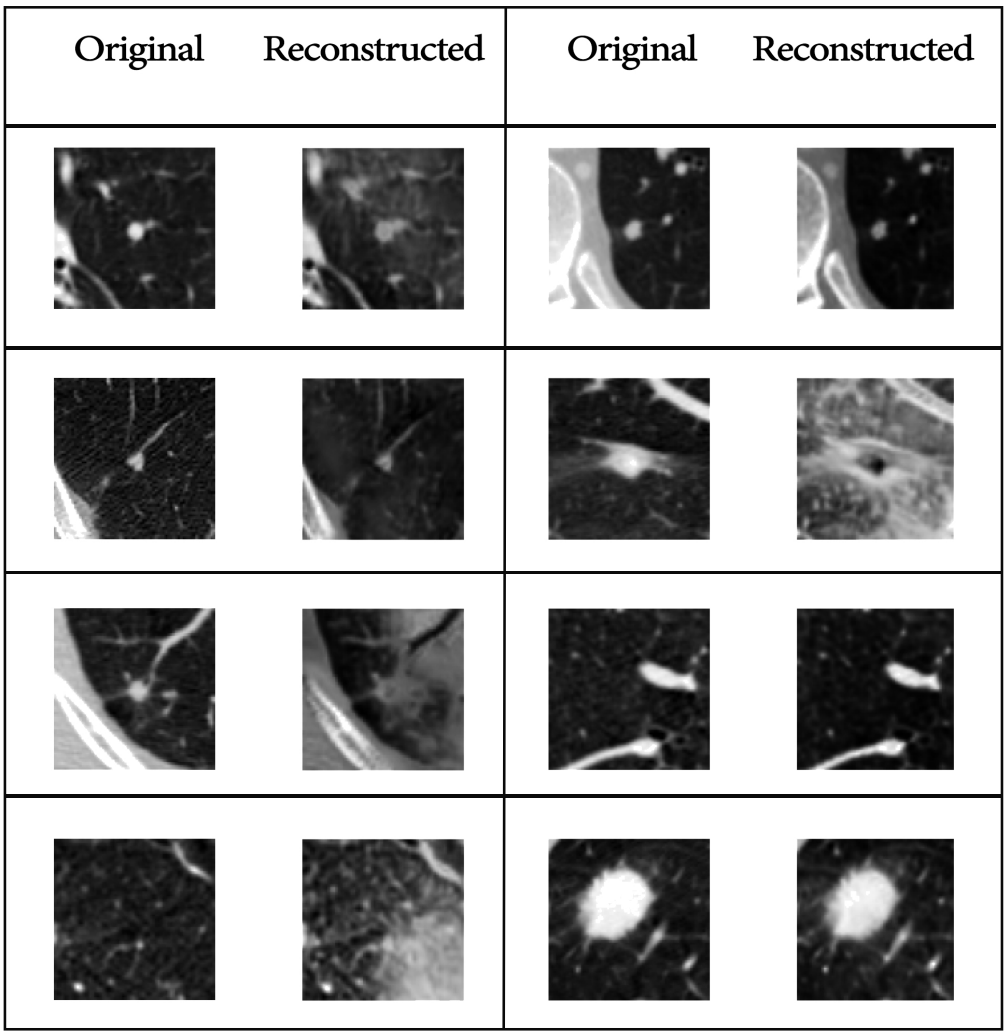


Figure 4.12: Reconstruction of patches from the Luna16 dataset that contain nodules

dominant textures, shape, and global structure of the model rather than features focused on anomalies.

