Medical image analysis with deep learning for computer-aided diagnosis in screening

Yutong Yan

DOCTORAT BIOLOGIE BRETAGNE SANTE LOIRE

UBO université de bretagne occidentale

LATIM

Thèse de doctorat de

L'UNIVERSITÉ DE BRETAGNE OCCIDENTALE

École Doctorale N° 605
Biologie, Santé
Spécialité : *Analyse et Traitement de l'Information et des Images Médicales*

Par
Yutong YAN

Medical image analysis with deep learning for computer-aided diagnosis in screening

Analyse d'images médicales par apprentissage profond pour le diagnostic assisté par ordinateur dans un contexte de dépistage

Thèse présentée et soutenue à Brest , le 08/10/2021
Unité de recherche : LaTIM UMR 1101, Inserm

Rapporteurs avant soutenance :

Diana MATEUS	Professeur des Universités, Centrale Nantes, LS2N
Carole LARTIZIEN	Directeur de recherche CNRS, CREATIS

Composition du Jury :

Rapporteurs :	Diana MATEUS	Professeur des Universités, Centrale Nantes, LS2N
	Carole LARTIZIEN	Directeur de recherche CNRS, CREATIS
Examinateurs :	Vincent NOBLET	Ingénieur de recherche CNRS, ICube
	Etienne DECENCIÈRE	Directeur de recherche, Mines ParisTech, CMM
Dir. de thèse :	Béatrice COCHENER	Professeur des Universités – Praticien Hospitalier, CHRU de Brest, UBO
Co-dir. de thèse :	Gouenou COATRIEUX	Professeur, IMT Atlantique, LaTIM
Encadrants de thèse :	Pierre-Henri CONZE	Maitre de conférences, IMT Atlantique, LaTIM
	Gwenolé QUELLEC	Chargé de recherche Inserm, LaTIM

Invité(s) :

Mathieu LAMARD	Ingénieur de recherche UBO, LaTIM
Michel COZIC	Medecom

Abstract

Computer-aided medical image analysis is essential to support clinicians in diagnosis, prognosis and therapy-related decisions through fast, repeatable and objective measurements made by computational resources. In particular, the latest development of artificial intelligence applied to diagnosis and screening represents a promising perspective. In this thesis, we addressed the current limitations of traditional computer-aided diagnosis (CAD) systems by providing efficient and fully-automated deep learning methods towards better interaction-free and more personalized medical care. In the contexts of breast cancer and diabetic retinopathy screening, we investigated three main challenges associated with computer-assisted medical image analysis: (1) identification and segmentation of lesions from high-resolution images, (2) multi-view information fusion for improved diagnosis, and (3) longitudinal prediction of severity grade changes. Our initial contribution to the first challenge was to propose an end-to-end mass segmentation pipeline that exploits long-range multi-scale spatial context through a cascade of convolutional encoder-decoders embedding the auto-context paradigm. Then, as a second contribution, we proposed a two-stage framework combining a deep coarse-scale mass localization involving a multi-scale fusion strategy and a fine-scale mass segmentation. The second challenge was addressed by fusing information arising from two standard mammography views, namely craniocaudal (CC) and mediolateral-oblique (MLO). Two methods were proposed towards this goal. First, a novel approach based on multi-task learning was introduced, combining mass classification with dual-view mass matching between CC/MLO mammograms. Then, we applied a label-efficient deep active learning approach that exploits dual-view consistency to mitigate the labeling workload of clinicians. These methods demonstrate the effectiveness of integrating multi-view information for detection or segmentation purposes. For the last challenge, we incorporated the prior screening of fundus images to address the referable diabetic retinopathy severity change detection. All these contributions can automatically analyze different medical images in various situations and are promising to provide relevant support for the development of the next generation of CAD systems.

Résumé

L'analyse d'images médicales assistée par ordinateur est cruciale pour l'aide au diagnostic, au pronostic et au suivi thérapeutique. En particulier, le récent développement de techniques issues de l'intelligence artificielle appliquées au diagnostic et au dépistage représente une perspective prometteuse. Pour faire face aux limites des systèmes traditionnels de diagnostic assisté par ordinateur (CAD), nous avons proposé dans cette thèse un ensemble de méthodes d'apprentissage profond efficaces et automatisées, visant à améliorer la prise en charge personnalisée des patients. Dans les contextes de dépistage du cancer du sein et de la rétinopathie diabétique, nous avons principalement étudié trois défis associés à l'analyse d'images médicales assistée par ordinateur : (1) l'identification et la segmentation de lésions à partir d'images acquises à haute résolution, (2) la fusion d'informations multi-vues pour un diagnostic amélioré, et (3) la prédiction longitudinale de changements de grade de sévérité. Notre contribution au premier défi a été de développer deux méthodes dédiées à la segmentation de masses à partir de mammographies natives, à haute résolution. Dans un premier temps, nous avons proposé un pipeline de segmentation entraîné de bout en bout consistant à exploiter le contexte spatial multi-échelle grâce à une cascade d'encodeur-décodeurs convolutifs exploitant le paradigme de l'auto-contexte. Ensuite, nous avons développé une approche alternative à deux étapes, combinant la localisation de masses basée sur l'image entière et exploitant une stratégie de fusion des prédictions effectuées à multiples résolutions et la segmentation de masses sur les régions d'intérêts extraites au moyen d'un réseau profond avec connexions imbriquées et denses. Le deuxième défi a été relevé en tirant profit des informations issues des vues craniocaudale (CC) et médiolatérale-oblique (MLO) des examens mammographiques. Deux méthodes ont ainsi été proposées. Tout d'abord, une nouvelle approche basée sur l'apprentissage multi-tâches a été introduite fournissant des détections de masses précises ainsi que des correspondances entre masses issues des deux vues. Ensuite, nous avons développé une approche d'apprentissage actif exploitant la cohérence inter-vues pour diminuer la charge d'annotations des cliniciens. Ces méthodes ont démontré l'efficacité de l'intégration d'informations issues de multiples vues pour la détection ou la segmentation. Pour le dernier défi, nous avons analysé des paires d'images de fond d'œil consécutives pour la détection de changements de grade de sévérité de la rétinopathie diabétique. Ces contributions permettent d'analyser automatiquement différentes images médicales dans diverses situations et promettent de fournir un support pertinent pour le développement de systèmes de CAD nouvelle génération.

Acknowledgements

First and foremost, I am very grateful to my supervisor Dr. Pierre-Henri Conze, for all the support and the encouragement during my master internship and my three-year PhD life. He taught me how to become a researcher, to formulate and solve a research problem, and to publish and present the research work. His patience and motivation guided me in all the time of research.

I would especially like to express my special thanks of gratitude to my director Prof. Béatrice Cochener and my co-director Prof. Gouenou Coatrieux, who gave me the opportunity to do this wonderful thesis, which leverages me to learn a great amount of knowledge and experience in the fields of medical image analysis and deep learning.

My sincere gratitude also goes to Dr. Gwenolé Quellec and Dr. Mathieu Lamard. Their immense knowledge and insights helped me a lot, whenever I encounter a bottleneck during my research, their valuable and constructive opinions always inspire me to move on.

I am especially grateful to Prof. Diana Mateus and Dr. Carole Lartizien for reviewing this manuscript. I appreciate their interest in my work as all of their insightful comments and suggestions. I would like to especially thank my thesis examiners Dr. Vincent Noblet and Dr. Etienne Decencière.Thanks for devoting time and effort to evaluate my work. I would also like to thank M. Cozic from Medecom for fruitful discussions. It is my honor to have you all as my defense committee.

I truly enjoyed my time in the ITI department of IMT Atlantique and LaTIM laboratory with all the lovely current and previous PhD students and post-docs that I have met. And special mention to Yue (Shandong Rosé) and Yingjie (Fujian Jennie), our little "Pink&Black" group in Brest. Although life is sometimes tough during the epidemic, we have spent so many happy and colorful days. I wish both of you a successful PhD graduation and a bright future. "For our friendship!"

I also appreciate the sea and beaches around our campus, which leave me a lot of good memories of Brest.

Finally, I would like to thank my parents for all the support you provide and efforts you made so that I can devote myself to my research without distraction. Last but not least, my special thanks to my boyfriend (soon be my fiancé and a PhD as well) Heng Zhang, thank you for your company and encouragement for more than eight years, from China to France, from undergraduate to PhD, without you, I can never be the best version of myself. 谢谢你这些年的陪伴和鼓励，跟你一起共同进步是我最幸福和幸运的一件事.

Table of Contents

List of Figures

List of Tables

Acronyms

acc	accuracy
ACR	American College of Radiology
AI	Artificial Intelligence
AL	Active Learning
AMD	Age related macular degeneration
ANN	Artificial Neural Network
AP	Average Precision
AP-HP	Public Assistance Hospitals of Paris
AUC	Area Under the receiver operating characteristics Curve
BI-RADS	Breast Imaging Reporting and Data System
CAD	Computer Aided Diagnosis
CC	Craniocaudal
CED	Convolutional Encoder-Decoder
CFP	Color Fundus Photography
cGAN	conditional GAN
CMCNet	Combined Matching and Classification Network
CNN	Convolutional Neural Network
DICOM	Digital Imaging and Communications in Medicine
Dim	Dimension
DL	Deep Learning
DR	Diabetic Retinopathy
FC	fully-connected
FCN	Fully Convolutional Network
FFDM	Full-field digital mammography
FN	False negative
FP	False positive
FPavg	False positive per image
FROC	Free Response Operating Characteristic
GAN	Generative Adversarial Network
GAP	Global Average Pooling
Grad-CAM	Gradient-weighted Class Activation Mapping
IoU	Intersection over Union
Inserm	French National Institute of Health and Medical Research

K-NN	K-nearest neighbors
L	Left
LaTIM	Laboratory of Medical Information Processing
LSTM	Long short-term memory
MDN	Mass detection network
MLO	Mediolateral-oblique
MLP	MultiLayer Perceptron
MSE	Mean Square Error
MSF	Multi-Scale Fusion
MSN	Mass segmentation network
NPDR	Non-proliferative Diabetic Retinopathy
PDR	Proliferative Diabetic Retinopathy
R	Right
rand	random
ReLU	Rectified Linear Unit
ROI	Regions of Interest
RPN	Region Proposal Network
SENet	Squeeze-and-Excitation Network
SGD	Stochastic gradient descent
SPL	Self-paced learning
STN	Spatial transformer network
SVM	Support Vector Machine
TN	True negative
TP	True positive
TPR	True positive rate
VGG	Visual Geometry Group
WHO	World Health Organization
YOLO	You Only Look Once

CHAPTER 1

INTRODUCTION

Contents

1.1 Context and motivations

With the continuous development and progress of non-invasive imaging technologies over the last decades, medical image analysis has become an indispensable tool in medical research, clinical disease screening, diagnosis and treatment. Medical image analysis deals with the in-depth study of one or more medical images in order to make a medical decision related to diagnosis, prognosis or therapy. When clinicians perform quantitative analysis or real-time monitoring of a specific internal tissue and organ, the objective is to answer questions based on what they observe from images. Is the patient normal or abnormal? What kind of disease does this patient have? What is the prognosis for this patient's disease? Is there a risk of recurrence? Which treatment would be most appropriate for the patient? It is essential for them to know the detailed information about this tissue or organ. However, manual analysis of medical images is limited by various factors. A large amount of raw medical images obtained clinically are usually high-resolution or 3-D images, which are difficult and time-consuming to interpret. Moreover, due to the complex conditions of image acquisition, images may be bad of quality (inadequate illumination, low contrast, presence of noise...). It is therefore difficult to extract quantitative and objective information from medical images, which causes single-reading less efficient and error-prone. In addition, it is impossible to avoid subjectivity when manually analyzing medical images, while a systematic double-reading of medical images is not always available due to a lack of experts. Accordingly, quantitative measurements can contribute to a better analysis of structures and functions in normal and abnormal cases while avoiding both intra- and inter-expert variability.

Currently, medical image analysis can benefit from precise, fast, repeatable and objective measurements made by computer technology. Using these computational resources, computer-assisted medical image analysis aims at providing clinicians and medical practitioners with the information they need to analyze and evaluate abnormalities in the shortest possible time. Efficient and precise medical imaging analysis is very crucial for identifying diseases at the earliest possible stage.

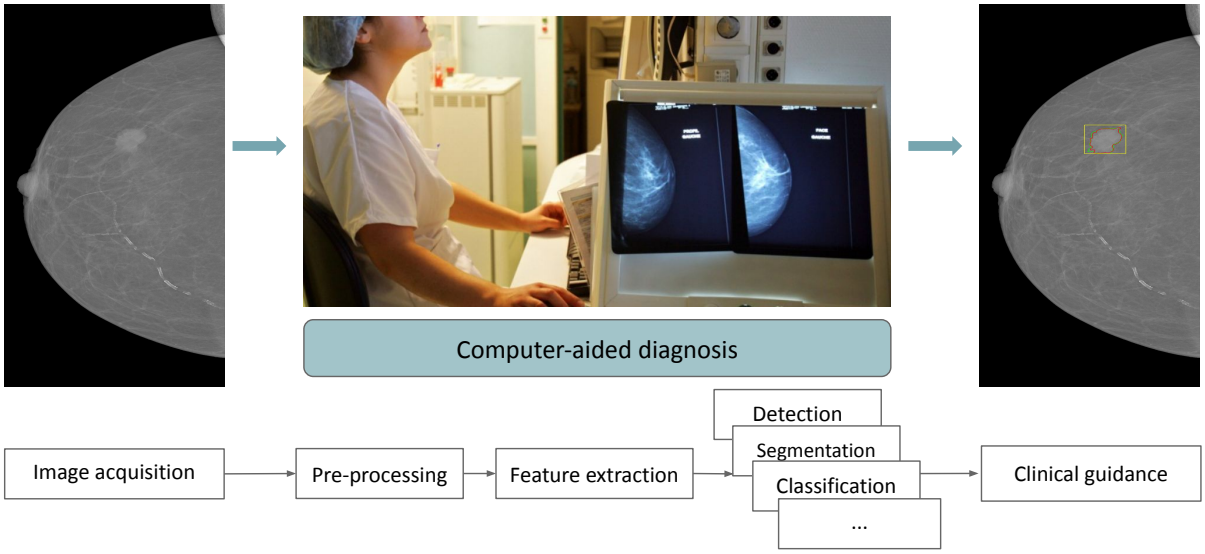


Figure 1.1 – General steps involved in a computer-aided diagnosis (CAD) system. The mammogram used for illustration is from the INbreast dataset (Moreira et al., 2012).

Under this circumstance, computer-aided diagnosis (CAD) has rapidly entered the radiology mainstream. CAD systems have been designed for supplemental lesion detection, classification and segmentation purposes as a "second opinion" to assist radiologists in their image interpretation tasks. Fig. 1.1 illustrates the general process flowchart of a typical CAD system. Generally, raw clinical images are firstly pre-processed to enhance the accuracy of the following feature extraction process. The output of a CAD system may be the likelihood of a single task or multiple tasks. The output of a CAD system may be to identify and mark suspicious areas, to outline potential lesions or to report the probability that the lesion is malignant, etc. It may also be a combination of multiple tasks. It is generally accepted that the first large-scale and systematic study and development of various CAD schemes started at the Kurt Rossmann Laboratory for Radiologic Image Research at the University of Chicago in the early 1980s (Doi, 2007). Subsequently, with the continuous progress of digital imaging and the optimization of machine performance, CAD systems have been greatly developed and applied to various image modalities, such as X-ray, ultrasound imaging, CT, MRI, etc. Abe et al. (2003) preliminarily demonstrated the usefulness of CAD. Afterwards, CAD systems were widely promoted and quickly adopted, and have further proven to outperform medical experts in certain tasks (Esteva et al., 2017; Gulshan et al., 2016). However, due to the quality requirements in clinical practice, some studies report that current CAD systems are for some applications inefficient and not automatic enough to significantly improve diagnosis guidance (Lehman et al., 2015). CAD is therefore still a field to be improved in medical image analysis.

In recent years, deep learning (DL) has achieved remarkable breakthroughs in medical image analysis through convolutional neural networks (CNN). Currently, artificial intelligence (AI) is actively being integrated into CAD in medical research. CAD become a good platform for introducing deep learning

methods. Recent CAD systems that employ deep learning demonstrate stronger robustness in clinical implementation than traditional methodologies (Asiri et al., 2019; Chan et al., 2020; Cong et al., 2020). The main advantage lies in avoiding the need for hand-crafted features by automatically learning representative features directly from data.

Despite potential benefits reported in the literature, CAD methods are generally subject to certain limitations and challenges in their applications (Santos et al., 2019). First of all, CAD systems integrated into routine requires high accuracy due to clinical requirements, i.e. high true positive rate combined with low false positive rate. Meanwhile, feasibility is also a key aspect that should not be overlooked towards efficient deployment. Another challenge is dealing with small lesions in high-resolution medical images. Pixel-wise segmentation has become a crucial task with numerous applications such as surgery planning, image-guided interventions or extraction of quantitative indices from images. Rather than using downsampled images or manually extracted regions of interests as in Al-antari et al. (2018), Byra et al. (2020), Caballo et al. (2020), Dhungel et al. (2017a), and Singh et al. (2020) or in Zhu et al. (2018), an ideal CAD system should be able to take into account the spatial context and detect lesions of any sizes to help with diagnosis, without any additional radiologist guidance. Moreover, in contrast to radiologists' practice, most CAD systems are based only on a single view, or a single image arising from time-series screenings. They analyze each image independently without considering the potential dependency information arising from multi-view or prior examinations, when available. This restricts the performance of CAD systems, and prevents them from reaching and exceeding the capabilities of human experts. Their practical applicability in realistic clinical scenarios is thus limited. To this end, eliminating those limitations and moving forwards on expanding the use of CAD tools in the daily routines of physicians are highly required.

In this context, the main objective of this work is to develop DL methods for medical image analysis and information fusion, able to detect, characterize, segment and predict the evolution of pathological structures from medical images. From a clinical perspective, automatic image processing for disease diagnosis and pathological follow-up would be beneficial to the adaptive management and therapeutic screening of each patient, towards more personalized medical care. DL applied to diagnosis and therapeutic follow-up represents a promising prospect. In this work, we mainly focus on two key clinical applications targeted in the research activities of LaTIM[1] (Laboratory of Medical Information Processing, UMR 1101) of French National Institute of Health and Medical Research (Inserm), consisting of the diagnosis of breast cancer (Sect. 2.1) and retinal pathologies such as diabetic retinopathy (DR) (Sect. 2.2).

1.2 Thesis outline

This manuscript is organized as follows:

Chapter 2 presents the two main targeted clinical applications in this work, including breast cancer

1. https://latim.univ-brest.fr

and diabetic retinopathy. For each of the two pathologies, we first introduce their clinical context and the current technical limitations. Then, the imaging modality and the employed datasets are described. Finally, we introduced in detail the pre-processing procedures which are employed to process each dataset.

In Chapter 3, we present deep learning related concepts that are exploited later in the thesis. In Sect. 3.1 we describe the key evolution and concepts of CNNs. In Sect. 3.2 we investigate several state-of-the-art CNN architectures that are often employed for the development of Computer-aided diagnosis (CAD) regarding classification, segmentation, detection and matching tasks. Sect. 3.3 gathers several deep learning strategies employed in this thesis.

Chapter 4 and Chapter 5 focus on developing highly automatic and efficient CAD systems for mammography analysis. Mammography is the main imaging modality used by radiologists to detect breast abnormalities. Chapter 4 presents two strategies, a "one-stage" approach and a "two-stage" pipeline, to tackle the problem of automated mass segmentation from high-resolution full mammograms. We further extend mammogram analysis by integrating multi-view information fusion in Chapter 5. In particular, Sect. 5.2 studies the dual-view benefits by presenting a multi-tasking network dedicated to breast mass matching, classification and segmentation; Sect. 5.3 proposes an active learning-based dual-view mammogram analysis approach where the dual-view prediction consistency is used as the selection criterion to maximize the mass detection and segmentation training performance while using the minimum amount of labeled data.

Chapter 6 integrates longitudinal information of images to help analyze the lesion evolution. Specifically, we address the referable DR severity change detection by analyzing the fusion of two consecutive longitudinal follow-up images. We first study several pre-training strategies (Sect. 6.3.2), then, we dive through an extensive exploration of image fusion schemes (Sect. 6.3.3) including early-fusion, intermediate-fusion and late-fusion to incorporate current and prior studies.

Finally, we conclude this thesis by discussing limitations and potential future works in Chapter 7.

CHAPTER 2

TARGETED CLINICAL APPLICATIONS

Contents

In this chapter, we focus on presenting the context of the two key clinical applications: the diagnosis of breast cancer (Sect. 2.1) and diabetic retinopathy (Sect. 2.2). Each section begins with the introduction of the clinical contexts of the pathology and the discussion of current limitations. Then, we describe the imaging modalities and the datasets employed to develop and evaluate our algorithms. Last but not least, we introduce in detail the pre-processing procedures applied to each dataset.