4.7 Conclusion

The work reported herewith analyzed the utilization of self-supervised methods from the diagnosis perspective. Similar works in the field focus on 3D segmentation

and domain specific pre-training, however, the MedNet model presented here has been trained on half a million images from different modalities and body-parts. The representations provided by the encoder work quite well in describing medical images regardless of their modality. In case of freezing layers and only training FC layers, it has been observed that the model is able to reach a much better AUC. It should be noted that the training procedure is faster given that most layers are not trained compared to a full training. Yet, the results on fully trained network show that the AUC is the same whether the model transfers from the ImageNet or MedNet. As seen from performed investigations, even though representations are much more descriptive, in case of rare abnormal details in the image the focus of the network is on general patterns that exist in medical images and their reconstruction, whereas ImageNet is better at classification since the major difference in the models occur in mid-final layers. Previous literature has showed that such models work better in case of 3D segmentation, but there are no ImageNet alternatives in that case for comparison. In conclusion, ImageNet weights are built based on classification task, which shows that the task the model has been trained on contributes more to transfer learning with respect to the gap between the modalities of source and target datasets.

4.8 Future work

Even though the pre-trained weights are not suitable for classification, the representations given by the network can be utilized in other scenarios. For example, the pre-trained weights can be used for transferring to other tasks such as segmentation or object detection in medical images. The model is generative, therefore, it can help in image retrieval and registration tasks.

This research shows the potential of multi-modal pre-training for medical images and the importance of selecting the right self-supervised task. For example, contrastive and clustering methods may improve the classification performance since the tasks are in coherence with respect to image reconstruction.

The reported work only focused on no reliance on the data, therefore including any supervision cues in the training stage would be out of scope. However, considering supervised cues can help in improving the model towards detecting abnormalities in medical images. For example, as mentioned earlier, datasets are partially labeled and they can be utilized in the training stage by including a penalty in the loss function to guide the training towards focusing on the most common target tasks in each specific modality/body part.

Chapter 5

Conclusion

Deep Learning has performed best in the supervised environment, benefiting from its ability to process sufficiently large data sets. The main challenge limiting the use of Deep Learning for medical images is the lack of sufficiently large annotated datasets for training. This research is about exploring different methods to help use practical data instead of curated and research-ready datasets. Following this idea, several avenues were explored that would benefit the use of Deep Learning frameworks in practice.

The first step was to quantitatively analyze the effects of noisy bounding boxes and the shortcomings of the widely known IoU. Real-world data are not annotated in the same way as curated datasets, as experts may use larger bounding boxes to provide context. Therefore, the approach proposed in Chapter 2 will relax the annotation requirements, which not only makes the training more robust, but also saves time and reduces annotation costs. The proposed method can be easily implemented and does not affect the complexity of the training. Consequently, deep neural networks can be robustly trained with routinely recorded annotations from radiologists.

Following this work, another challenge in mammography was tackled. In this field, the radiologist uses both CC and MLO views for diagnosis. In the work presented, an attempt was made to mimic this behavior by developing a multi-stream object detector that uses both images as input. Therefore, a step towards matching the two images was chosen, which is challenging since it should be done in an unsupervised manner. The system consists of an affine registration network for global alignment of the images, followed by a deformable registration module that handles local and non-affine transformations due to tissue compression. The main innovation is the use of generalized IoU in the loss to incorporate supervised cues to the location of masses present in both images. The framework provides the opportunity to use the information scattered in both views.

Eventually, research focused on the application of self-monitored methods in the medical field. The fact that self-monitored labels do not require manual labeling

allows training on large data sets. To this end, a sufficiently large dataset of half a million images was acquired for pre-training ResNet50 with a framework called Model Genesis[cite] to analyze this method in the medical domain. It has been shown that while the representation works quite well for describing images in terms of the famous pre-trained ImageNet weights, ImageNet still outperforms it when applied to the target task without freezing layers. Further research has shown that this is due to the fact that earlier layers tend to extract the same information, mostly textures. Due to the task ImageNet is pre-trained on, classification, the model looks for objects in the middle of the image in the last layers. The pre-trained medical model, on the other hand, focuses on details to reconstruct the image. Therefore, when transferred to the target task, image classification, the pre-trained ImageNet weights require much less adaptation effort. The research conducted has shown that while self-supervised techniques work well in describing images from different modalities and body parts, they can lead to negative transfer with respect to the target task. Therefore, it is not enough to just close the gap in the area before training, but the task itself is very important.