2.1 Breast cancer

2.1.1 Clinical context

Breast cancer is ranked as the leading cause of global cancer incidence among women in 2020, with an estimated 2.3 million new cases, representing about 25% of all cancers in women (Sung et al., 2021). It is also the leading cause of cancer death among women from ages 20 to 59 (Singel et al., 2018). To reduce the mortality rate, it is recommended that women over the age of 40 or women with breast tumors

who recover after treatment should have a breast screening every year. Digital X-ray mammography is recognized as a key imaging modality for radiologists to detect breast abnormalities since it allows early detection of breast cancer in women who have no symptoms, and helps women prevent breast cancer and get promptly treated. The development of massive screening has allowed earlier diagnosis and better cancer management with a significant improvement in terms of survival (Myers et al., 2015). Nonetheless, the proportion of women recalled for further examinations after screening remains significant (100 per 1000) while only 5 are truly affected. Moreover, due to the lack of a second reading by other radiologists, a substantial number of them are given heavy treatments by mistake (Myers et al., 2015).

Mammography analysis is mainly related to the detection and classification of lesions including masses, calcifications, asymmetries of the two breasts or distortion of breast tissues. Among those abnormalities, breast masses are the most important clinical symptoms of carcinomas. Characterized by medium gray to white regions within the breast area, masses exhibit a great diversity of size, shape (irregular, oval, lobulated, round), contours (circumscribed, ill-defined, spiculated, obscured) and texture, which makes them difficult to be distinguished from surrounding healthy tissues. Texture, shape and margin characteristics of masses play a key role for further breast tissue analysis (Virmani, Agarwal, et al., 2019) such as benign and malignant classification, lesion segmentation or cancer evolution prediction. The standard terminology of breast cancer severity used in mammography reports is named BI-RADS (the Breast Imaging Reporting and Data System) (D'Orsi, 1996), which was developed by the American College of Radiology (ACR) based on the level of suspicious findings.

Mammograms are usually analyzed manually by a radiologist. This task is time-consuming and prone to strong inter-expert variability (Hamidinekoo et al., 2018), resulting in up to 10%–30% undetected lesions (Moreira et al., 2012). Moreover, it is difficult and impractical for clinicians to perform double reading in most screening situations. To assist clinicians for mammogram interpretation and also to avoid time-consuming and tedious second opinions, there is an urgent need for an efficient and automatic Computer-aided diagnostic (CAD) system which is able to automatically detect and segment breast masses from the full mammogram.

2.1.2 Digital mammography

Digital mammography, also known as full-field digital mammography (FFDM), replaces X-ray films with solid-state detectors which convert X-rays into electrical signals. These electrical signals can be printed on special films similar to conventional film-screen mammograms. They can also be preserved in digital format and be displayed on a computer screen, facilitating radiologists to review and storage. Current FFDM systems are greatly improved with higher image resolution and contrast, making it easier to view dense breast tissue and small tumors (Stanford Health Care, 2017).

Standard mammography views include the bilateral craniocaudal (CC) and the mediolateral-oblique (MLO) (Fig. 2.1), which are the two views concerned in this study. The CC view is obtained by sending X-rays from a source above the breast to a detector below the breast (Fig. 2.2 (a)). The MLO view is

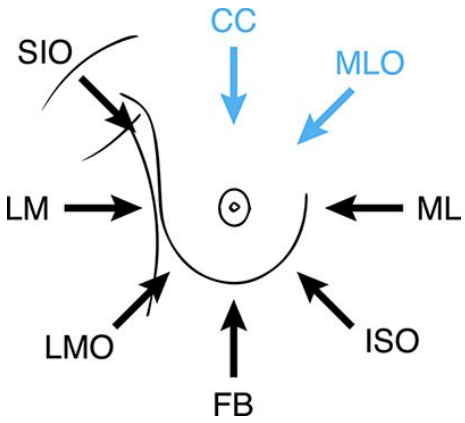
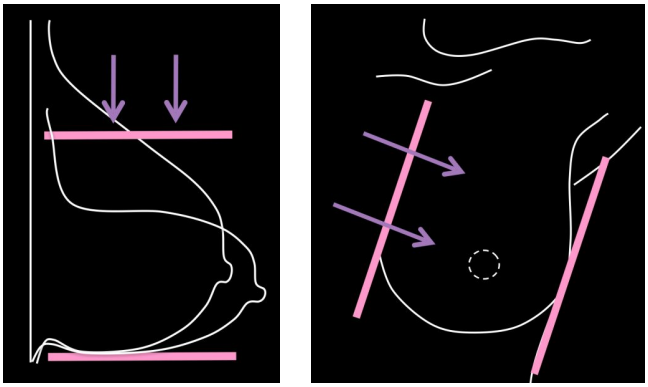


Figure 2.1 – Standard mammogram image views [2].The craniocaudal (CC) and mediolateral-oblique (MLO) views are in blue.



(a) (b)

Figure 2.2 – Acquisition of the (a) craniocaudal (CC) and (b) mediolateral-oblique (MLO) views.

produced by passing X-rays from the medial-superior to the lateral-inferior aspect of the breast (taken under 45°) (Fig. 2.2 (b)). The CC and MLO views are currently the two most common mammogram image views. Additional standard views include lateromedial (LM), mediolateral (ML), lateromedial oblique (LMO), inferomedial to superolateral oblique (ISO), from below (FB) and superior inferior oblique (SIO).

Digital mammograms are usually saved in the DICOM (Digital Imaging and Communications in Medicine) format that gathers both a set of images and the meta-data related to the acquisition process.

2.1.3 Datasets

Several publicly-available mammography databases such as the MIAS (the Mammographic Image Analysis Society Digital Mammogram Database) (Suckling J, 1994), INbreast (Moreira et al., 2012), DDSM-CBIS (Digital Database for Screening Mammography) (Lee et al., 2017) or the BancoWeb

2. Image extracted from https://www.synapse.org/#!Synapse:syn4224222/wiki/401750

LAPIMO Database (Matheus & Schiabel, 2011) are being widely used by researchers and specialists in breast cancer research. MIAS and BancoWeb LAPIMO were not used in our work, since they consist only of annotations in the form of regions of interest (ROIs), which is not sufficient for us to learn tasks that require mass delineations. Therefore, we employ the INbreast (Moreira et al., 2012) and the DDSM-CBIS (Lee et al., 2017) datasets in our mammography analysis studies.

2.1.3.1 INbreast

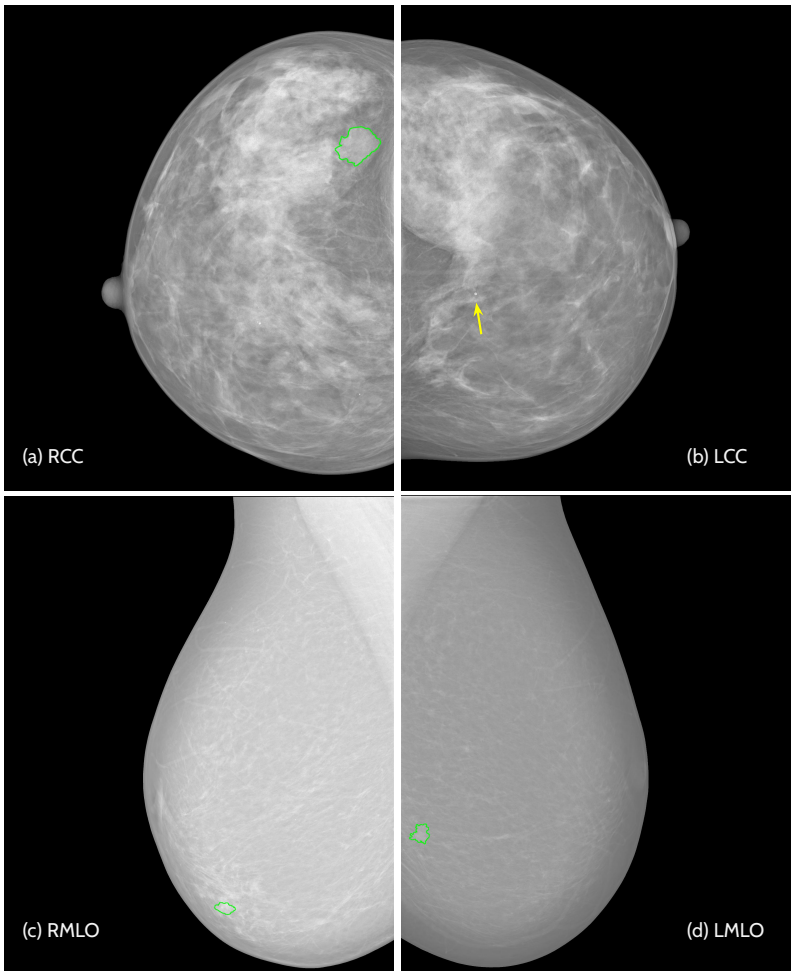
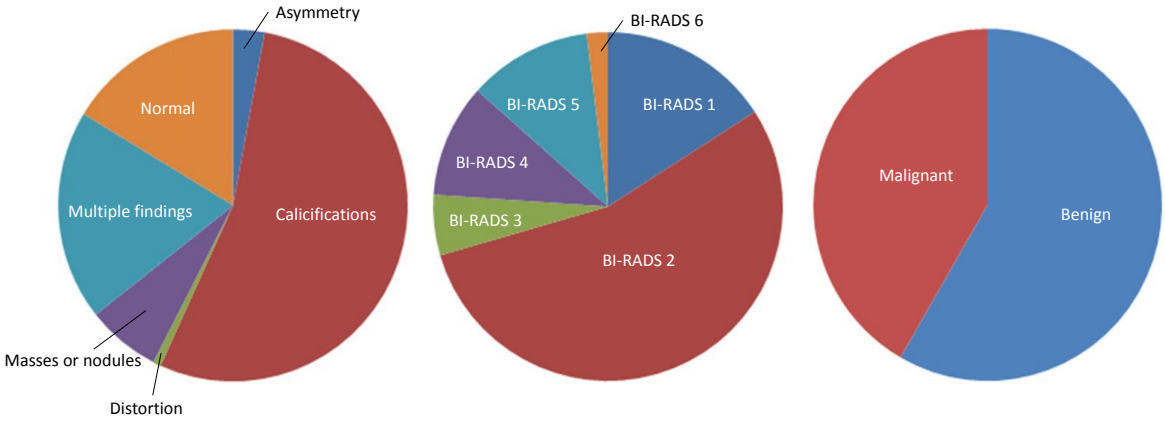


Figure 2.3 – Mammogram examples from the INbreast (Moreira et al., 2012) dataset: (a) craniocaudal (CC) view of the right breast; (b) CC view of the left breast; (c) mediolateral-oblique (MLO) view of the right breast; (d) MLO view of the left breast. Green lines indicate mass delineations, yellow arrow indicates calcifications.

INbreast [3] (Moreira et al., 2012) is a full-field digital mammography (FFDM) database which contains a total of 410 mammograms from 115 examinations, acquired and collected from the hospital S. João

3. https://doi.org/10.1016/j.acra.2011.09.014



(a) (b) (c)

Figure 2.4 – Charts describing the distribution of (a) different abnormalities (b) the BI-RADS (Breast Imaging Reporting and Data System) images and (c) benign/malignant cases in the INbreast (Moreira et al., 2012) database.

(Centro Hospitalar de S. João, CHSJ) in Porto, Portugal, acquired between April 2008 and July 2010. Two image resolutions are included: 3328×4084 or 2560×3328 pixels, depending on the breast size of the patient. INbreast images are provided in DICOM format (.dcm).

A total of 90 out of 115 cases contains four images per case from women affected by breast cancer in bilateral breasts, corresponding to the four standard views used in screening mammography: R-CC, L-CC, R-MLO, and L-MLO where L and R respectively stand for left and right. Fig. 2.3 shows examples of four standard views in INbreast dataset. The remaining 25 cases contain two images per case, from patients undergoing mastectomy. Four types of lesions (masses, calcifications, asymmetry and distortions) are included and identified. In addition, expert annotations such as breast density of ACR standard and BI-RADS scale as well as contextual information fields (patient age, screening date, type of lesion ...) are included in diagnosis reports. The overall distributions of different abnormalities, the BI-RADS rates and benign/malignant cases in this database are shown in Fig. 2.4.

Among 410 mammograms, 107 mass cases for which accurate contours made by specialists are employed. This dataset was used in Chapter 4 for training and test, and in Chapter 5 as the test set. The data distribution will be described in detail in the corresponding chapters.

2.1.3.2 DDSM-CBIS

The DDSM (Digital Database for Screening Mammography) database (Bowyer et al., 1996) was primarily developed by the Breast Cancer Research Program of the U.S. Army Medical Research and Materiel Command in 1997. The original DDSM database contains a total of 10,480 images of 2,620 scanned film mammography studies, including normal, benign and malignant cases, with two images (CC and MLO) from each laterality. Although being the largest and the most used database, DDSM

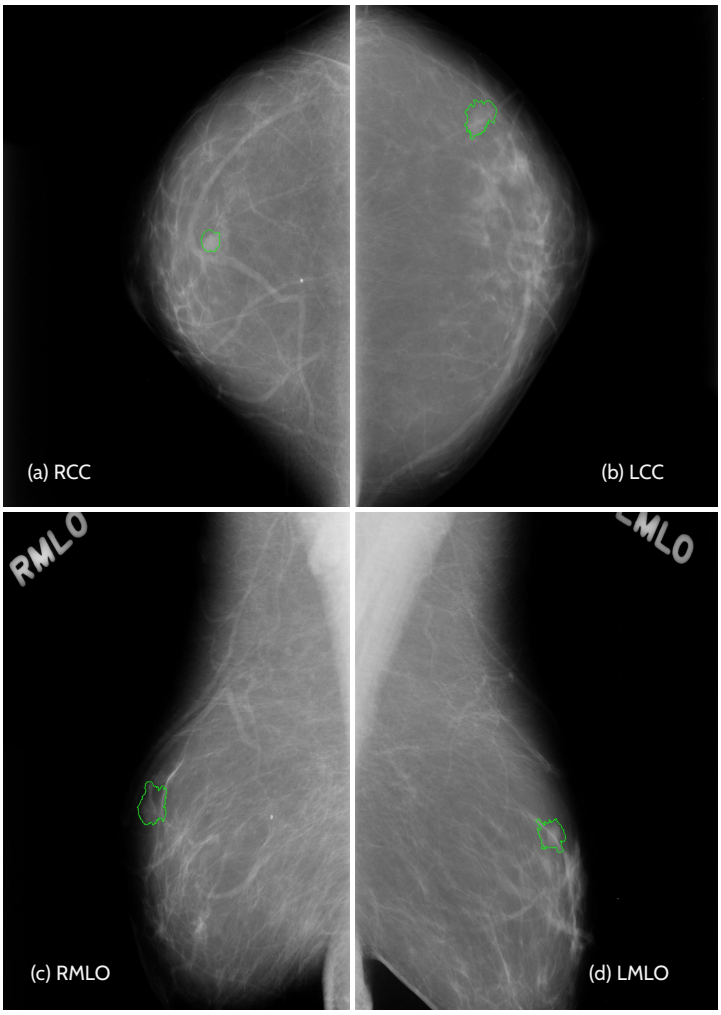


Figure 2.5 – Mammogram examples from the DDSM-CBIS dataset: (a) craniocaudal (CC) view of the right breast; (b) CC view of the left breast; (c) mediolateral-oblique (MLO) view of the right breast; (d) MLO view of the left breast. Green lines indicate mass delineations.

has several limitations such as its non-standard compression, lack of update and maintenance, lack of segmentation of lesions, its obsolescence regarding future CAD systems, etc. To address these limitations, Lee et al. (2017) developed DDSM-CBIS [4] (Curated Breast Imaging Subset of DDSM), an updated and standardized subset of the DDSM database. DDSM-CBIS includes 1514 images containing masses.

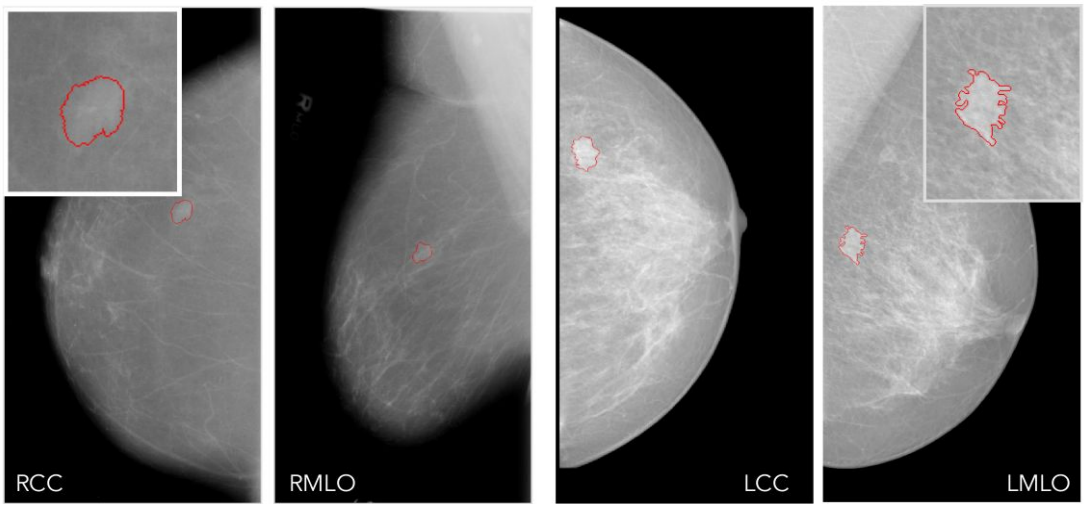
In DDSM-CBIS, mammograms are saved in DICOM format. Both MLO and CC views of the mammograms are included (Fig. 2.5). Meta-data from all mass and calcification cases are gathered and reformatted into .csv files. Patient and pathology information such as patient age, breast density in ACR and BI-RADS scale are included in data description reports. Coarse ground truth manual delineations of masses and calcifications are also saved as binary DICOM images, which can be extracted as binary masks

4. https://doi.org/10.7937/K9/TCIA.2016.7O02S9CY

of the same size as their associated mammograms. However, since the annotations for the abnormalities were provided to indicate a general position of lesions, the precision is insufficient for validating or comparing segmentation algorithms (Song et al., 2010). In this work, we used the total amount of images containing masses, i.e., 1514 mammograms. This dataset was used in Chapter 4 and Sect. 5.1 as the training set, and in Sect. 5.2 as the training set as well as the simulated unlabeled pool. More details regarding data distribution will be introduced in the corresponding chapters.

2.1.3.3 INbreast versus DDSM-CBIS

Compared to DDSM-CBIS which converts traditional film-screen mammograms to digital mammograms, INbreast is more clinically relevant since most screening procedures use full-digital mammograms. Moreover, the ground truth delineations of lesions in DDSM-CBIS are less precise than those in INbreast (Fig. 2.6). Therefore, INbreast is adopted as the primary training and evaluation dataset in this work, whereas DDSM-CBIS is mostly used as a pre-training or auxiliary training set. The comparison of statistical details of these two datasets are presented in Tab 2.1.

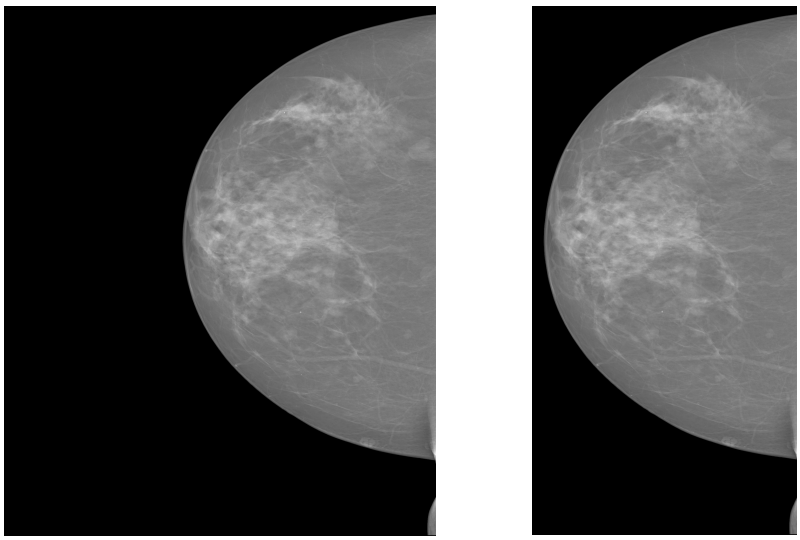


DDSM-CBIS (Lee et al., 2017) INbreast (Moreira et al., 2012)

Figure 2.6 – Comparison between mass ground truth delineations (red) from DDSM-CBIS (left) and INbreast (right) datasets.

Dataset	INbreast	DDSM-CBIS
Number of mammograms	410	10239
Number of mass cases	107	1514
Mass delineations	precise	coarse
Pixel size (μm)	70	43.5
Contrast resolution (bit)	14	12

Table 2.1 – Statistics of INbreast (Moreira et al., 2012) and DDSM-CBIS (Lee et al., 2017).



original INbreast image pre-processed 2048 × 1024 image

Figure 2.7 – Mammogram pre-processing example from original to 2048x1024 images.

2.1.3.4 Mammogram pre-processing

Due to computational limitations or in order to speed up the process, mammograms are often resized to a lower resolution as a pre-processing step (Al-antari et al., 2018; Dhungel et al., 2017b). In this study, in order to preserve the original resolution as much as possible towards more precise subsequent detection, segmentation or classification results, the only pre-processing step applied to the whole image is to crop most of the blank area (a whole non-breast region) from the original image and resize the remaining area to 2048×1024 (see Fig. 2.7).