Bibliography

[1] Richa Agarwal et al. "Automatic mass detection in mammograms using deep convolutional neural networks". In: *J. Med. Imaging* 6.3 (2019), p. 031409.

[2] F Alfano et al. "Prone to Supine Surface Based Registration Workflow for Breast Tumor Localization in Surgical Planning". In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE. 2019, pp. 1150–1153.

[3] Alexander Andreopoulos and John K Tsotsos. "Efficient and generalizable statistical models of shape and appearance for analysis of cardiac MRI". In: *Medical image analysis* 12.3 (2008), pp. 335–357.

[4] Michela Antonelli et al. "The medical segmentation decathlon". In: *arXiv preprint arXiv:2106.05735* (2021).

[5] Miguel Areia et al. "Cost-effectiveness of artificial intelligence for screening colonoscopy: a modelling study". In: *The Lancet Digital Health* (2022).

[6] Paola Armaroli et al. "A randomised controlled trial of Digital Breast Tomosynthesis versus Digital Mammography as primary screening tests: screening results over subsequent episodes of the Proteus Donna study". In: *International Journal of Cancer* (2022).

[7] Samuel G Armato III et al. "The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans". In: *Medical physics* 38.2 (2011), pp. 915–931.

[8] Shekoofeh Azizi et al. "Big self-supervised models advance medical image classification". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 3478–3488.

[9] Guha Balakrishnan et al. "VoxelMorph: a learning framework for deformable medical image registration". In: *IEEE transactions on medical imaging* 38.8 (2019), pp. 1788–1800.

[10] Andriy I Bandos et al. "Area under the free-response ROC curve (FROC) and a related summary index". In: *Biometrics* 65.1 (2009), pp. 247–256.

[11] Chandradeep Bhatt et al. “The state of the art of deep learning models in medical science and their challenges”. In: *Multimedia Systems* 27.4 (2021), pp. 599–613.

[12] Nicholas Bien et al. “Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet”. In: *PLoS medicine* 15.11 (2018), e1002699.

[13] Rishi Bommasani et al. “On the opportunities and risks of foundation models”. In: *arXiv preprint arXiv:2108.07258* (2021).

[14] Mireille Broeders et al. “The impact of mammographic screening on breast cancer mortality in Europe: a review of observational studies”. In: *Journal of medical screening* 19.1_suppl (2012), pp. 14–25.

[15] Mathilde Caron et al. “Unsupervised learning of visual features by contrasting cluster assignments”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 9912–9924.

[16] Kenny H Cha et al. “Evaluation of data augmentation via synthetic images for improved breast mass detection on mammograms using deep learning”. In: *Journal of Medical Imaging* 7.1 (2019), p. 012703.

[17] Kenny H Cha et al. “Reducing overfitting of a deep learning breast mass detection algorithm in mammography using synthetic images”. In: *Medical Imaging 2019: Computer-Aided Diagnosis*. Vol. 10950. International Society for Optics and Photonics. 2019, p. 1095004.

[18] Liang-Chieh Chen et al. “DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40 (2018), pp. 834–848.

[19] Ting Chen et al. “A simple framework for contrastive learning of visual representations”. In: *International conference on machine learning*. PMLR. 2020, pp. 1597–1607.

[20] Yaru Chen et al. “Professionals’ responses to the introduction of AI innovations in radiology and their implications for future adoption: a qualitative study”. In: *BMC Health Services Research* 21.1 (2021), pp. 1–9.

[21] Veronika Cheplygina, Marleen de Bruijne, and Josien PW Pluim. “Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis”. In: *Medical image analysis* 54 (2019), pp. 280–296.

[22] Sasank Chilamkurthy et al. “Development and validation of deep learning algorithms for detection of critical findings in head CT scans”. In: *arXiv preprint arXiv:1803.05854* (2018).

[23] Francois Chollet. *Deep learning with Python*. Simon and Schuster, 2021.

[24] Kenneth Clark et al. "The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository". In: *Journal of digital imaging* 26.6 (2013), pp. 1045–1057.

[25] JJJ Condon et al. "Replication of an open-access deep learning system for screening mammography: Reduced performance mitigated by retraining on local data". In: *medRxiv* (2021).

[26] Karin Dembrower et al. "Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective simulation study." In: *The Lancet. Digital health* 2 9 (2020), e468–e474.

[27] *Digital Mammography DREAM Challenge, 2017*. https://www.synapse.org/#!Synapse:syn4224222/wiki/434546. Accessed: 2019-11-26.

[28] Carl Doersch and Andrew Zisserman. "Multi-task self-supervised visual learning". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2051–2060.

[29] Richard O Duda and Peter E Hart. "Use of the Hough transformation to detect lines and curves in pictures". In: *Communications of the ACM* 15.1 (1972), pp. 11–15.

[30] Saskia van Engeland et al. "A comparison of methods for mammogram registration". In: *IEEE Transactions on Medical Imaging* 22.11 (2003), pp. 1436–1444.

[31] Linus Ericsson, Henry Gouk, and Timothy M Hospedales. "How well do self-supervised models transfer?" In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 5414–5423.

[32] Andre Esteva et al. "A guide to deep learning in healthcare". In: *Nature medicine* 25.1 (2019), pp. 24–29.

[33] Sina Famouri, Lia Morra, and Fabrizio Lamberti. "A Deep Learning Approach for Efficient Registration of Dual View Mammography". In: *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*. Springer. 2020, pp. 162–172.

[34] Sina Famouri et al. "Breast Mass Detection With Faster R-CNN: On the Feasibility of Learning From Noisy Annotations". In: *IEEE Access* 9 (2021), pp. 66163–66175.

[35] Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. "Robustness of classifiers: from adversarial to random noise". In: *Advances in neural information processing systems* 29 (2016).

[36] Benoıt Frénay and Ata Kabán. "A Comprehensive Introduction to Label Noise". In: ESANN 2014 proceedings, 2014.

[37] Benoıt Frénay and Michel Verleysen. "Classification in the presence of label noise: a survey". In: *IEEE T. Neur. Net. Lear.* 25.5 (2014), pp. 845–869.

[38] Roberta Fusco, Vincenza Granata, and Antonella Petrillo. "Introduction to special issue of radiology and imaging of cancer". In: *Cancers* 12.9 (2020), p. 2665.

[39] Jiyang Gao et al. "NOTE-RCNN: noise tolerant ensemble RCNN for semi-supervised object detection". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2019, pp. 9508–9517.

[40] Ross Girshick et al. "Rich feature hierarchies for accurate object detection and semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 580–587.

[41] Julien Guiot et al. "A review in radiomics: Making personalized medicine a reality via routine imaging". In: *Medicinal Research Reviews* 42.1 (2022), pp. 426–440.

[42] Yujun Guo et al. "Breast image registration techniques: a survey". In: *Medical and Biological Engineering and Computing* 44.1-2 (2006), pp. 15–26.

[43] Lubomir Hadjiiski et al. "Automated registration of breast lesions in temporal pairs of mammograms for interval change analysis—Local affine transformation for improved localization". In: *Medical physics* 28.6 (2001), pp. 1070–1079.

[44] Fatemeh Haghighi et al. "Dira: Discriminative, restorative, and adversarial learning for self-supervised medical image analysis". In: *arXiv preprint arXiv:2204.10437* (2022).

[45] Bo Han et al. "Co-teaching: Robust training of deep neural networks with extremely noisy labels". In: *Advances in Neural Information Processing Systems*. 2018, pp. 8527–8537.

[46] Grant Haskins, Uwe Kruger, and Pingkun Yan. "Deep learning in medical image registration: a survey". In: *Machine Vision and Applications* 31.1 (2020), p. 8.

[47] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[48] Michael Heath et al. "The digital database for screening mammography". In: *Proceedings of the 5th international workshop on digital mammography*. Medical Physics Publishing. 2000, pp. 212–218.