In the subsequent proposed methods that incorporate ROI processing, mammogram patches are normalized according to the dataset mean and standard deviation (Sect. 4.2). Specifically, in Chapter 5, we implemented the proposed methods using the PyTorch [5] library. We first scaled pixels between 0 and 1, then, the single channel of grayscale image was copied to three channels and normalized with respect to the ImageNet dataset (Russakovsky et al., 2015) as done in Torch [6].

2.2 Diabetic retinopathy

2.2.1 Clinical context

Statistics from the International Diabetes Federation (Saeedi et al., 2019) show that the global prevalence of diabetes in 2019 is estimated to be 9.3% (463 millions) and will rise to 10.9% by 2045 (700 millions). As a common and high-risk complication of diabetes, diabetic retinopathy (DR) is a

5. https://pytorch.org/
6. https://github.com/keras-team/keras-applications/blob/master/keras_applications/imagenet_utils.py#L52-L55

(a) microaneurysm (b) hemorrhage (c) exudate

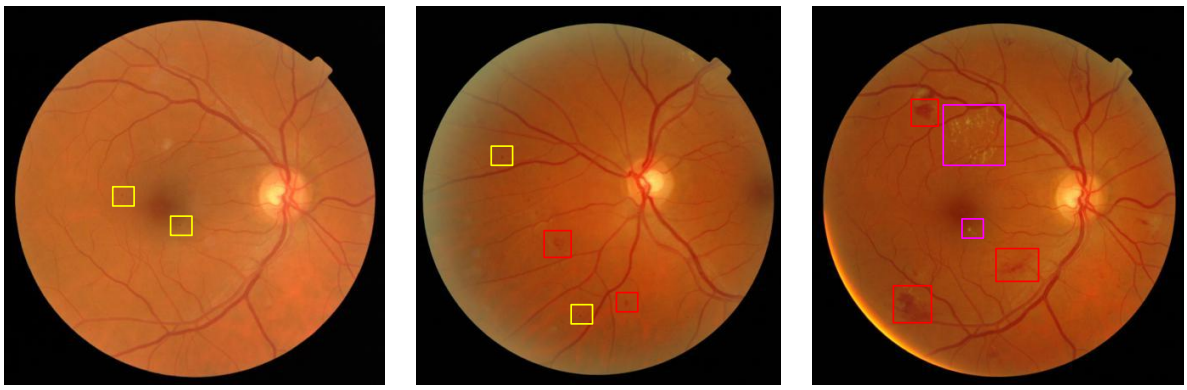
Figure 2.8 – Clinical signs of diabetic retinopathy. Images from OPHDIAT (Massin et al., 2008) dataset.

leading cause of visual impairment and blindness worldwide (Ogurtsova et al., 2017), affecting 146 million of people according to the report of the World Health Organization (WHO) in 2019 (World Health Organization, 2019). The overall prevalence of DR is up to 27.0%, comprising non-proliferative DR (NPDR) for 25.2% and proliferative DR (PDR) for 1.4% (Thomas et al., 2019). A regular annual DR screening for diabetic patients is recommended since the initial stages of the disease are subtle and hardly detectable. Moreover, considering the increasing numbers of diabetic patients and the limited numbers of specialists, developing an automatic CAD system that can analyze eye screening in an efficient and automatic fashion is a worthwhile endeavor.

In general retinal screening, color fundus photography (CFP) is commonly used for DR diagnosis by examining the presence of retinal lesions such as microaneurysms, hemorrhages, soft exudates and hard exudates. Microaneurysms (Fig. 2.8 (a)) are the first sign of DR. They consist of dilations of the venous end of retinal capillaries, appearing as small dark red dots detached from blood vessels, usually between 10 and 100 μm. Hemorrhages (Fig. 2.8 (b)) are blood leaks which appear like dark red regions within the retina. Exudates (Fig. 2.8 (c)) are the accumulation of lipid deposits in the retina, which appear as yellow bright areas in color retinal images. The proposed international clinical DR severity scale includes: no apparent retinopathy, mild NPDR, moderate NPDR, severe NPDR, and PDR (Wilkinson et al., 2003). NPDR is the early-to-middle stage of DR and is a progressive microvascular disease characterized by small vessel damages and occlusions. PDR corresponds to the period of potential visual loss due to massive hemorrhage. Fig. 2.9 illustrates the evolution from mild to severe NPDR.

The Laboratory of Medical Information Processing (LaTIM) has focused on DR since 2005. LaTIM has participated in the MESSIDOR [7] project for the development of a reference DR database under a consortium comprising AP-HP (Public Assistance Hospitals of Paris), ADCIS [8] and Mines ParisTech.

7. http://www.adcis.net/en/third-party/messidor2/
8. https://www.adcis.net/en/home/



(a) mild NPDR (b) moderate NPDR (c) severe NPDR

Figure 2.9 – Evolution from mild to severe NPDR. Yellow, red and magenta boxes respectively highlight microaneurysms, hemorrhages and exudates. Images from OPHDIAT (Massin et al., 2008) dataset.

2.2.2 Color fundus photography (CFP)

Color Fundus Retinal Photography uses specially designed fundus cameras in order to record their conditions, document the presence of retina abnormality signs related to diabetic retinopathy (DR), age related macular degeneration (AMD), macular edema or retinal detachment, and monitor their progression over time.

A fundus camera is a specialized low power microscope with an attached camera that enables illuminating and imaging the retina at the same time. Through the dilated pupil, fundus cameras photograph the inside surface of the eye, including the retina, the retinal vasculature, the optic nerve head (optic disc), the macula, and the posterior pole. Fig. 2.10 shows an example of a fundus photograph from a normal left eye with a clear visualization of these structures.

Currently, advances in fundus imaging and technology have allowed modern fundus cameras to capture high-resolution digital images with automated eye alignment and electronic illumination control. The ultra-wide field retinal imaging (Patel et al., 2020) is able to capture up to 200° of the fundus or approximately 82% of the retina in a single capture. Fundus imaging remains the primary method of retinal imaging at documenting retinal abnormalities thanks to its safety and cost-effectiveness (Abràmoff et al., 2010).

2.2.3 Dataset

2.2.3.1 OPHDIAT

The OPHDIAT dataset is a massive CFP database collected from the OphDiaT (Ophthalmology Diabetes Telemedicine) network (Massin et al., 2008), which was established in 2004 at AP-HP for diabetic retinopathy screening. This database was constructed by extracting examinations between 01/01/2004 and 01/10/2017, in an anonymized fashion. A total of 164,659 examinations were collected over the defined

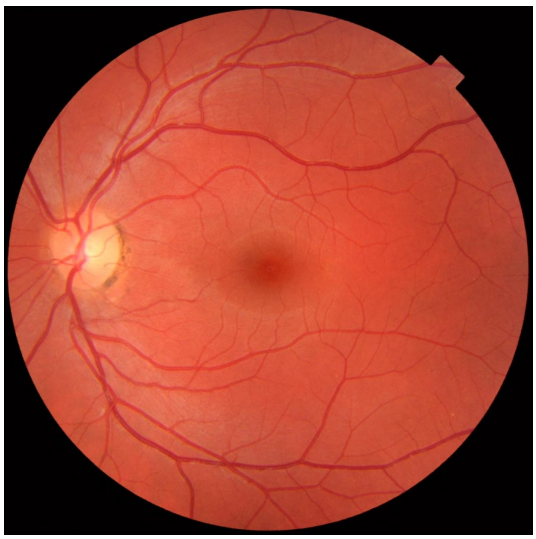
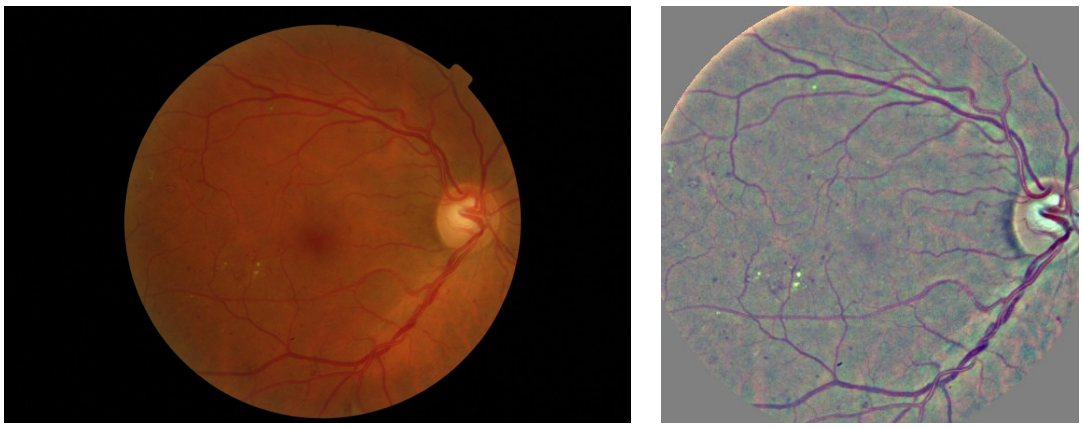


Figure 2.10 – Fundus photograph of normal left eye with no sign of disease or pathology. Image acquired at Gävle Hospital in Sweden on a healthy 25-year-old male volunteer (Haggstrom et al., 2014)

period, from 101,383 different patients. About 763,848 CFP images were interpreted, and about 673,017 CFP images were assigned with DR severity grades, including normal, mild NPDR, moderate NPDR, severe NPDR, PDR and high risk PDR. Image size varies from 1440×960 to 3504×2336 pixels. Each examination contains at least two images for each eye. Images laterality ("left eye" or "right eye") were then identified using the algorithm developed in Quellec et al. (2019). Expert annotations (DR severity grade for each eye along with the related text comments) as well as contextual information fields (patient age, screening date, diabetes history...) are included in the diagnosis reports. Double reading was adopted to ensure the annotation quality, and 6,850 exams were read at least twice. In the case of disagreement, a senior ophthalmologist has read for the third time to take a final decision. This dataset was used in Chapter 6 for training, validation and test purposes. The data distribution is also described in Chapter 6.

2.2.3.2 Retinal image pre-processing

In view of the diversity of image resolution, color, contrast, illumination, etc. presented in the OPHDIAT database, several pre-processing steps are performed as specified by Quellec et al. (2017). Firstly, images are adaptively cropped to the width w of the field of view (the eye area in CPF image). Secondly, in order to attenuate the great intensity variations of the dataset, the background of images is estimated by a large Gaussian filter in each color channel with standard deviation of 8.5 pixels, then subtracted from the image. Finally, the field of view is eroded by 5% to eliminate illumination artifacts around its edges. The resulting images are resized and cropped to $w \times w$ pixels, and are then adjusted to adapted sizes depending on the employed deep model. Fig. 2.11 shows an example of a pre-processed retinal image.



(a) original fundus photograph (b) pre-processed image

Figure 2.11 – Retinal image pre-processing example. Images from OPHDIAT (Massin et al., 2008) dataset.

2.3 Conclusion

In this chapter, we presented the two targeted clinical applications, including breast cancer and diabetic retinopathy. For each application, we introduced their clinical background and the related limitations we aim to target. Then, we described the relevant databases used in this work. We carefully compared the two mammography databases INbreast and DDSM-CBIS. In addition, we provided the pre-processing approaches followed to normalize these datasets.

CHAPTER 3

DEEP LEARNING BACKGROUND

Contents

In this chapter, we present deep learning related concepts that are exploited later in this thesis. In Sect. 3.1, we describe the key evolution and concepts of convolutional neural networks (CNN). In Sect. 3.2 we investigate several state-of-the-art CNN architectures that are often employed for the development of computer-aided diagnosis (CAD) systems for classification, segmentation, detection and/or matching purposes. Sect. 3.3 gathers several deep learning strategies employed in this thesis. Sect. 3.4 concludes this chapter.

3.1 Deep learning concepts

Whatever the application under investigation, image processing techniques usually aim at detecting and extracting representative image features such as specific patterns or textures. The extraction of representative features from data is critical in building successful machine learning models. As a result, traditional machine learning methods are employed to focus on discovering, understanding, characterizing and improving hand-crafted features that can be extracted from images, by means of Support Vector

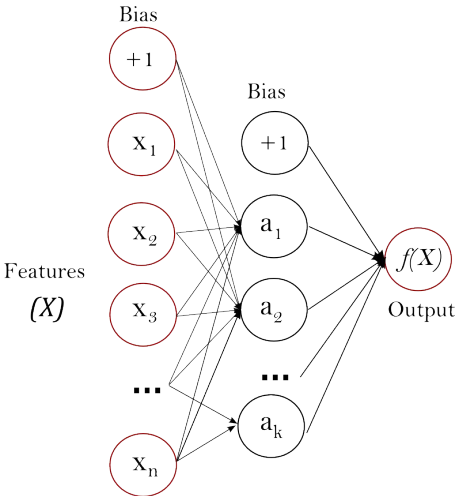


Figure 3.1 – One hidden layer multi-layer perceptron (MLP).

Machine (SVM) (Hearst et al., 1998), Scale Invariant Feature Transform (SIFT) (Lowe, 1999), or K-NN (Cover & Hart, 1967) algorithms.

However, for many tasks, it is difficult to determine what features need to be extracted. One solution is to rely on representation learning, where machine learning algorithms not only determine the mapping from representation to output, but also the representation itself. Representation learning tends to offer superior performance compared to strategies based on hand-designed representations (Bengio et al., 2017). The deep learning paradigm, which is a sub-class of machine learning, enables to implicitly learn features through neural networks, where deep layers in these neural networks act as a set of feature extractors to automatically produce generic representations that are independent of a specific classification task (LeCun et al., 2015).

Deep learning models are typically based on a set of artificial neural networks (ANNs). ANNs are computing systems inspired by biological neural networks of the human brain. The power of neural networks comes from their ability to learn an accurate representation from training data and relate it to the desired output. An example of a simple ANN is a neuron or a perceptron, which is defined as:

$$f_{neuron}(x) = g(W \cdot x) + B \tag{3.1}$$

where x is the input, g denotes a non-linear activation function, W and B respectively represent the weight and bias, and $\cdot$ is the doc product operation. The neuron attempts to find the combination of weights and bias that approximates the relation between the input x and the corresponding output. The solution space of a neuron is limited in its linear separability. In order to extend the solution space, the multi-layer perceptron (MLP) was designed to connect several perceptrons and to map a set of input values to output values. An MLP is formed by many neurons grouped into one or more non-linear layers, called hidden

layers. An example of MLP with one hidden layer can be described as:

$$f_{MLP}(x) = g_2(W_2 \cdot g_1(W_1 \cdot x + B_1) + B_2) \tag{3.2}$$

where g_i, W_i, B_i respectively denote the activation function, weight and bias of the i^{th} layer.

The predictive capability of neural networks comes from the hierarchical or multi-layered structure of MLP. Fig. 3.1 shows a one hidden layer MLP with a set of input neurons x_i with $i \in \{1, , n\}$. Each neuron in the hidden layer transforms the values from the previous layer with a weighted linear summation $\sum_{i=1}^{n} w_i x_i$, followed by a non-linear activation function $g(\cdot)$. Each neuron of an MLP is fully connected to all neurons in the following layer. However, MLP is restricted to one-dimensional training sets. In order to better represent higher dimensional patterns (e.g. edges, contours), convolution operations can be used to enhance neural networks.

3.1.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) is by far the most popular and extensively used deep learning paradigm. The emergence of CNNs is inspired by biological processing, because the connection pattern between neurons is similar to the tissue of the visual cortex of animals. CNNs are commonly trained in a supervised manner and require a large amount of labeled data. CNNs can be considered as variants of MLPs. A simple example of a two-convolutional-layer CNN can be described as:

$$f_{CNN}(x) = g_2(W_2 * g_1(W_1 * x + B_1) + B_2) \tag{3.3}$$

where $*$ represents a convolution operation, g_i, W_i and B_i respectively denote the activation function, weights and bias of the i^{th} layer.

A CNN model is normally composed of a series of layers including convolutional layers, pooling layers, activation layers and fully-connected layers. Depending on different purposes, other types of layer can be employed including batch normalization (Ioffe & Szegedy, 2015), regularization, Global Average Pooling (GAP), etc. Here, we briefly introduce the CNN layers involved in this work.

Convolutional layer. Convolutional layers extract feature maps from the input through learnable filters (kernels). Each filter outputs a weighted sum of each element of the input to its local neighbors. The local size of the input image related to the output is named the receptive field. Receptive field is an important concept in CNNs which is used to represent the range of perception by neurons of the original image within the network (Fig. 3.2). Since convolutional layers are locally connected through a sliding filter, a neuron can not perceive all the information from the original image. The larger a neuron's receptive field, the larger the range of the original image it can perceive, the more global and higher-level features it may contain. Conversely, the smaller the receptive field, the features it contains tend to be more localized and detailed.

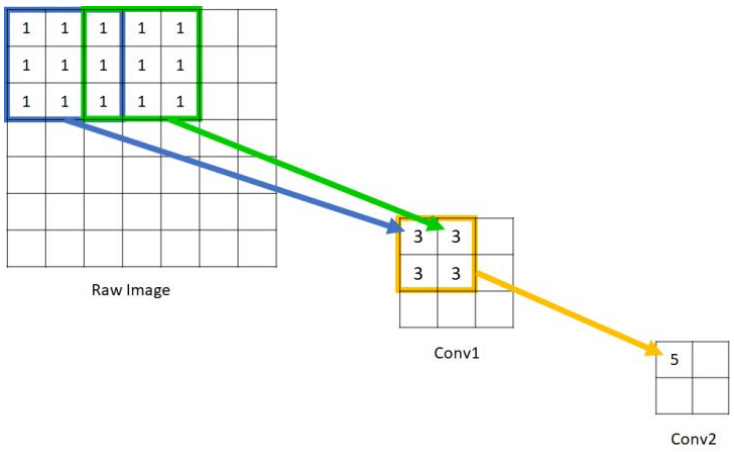


Figure 3.2 – An example of the receptive field after two convolution filters (one 3×3 filter followed by one 2×2 filter).

Pooling layer. The pooling layer downsamples the dimensions of the feature maps by keeping the most significant information. Pooling layers are usually necessary in deep learning architectures to reduce the dimensionality. By gradually reducing the spatial size of the representation, the pooling operation allows the extraction of multi-scale features, reducing the amount of parameters and computation in the network, thereby suppressing overfitting. The two most commonly used pooling operations are max pooling and average pooling. As the name implies, max pooling keeps the maximum value of the pooling range after downsampling whereas average pooling keeps the average value.

Activation layer. The convolutional layer only generates linear activation responses. In order to extend a network to represent the non-linearity, it is necessary to add non-linear activation functions such as ReLU (Eq. 3.4), sigmoid (Eq. 3.5), softmax (Eq. 3.6) or hyperbolic tangent (Eq. 3.7), etc. ReLU (Eq. 3.4) activation function preserves the properties of linear models for optimization, while modeling non-linear transformations as well. Another group of non-linear layers such as sigmoid, softmax or hyperbolic tangent function are usually used at the end of the networks for predicting a probability distribution. The sigmoid function sigmoid (Eq. 3.5) is commonly used for logistic regression, while the softmax function (Eq. 3.6) is an extension of the logistic regression model to multi-classification problems. The hyperbolic tangent function (Eq. 3.7) can be used to normalize data into the range [-1, 1].

$$ReLU(x) = max(0, x) \tag{3.4}$$

$$sigmoid(x) = \frac{1}{1 + e^{-x}} \tag{3.5}$$

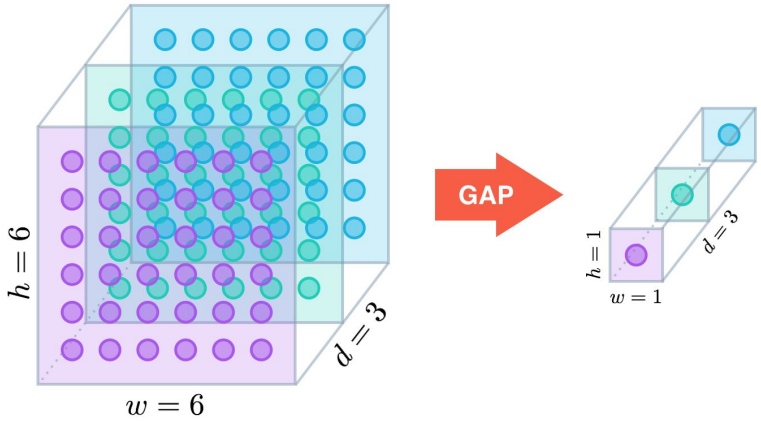


Figure 3.3 – Global Average Pooling.

$$softmax(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}} \tag{3.6}$$

$$tanh(x) = \frac{1 - e^{-2x}}{1 + e^{-2x}} \tag{3.7}$$

Fully-connected layer. Fully-connected layers act as a "classifier" in the entire convolutional neural network. The aforementioned layers map the input data to the hidden-layer feature space, while the fully-connected layer maps the learned distributed feature representation to the sample label space. Similar to MLP, each neuron in the fully-connected layer is connected to all neurons in the previous layer. The purpose of fully connected layers is to classify the output feature maps of the CNN into various classes.