[49] Dan Hendrycks et al. "Using trusted data to train deep networks on labels corrupted by severe noise". In: *Advances in neural information processing systems*. 2018, pp. 10456–10465.

[50] Regina J Hooley. "Breast density legislation and clinical evidence". In: *Radiologic Clinics* 55.3 (2017), pp. 513–526.

[51] Mohammad Reza Hosseinzadeh Taher et al. "A Systematic Benchmarking Analysis of Transfer Learning for Medical Image Analysis". In: *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*. Springer, 2021, pp. 3–13.

[52] Yipeng Hu et al. "Weakly-supervised convolutional neural networks for multimodal image registration". In: *Medical image analysis* 49 (2018), pp. 1–13.

[53] Eui Jin Hwang et al. "Use of artificial intelligence-based software as medical devices for chest radiography: a position paper from the Korean Society of Thoracic Radiology". In: *Korean Journal of Radiology* 22.11 (2021), p. 1743.

[54] *Internet Brain Segmentation Repository (IBSR)*. URL: http://www.cma.mgh.harvard.edu/ibsr/.

[55] Jeremy Irvin et al. "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison". In: *arXiv preprint arXiv:1901.07031* (2019).

[56] Max Jaderberg et al. "Spatial Transformer Networks". In: *Advances in Neural Information Processing Systems 28*. 2015, pp. 2017–2025.

[57] Hwejin Jung et al. "Detection of masses in mammograms using a one-stage object detector based on a deep convolutional neural network". In: *PloS one* 13.9 (2018).

[58] Davood Karimi et al. "Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis". In: *Medical Image Analysis* 65 (2020), p. 101759.

[59] A. Emre Kavur et al. "CHAOS Challenge - combined (CT-MR) healthy abdominal organ segmentation". In: *Medical Image Analysis* 69 (2021), p. 101950.

[60] Mahdi Khosravy et al. "Deep Face Recognizer Privacy Attack: Model Inversion Initialization by a Deep Generative Adversarial Data Space Discriminator". In: *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE. 2020, pp. 1400–1405.

[61] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[62] Marc D Kohli, Ronald M Summers, and J Raymond Geis. "Medical image data and datasets in the era of machine learning: Whitepaper from the 2016 C-MIMI meeting dataset session". In: *Journal of digital imaging* 30.4 (2017), pp. 392–399.

[63] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. "Revisiting self-supervised visual representation learning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 1920–1929.

[64] Thijs Kooi et al. "Large scale deep learning for computer aided detection of mammographic lesions". In: *Medical image analysis* 35 (2017), pp. 303–312.

[65] David Kügler and Anirban Mukhopadhyay. "How Bad is Good enough: Noisy annotations for instrument pose estimation". In: *arXiv preprint arXiv:1806.07836* (2018).

[66] Thomas C Kwee and Robert M Kwee. "Workload of diagnostic radiologists in the foreseeable future based on recent scientific advances: growth expectations and role of artificial intelligence". In: *Insights into imaging* 12.1 (2021), pp. 1–12.

[67] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *nature* 521.7553 (2015), pp. 436–444.

[68] Rebecca Sawyer Lee et al. "A curated mammography data set for use in computer-aided detection and diagnosis research". In: *Scientific data* 4.1 (2017), pp. 1–9.

[69] Rebecca Sawyer Lee et al. *Curated Breast Imaging Subset of DDSM [Dataset]. The Cancer Imaging Archive*. https://doi.org/10.7937/K9/TCIA.2016.7O02S9CY. 2016.

[70] Daniel Lévy and Arzav Jain. "Breast mass classification from mammograms using deep convolutional neural networks". In: *arXiv preprint arXiv:1612.00542* (2016).

[71] Hongming Li and Yong Fan. "Non-rigid image registration using self-supervised fully convolutional networks without training data". In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE. 2018, pp. 1075–1078.

[72] Rui Liao et al. "An artificial agent for robust image registration". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 31. 1. 2017.

[73] Tsung-Yi Lin et al. "Focal Loss for Dense Object Detection". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42 (2020), pp. 318–327.