Batch normalization layer. The idea of batch normalization is to normalize the inputs of each layer in order to have a mean output activation of 0 and a standard deviation of 1. Batch normalization is a technique for improving the performance and stability of CNNs.

Global Average Pooling layer. Global average pooling (GAP) was originally proposed in M. Lin et al. (2013), with the goal of minimizing overfitting by reducing the total number of parameters in the model. Overfitting often occurs when the training data is not big enough, or when the model is overtrained. During the training process, the complexity of the model increases, the error in training data decreases, while validation data errors rise. GAP layers are used to turn a three-dimensional ($h \times w \times d$) tensor into a feature vector ($1 \times 1 \times d$) (Fig. 3.3).

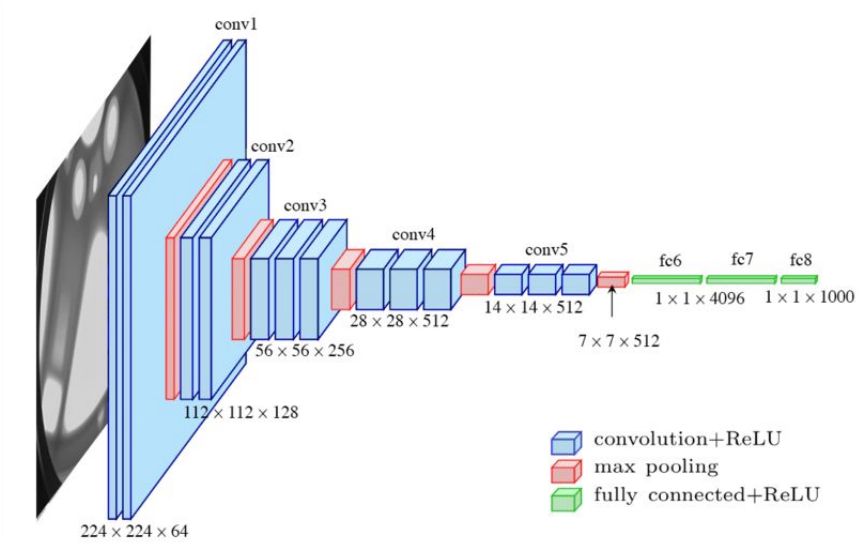


Figure 3.4 – The architecture of VGG16.

3.2 State-of-the-art CNN architectures for medical image analysis

From the initial definition to the current highly sophisticated architecture, CNN has undergone significant evolution. In 1989, CNN received attention for the first time (LeCun et al., 1989) for a 3-layer "ConvNet". Since the AlexNet (Krizhevsky et al., 2012) made a breakthrough in the image classification competition in 2012, deep learning has entered a period of rapid development. During the following years, a variety of CNN models targeting different image processing tasks have been proposed. Deep learning has also rapidly developed into a research hotspot in image analysis, with an incredible amount of research papers published every year. In this section, we will introduce some state-of-the-art CNN architectures that are employed or mentioned in this thesis. These models are also widely investigated in the field of medical image analysis, as well as in the design of computer-aided diagnosis systems.

3.2.1 Image classification

The image classification network can also be defined as a feature extraction network, whose role is to encode the input image as a high-level feature representation in the latent space. Deep networks in this category, such as VGG (Simonyan & Zisserman, 2014), ResNet (He et al., 2016a) or Inception (Szegedy et al., 2015), are commonly adopted as a component or a backbone of other customized variants or as a comparison baseline.

VGG. The VGG model (Simonyan & Zisserman, 2014) was proposed by the Visual Geometry Group of the University of Oxford at ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2014.

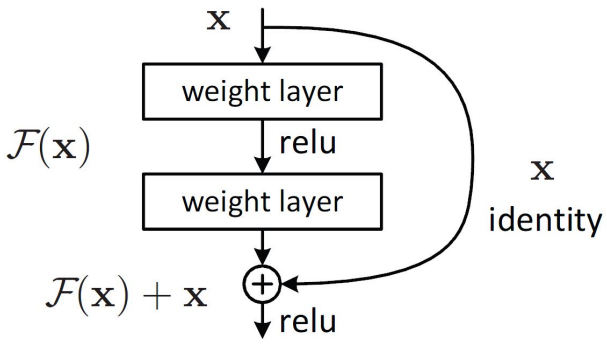
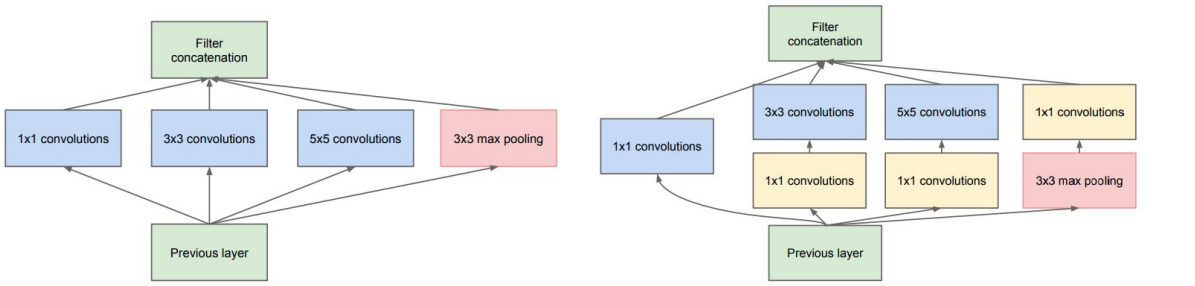


Figure 3.5 – A residual block (He et al., 2016a).

The VGG model can be regarded as a basic model of convolutional neural networks thanks to the neat and standardized architecture. The core of VGG is five groups of convolution operations (Fig. 3.4) composed of convolution and pooling layers. After each group of convolution, followed by three fully-connected layers for classification. The size of the convolution kernel used in VGG is 3×3, which is also one of the most common convolution kernel sizes in CNNs. The number of channels in the feature map rises from 64 to 512. It should be noted that, based on the number of convolutional layers in each group, there are 11, 13, 16, 19 and other types of VGG models. The most commonly used ones are VGG16 and VGG19. An improvement of VGG16 compared to the previous success AlexNet (Krizhevsky et al., 2012) is the use of consecutive 3×3 convolution kernels to replace larger convolution kernels (11×11, 7×7, 5×5). This enables to increase the network depth while keeping the same receptive field. Thus, the main advantage of VGG is that it can learn more complex patterns by using multiple non-linear layers to increase the network depth, while the computational cost is smaller.

ResNet. ResNet was proposed by He et al. (2016a) which won the challenge of ILSVRC 2015 in image classification, detection, and localization. ResNet adopts the design of the residual module (Fig. 3.5), which solves the problem of gradient vanishing/exploding when the convolutional network gets deeper. The design of neural networks generally focuses on increasing the network depth, since the depth greatly impacts the network performance. However, the deeper the neural network, the more difficult for training due to gradient vanishing or exploding during back-propagation. To solve this limitation, a skip connection (Fig. 3.5) was designed to add the input x from the previous layer to the next layer without any modification of the input. The output is thus $H(x) = F(x) + x$, accordingly, the weight layer is actually learning a residual mapping $F(x) = H(x) - x$. Even if the gradient vanished in the weight layer, we can still transfer the x back to an earlier layer. This simple step enables training to converge faster, and to successfully train much deeper CNNs. Depending on the number of repetitions of the residual module and the number of layers in each module, ResNet can be designed to lighter (ResNet18, ResNet34) or heavier (ResNet50, ResNet101, ResNet152) networks.



(a) Inception module, naïve version (b) Inception module with dimension reductions

Figure 3.6 – Inception module (Szegedy et al., 2015).

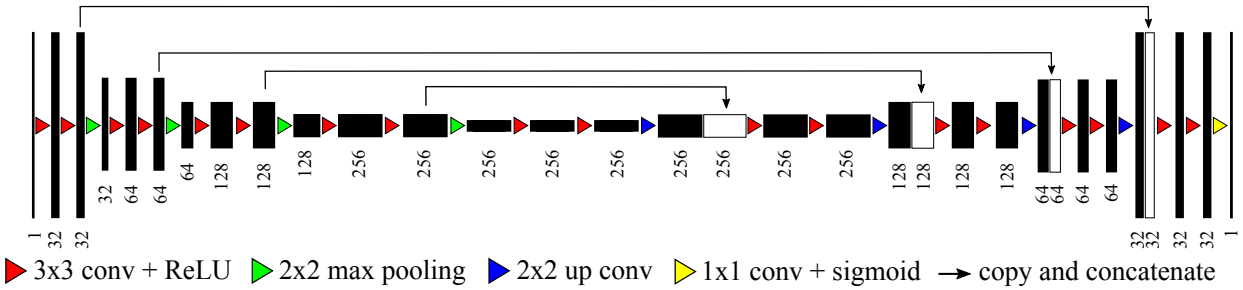


Figure 3.7 – An example of deep convolutional encoder-decoder (CED) architecture (Yan et al., 2019b).

Inception. The "Inception module" was firstly introduced by Szegedy et al. (2015) as the core component of GoogLeNet. GoogLeNet is also considered as the Inception-V1. The basic structure of the Inception module consists of four parallel components: a 1×1 convolution, a 3×3 convolution, a 5×5 convolution and a 3×3 maximum pooling (Fig. 3.6 (a)). By merging pooling layer outputs with convolution layers outputs, each layer in the network can learn "sparse" (3×3, 5×5) or "fine" (1×1) features, which not only increases the width of the network, but also increases the network's adaptability to multiple scales. Szegedy et al. (2015) also proposed an improved architecture using additional 1×1 convolutions (Fig. 3.6 (b)) to avoid the inevitable parameter growth occurring from stage to stage. Based on GoogLeNet, several improved versions of Inception were derived later. Inception-V2 (Ioffe & Szegedy, 2015) used two 3×3 convolutions instead of 5×5 convolutions to reduce the amount of parameters, and applied the batch normalization to accelerate the convergence and stabilize the training. Inception-V3 (Szegedy et al., 2016) adopted the idea of factorization to further reduce the amount of parameters, where a large two-dimensional convolution were split into two smaller convolutions (e.g. split 7×7 convolution into 1×7 convolution and 7×1 convolution). Afterwards, Inception-V4 and Inception-ResNet were designed (Szegedy et al., 2017) to further optimize the network performance.