[74] Tsung-Yi Lin et al. "Focal loss for dense object detection". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988.

[75] Geert Litjens et al. "A survey on deep learning in medical image analysis". In: *Medical image analysis* 42 (2017), pp. 60–88.

[76] Yuhang Liu et al. "Act Like a Radiologist: Towards Reliable Multi-view Correspondence Reasoning for Mammogram Mass Detection". In: *IEEE transactions on pattern analysis and machine intelligence* PP (2021).

[77] JB Antoine Maintz and Max A Viergever. "A survey of medical image registration". In: *Medical image analysis* 2.1 (1998), pp. 1–36.

[78] Daniel S Marcus et al. "Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults". In: *Journal of cognitive neuroscience* 19.9 (2007), pp. 1498–1507.

[79] Mohammed A Al-masni et al. "Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system". In: *Computer methods and programs in biomedicine* 157 (2018), pp. 85–94.

[80] Scott Mayer McKinney et al. "International evaluation of an AI system for breast cancer screening". In: *Nature* 577.7788 (2020), pp. 89–94.

[81] Shun Miao et al. "Dilated FCN for multi-agent 2D/3D medical image registration". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1. 2018.

[82] John D Miller et al. "Self-Supervised Deep Learning to Enhance Breast Cancer Detection on Screening Mammography". In: *arXiv preprint arXiv:2203.08812* (2022).

[83] Inês C Moreira et al. "Inbreast: toward a full-field digital mammographic database". In: *Academic radiology* 19.2 (2012), pp. 236–248.

[84] Lia Morra, Silvia Delsanto, and Loredana Correale. *Artificial intelligence in medical imaging: From theory to clinical practice*. CRC Press, 2019.

[85] Lia Morra et al. "Breast cancer: computer-aided detection with digital breast tomosynthesis". In: *Radiology* 277.1 (2015), pp. 56–63.

[86] Lia Morra et al. "Bridging the gap between natural and medical images through deep colorization". In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE. 2021, pp. 835–842.

[87] Nagarajan Natarajan et al. "Learning with noisy labels". In: *Advances in neural information processing systems*. 2013, pp. 1196–1204.

[88] Sisse Njor et al. "Breast cancer mortality in mammographic screening in Europe: a review of incidence-based mortality studies". In: *Journal of medical screening* 19.1_suppl (2012), pp. 33–41.

[89] William C Ou, Dogan Polat, and Basak E Dogan. "Deep learning in breast radiology: current progress and future directions". In: *European Radiology* 31.7 (2021), pp. 4872–4885.

[90] Shaked Perek et al. "Siamese Network for Dual-View Mammography Mass Matching". In: *RAMBO+BIA+TIA@MICCAI*. 2018.

[91] Shaked Perek et al. "Siamese network for dual-view mammography mass matching". In: *Image Analysis for Moving Organ, Breast, and Thoracic Images*. Springer, 2018, pp. 55–63.

[92] Nicholas Petrick et al. "Evaluation of computer-aided detection and diagnosis systemsa". In: *Medical physics* 40.8 (2013).

[93] Chen Qin et al. "Joint learning of motion estimation and segmentation for cardiac MR image sequences". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 472–480.

[94] Maithra Raghu et al. "Transfusion: Understanding transfer learning for medical imaging". In: *Advances in neural information processing systems* 32 (2019).

[95] Pranav Rajpurkar et al. "Mura: Large dataset for abnormality detection in musculoskeletal radiographs". In: *arXiv preprint arXiv:1712.06957* (2017).

[96] José Luis Raya-Povedano et al. "AI-based Strategies to Reduce Workload in Breast Cancer Screening with Mammography and Tomosynthesis: A Retrospective Evaluation". In: *Radiology* 300.1 (2021), pp. 57–65.

[97] Joseph Redmon et al. "You only look once: Unified, real-time object detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 779–788.

[98] Scott Reed et al. "Training deep neural networks on noisy labels with bootstrapping". In: *arXiv preprint arXiv:1412.6596* (2014).

[99] Shaoqing Ren et al. "Faster R-CNN: Towards real-time object detection with region proposal networks". In: *Advances in neural information processing systems*. 2015, pp. 91–99.

[100] Yinhao Ren et al. "Retina-Match: Ipsilateral Mammography Lesion Matching in a Single Shot Detection Pipeline". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2021, pp. 345–354.

[101] Hamid Rezatofighi et al. "Generalized intersection over union: A metric and a loss for bounding box regression". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 658–666.

[102] Dezső Ribli et al. “Detecting and classifying lesions in mammograms with Deep Learning”. In: *Scientific reports* 8.1 (2018), p. 4165.

[103] Abi Rimmer. “Radiologist shortage leaves patient care at risk, warns royal college”. In: *BMJ: British Medical Journal (Online)* 359 (2017).

[104] Alejandro Rodriguez-Ruiz et al. “Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists”. In: *Journal of the National Cancer Institute* 111 (Mar. 2019).

[105] David Rolnick et al. “Deep learning is robust to massive label noise”. In: (2017). arXiv: 1705.10694.

[106] Miguel Romero et al. “Targeted transfer learning to improve performance in small medical physics datasets”. In: *Medical Physics* 47.12 (2020), pp. 6246–6256.

[107] Holger R Roth et al. “A new 2.5 D representation for lymph node detection using random sets of deep convolutional neural network observations”. In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2014, pp. 520–527.

[108] Daniela Sacchetto et al. “Mammographic density: Comparison of visual assessment with fully automatic calculation on a multivendor dataset”. In: *European radiology* 26.1 (2016), pp. 175–183.

[109] Berkman Sahiner et al. “Deep learning in medical imaging and radiation therapy”. In: *Medical physics* 46.1 (2019), e1–e36.

[110] Mattie Salim et al. “External Evaluation of 3 Commercial Artificial Intelligence Algorithms for Independent Assessment of Screening Mammograms”. In: *JAMA Oncology* 6 (Aug. 2020).

[111] Maha Sallam and Kevin Bowyer. “Registering time sequences of mammograms using a two-dimensional image unwarping technique”. In: *Second International Workshop on Digital Mammography*. 1994, pp. 121–130.

[112] Maurice Samulski and Nico Karssemeijer. “Optimizing case-based detection performance in a multiview CAD system for mammography”. In: *IEEE Transactions on Medical Imaging* 30.4 (2011), pp. 1001–1009.

[113] Thomas Schaffter et al. “Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms”. In: *JAMA network open* 3.3 (2020), e200265–e200265.

[114] Ioannis Sechopoulos. “A review of breast tomosynthesis. Part I. The image acquisition process”. In: *Medical physics* 40.1 (2013), p. 014301.

[115] Debapriya Sengupta, Phalguni Gupta, and Arindam Biswas. “A survey on mutual information based medical image registration algorithms”. In: *Neurocomputing* 486 (2022), pp. 174–188.

[116] Arnaud Arindra Adiyoso Setio et al. "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge". In: *Medical image analysis* 42 (2017), pp. 1–13.

[117] Yiqiu Shen et al. "An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization". In: *Medical image analysis* 68 (2021), p. 101908.

[118] Saeed Shurrab and Rehab Duwairi. "Self-supervised learning methods and applications in medical imaging analysis: A survey". In: *arXiv preprint arXiv :2109.08685* (2021).

[119] Priscilla J Slanetz, Phoebe E Freer, and Robyn L Birdwell. "Breast-density legislation–practical considerations." In: *The New England Journal of Medicine* 372.7 (2015), pp. 593–595.

[120] Leslie N Smith. "Cyclical learning rates for training neural networks". In: *2017 IEEE winter conference on applications of computer vision (WACV)*. IEEE. 2017, pp. 464–472.

[121] Martina Sollini et al. "Artificial intelligence and hybrid imaging: the best match for personalized medicine in oncology". In: *European Journal of Hybrid Imaging* 4.1 (2020), pp. 1–22.

[122] Sainbayar Sukhbaatar et al. "Training convolutional networks with noisy labels". In: *arXiv preprint arXiv:1406.2080* (2014).

[123] Tung Tran and Ramakanth Kavuluru. "Distant supervision for treatment relation extraction by leveraging MeSH subheadings". In: *Artificial intelligence in medicine* 98 (2019), pp. 18–26.

[124] Tuan Truong, Sadegh Mohammadi, and Matthias Lenga. "How transferable are self-supervised features in medical image classification tasks?" In: *Machine Learning for Health*. PMLR. 2021, pp. 54–74.

[125] Pascal Vagssa et al. "Pectoral muscle deletion on a mammogram to aid in the early diagnosis of breast cancer". In: *International Journal of Engineering, Science and Technology* 12.3 (2020), pp. 57–65.

[126] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE". In: *Journal of machine learning research* 9.11 (2008).

[127] Guido Van Schie et al. "Correlating locations in ipsilateral breast tomosynthesis views using an analytical hemispherical compression model". In: *Physics in Medicine & Biology* 56.15 (2011), p. 4715.

[128] Max A Viergever et al. *A survey of medical image registration–under review*. 2016.

[129] Bob D de Vos et al. “End-to-end unsupervised deformable image registration with a convolutional neural network”. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2017, pp. 204–212.

[130] Xiaosong Wang et al. “ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2097–2106.

[131] Michael A Wirth, Jay Narhan, and Derek WS Gray. “Nonrigid mammogram registration using mutual information”. In: *Medical Imaging 2002: Image Processing*. Vol. 4684. International Society for Optics and Photonics. 2002, pp. 562–573.

[132] N. Wu et al. “Improving the Ability of Deep Neural Networks to Use Information from Multiple Views in Breast Cancer Screening”. In: *Proceedings of machine learning research* 121 (2020), pp. 827–842.

[133] Nan Wu et al. “Deep neural networks improve radiologists’ performance in breast cancer screening”. In: *IEEE transactions on medical imaging* (2019).

[134] Nan Wu et al. “The NYU breast cancer screening dataset V1. 0”. In: *New York Univ., New York, NY, USA, Tech. Rep* (2019).

[135] Yang Xin et al. “Machine learning and deep learning methods for cybersecurity”. In: *IEEE Access* 6 (2018), pp. 35365–35381.

[136] Ke Yan et al. “DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning”. In: *J. Med. Imaging* 5.3 (2018), p. 036501.

[137] Zhicheng Yang et al. “MommiNet: Mammographic Multi-view Mass Identification Networks”. In: *MICCAI*. 2020.

[138] P. Yi et al. “DeepCAT: Deep Computer-Aided Triage of Screening Mammography”. In: *Journal of Digital Imaging* (2021), pp. 1–9.

[139] Syed Sahil Abbas Zaidi et al. “A Survey of Modern Deep Learning based Object Detection Models”. In: *arXiv preprint arXiv:2104.11892* (2021).

[140] Federica Zanca et al. “Correlation of free-response and receiver-operating-characteristic area-under-the-curve estimates: Results from independently conducted FROC/ROC studies in mammography”. In: *Medical physics* 39.10 (2012), pp. 5917–5929.

[141] Chiyuan Zhang et al. “Understanding deep learning requires rethinking generalization”. In: *arXiv preprint arXiv:1611.03530* (2016).

[142] Zhaohui Zheng et al. “Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression”. In: *arXiv preprint arXiv:1911.08287* (2019).

[143] Zongwei Zhou et al. “Models genesis: Generic autodidactic models for 3d medical image analysis”. In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2019, pp. 384–393.

[144] Xinrui Zhuang et al. “Self-supervised feature learning for 3d medical images by playing a rubik’s cube”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 420–428.

This Ph.D. thesis has been typeset by means of the TEX-system facilities. The typesetting engine was pdfLATEX. The document class was `toptesi`, by Claudio Beccari, with option `tipotesi=scudo`. This class is available in every up-to-date and complete TEX-system installation.