3.2.2 Image segmentation

In the domain of medical imaging, image segmentation often refers to semantic segmentation, which consists in classifying each pixel of an image into a certain class label. It can therefore be considered as a dense pixel-wise classification problem. However, CNN also struggled in dealing with such problems. Despite pooling layers within deep models allow to increase the receptive field to aggregate the context as well as to reduce the number of parameters, semantic segmentation also requires the context information to be preserved. In 2014, fully convolutional networks (FCN) (Long et al., 2015) popularized CNN architectures for dense predictions without any fully-connected layers. They added skips-connections between layers to fuse coarse semantic context with local appearance information to improve over the coarseness of up-sampling. This end-to-end architecture allowed segmentation maps to be generated for images of arbitrary sizes. Afterwards, huge efforts have been devoted to automatic segmentation based on variants of FCN. Derived architectures comprise a regular FCN to extract multi-scale features, followed by an up-sampling part that enables to recover the input resolution using deconvolutional layers. Fully connected layers are removed, whereas a pixel-wise classification layer is applied at the end to generate the final segmentation mask. This can be considered as the earliest form of convolutional encoder-decoder (CED). Fig. 3.7 shows an example of a typical CED architecture. Almost all the subsequent semantic segmentation methods follow this paradigm. In this section, we will introduce the most well-known CED architectures: U-Net (Ronneberger et al., 2015a), SegNet (Badrinarayanan et al., 2017) and a state-of-the-art UNet++ (Z. Zhou et al., 2018) which was employed in one of our works (Chapter 4).

U-Net. U-Net (Ronneberger et al., 2015a) is a widely used CED in the medical image analysis community. Its architecture (Fig. 3.7) is made of a contraction path (the encoder) which gradually reduces the spatial dimension using pooling layers and a symmetric expansion path (the decoder) which gradually recovers the object details and spatial dimension. To improve localization accuracy, U-Net employs skip connections which concatenate features between the contracting and expanding paths. By allowing information to flow from low to high-level feature maps, a faster convergence can be achieved. U-Net consists of sequential layers including 3×3 convolutional layers followed by Rectified Linear Unit (ReLU) activations. Reducing the spatial size is handled by 2×2 max pooling layers. The first convolutional layer generates 32 channels. This number doubles after each pooling as the network deepens.

SegNet. Instead of copying the encoder features as in U-Net (Ronneberger et al., 2015a), SegNet (Badrinarayanan et al., 2017) copied the max-pooling indices received from the encoder to the corresponding decoder to perform the non-linear up-sampling of the input feature maps. This makes SegNet more memory efficient than standard FCN. The architecture of the encoder network is identical to the 13 convolutional layers in the VGG16 network (Simonyan & Zisserman, 2014). SegNet requires more training samples and longer training time than U-Net. As a consequence, U-Net is more employed in the medical imaging community.

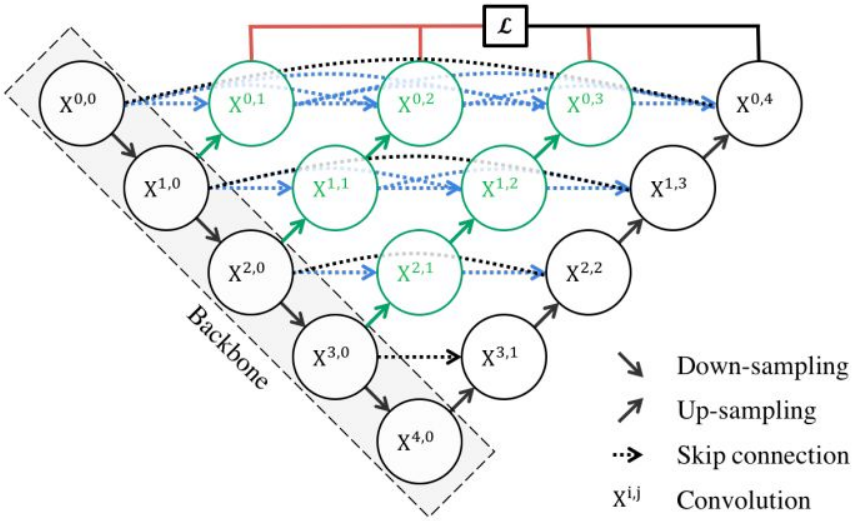


Figure 3.8 – UNet++: nested U-Net architecture for medical image segmentation. Image extracted from Z. Zhou et al. (2018).

UNet++. Rather than using simple shortcuts as in U-Net (Ronneberger et al., 2015a) or SegNet (Badrinarayanan et al., 2017), UNet++ (Z. Zhou et al., 2018) re-designed the skip pathways through a series of nested dense convolutional blocks as a convolutional pyramid to enhance feature fusion (Fig. 3.8). The number of convolution layers depends on the pyramid level. Concatenating intermediate subsequent layers bridges the semantic gap between feature maps. Then, a deep supervision is applied to prevent gradient vanishing issues in the middle part of the model during back-propagation and gather multi-depth outputs to ensure a better segmentation accuracy. Nevertheless, the nested dense skip connections as well as the deep supervision also result in a larger number of parameters, which is computationally more expensive than the standard U-Net.

3.2.3 Image detection

There are two main object detector categories: two-stage and one-stage object detectors. The dominant paradigm in modern object detection is based on two-stage approaches, including R-CNN (Region-based Convolutional Neural Network) (Girshick et al., 2014), Fast R-CNN (Girshick, 2015), Faster R-CNN (Ren et al., 2015), Mask R-CNN (He et al., 2017), R-FCN (Dai et al., 2016) etc. One-stage object detectors including YOLO (Redmon et al., 2016), SSD (W. Liu et al., 2016) and RetinaNet (T.-Y. Lin et al., 2017b) are preferred in the industrial domain for its good trade-off between accuracy and efficiency.

3.2.3.1 Two-stage detector

Namely, two-stage detectors have two stages to perform the detection: the first stage generates a sparse set of candidate region proposals that contains most of the objects, while filtering out the majority of negative locations (background). Then, the second stage classifies the region proposals into foreground classes or background and performs bounding box regression to refine the location and the size of objects.

Faster R-CNN. Faster R-CNN (Ren et al., 2015) is a representative and well-performing two-stage detector. Detectors before Faster R-CNN such as R-CNN (Girshick et al., 2014) or Fast R-CNN (Girshick, 2015) used classic algorithms such as selective search (Uijlings et al., 2013) to generate region proposals, which cost enormous computational resources. Faster R-CNN was introduced to include the generating region proposal process into the CNN architecture. Faster R-CNN introduced a Region Proposal Network (RPN) that simultaneously predicts object bounding boxes and scores at each position. RPN is a fully convolutional neural network that shares full-image convolutional features with the detection network, so it requires only a small additional computation cost for generating region proposals. This is a fundamental work for object detection because almost all the two-stage detectors after it adopt a similar structure. Faster R-CNN also introduced the concept of anchor boxes, which acted as references as well as spatial constraints during classification and regression process. This "anchor-based" mechanism was also adopted in many recent proposed object detectors (e.g. (Dai et al., 2016; T.-Y. Lin et al., 2017a; T.-Y. Lin et al., 2017b; W. Liu et al., 2016; Redmon et al., 2016; H. Zhang et al., 2020, 2021)) to generate region proposals
.

3.2.3.2 One-stage detector

Despite the excellent detection performance, training and inference of two-stage detection models are usually less efficient owing to its complex network architecture. Thus, these models are not very applicable to high-resolution medical image analysis for computational efficiency consideration. Accordingly, one-stage object detectors were introduced as an alternative.

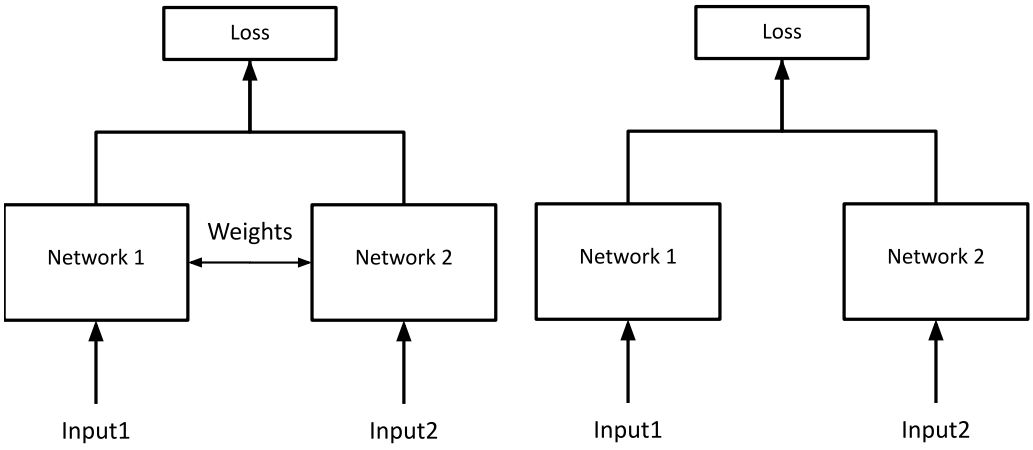
YOLO. YOLO for You-Only-Look-Once (Redmon et al., 2016), is an extremely efficient single-stage object detection model outperforming several more complex two-stage models in terms of detection speed and accuracy. Rather than performing independent processing for each potential region, YOLO posed detection as a regression problem (called "single-shot detection") and performed predictions for all objects at once with a single convolutional neural network applied to the entire image. For this reason, YOLO can see the larger context of the entire image and makes fewer background patch errors than the region-based methods. YOLO is extremely fast at test time, so that it can be used for real-time detection. Recently proposed YOLOv3 (Redmon & Farhadi, 2018) achieved higher accuracy and a much faster detection speed compared with more complex state-of-the-art detectors.

SSD. The main idea of SSD (W. Liu et al., 2016) is to combine Faster R-CNN (Ren et al., 2015) and YOLO (Redmon et al., 2016): perform the one-stage approach and implement the concept of anchor boxes. Concretely, it used the convolution layers of VGG16 (Simonyan & Zisserman, 2014) network as the feature extractor, and added convolutional feature layers to the end of the truncated base network. These layers progressively decreased in size and allow predictions of detections at multiple scales. Then, SSD associated a set of default bounding boxes (anchor boxes) with each feature map cell, for multiple feature maps at the top of the network. After that, for each anchor box, SSD computed scores for each class and the 4 offsets relative to the original default box shape. Feature maps from different levels within a network are known to have different spatial resolutions and receptive field sizes. Performing convolution on these different feature maps can therefore detect objects of different scales.

RetinaNet. RetinaNet (T.-Y. Lin et al., 2017b) is the current state-of-the-art one-stage detector, since it manages to match the accuracy of two-stage detectors while running at similar speeds with respect to other one-stage methods. As all other one-stage detectors, RetinaNet consists of a backbone network and two task-specific sub-networks (one for object classification and the other for bounding box regression). RetinaNet adopted a feature pyramid network (FPN) (T.-Y. Lin et al., 2017a) on top of the ResNet (He et al., 2016a) architecture as its backbone. To this backbone, RetinaNet attached two sub-networks, one for classifying anchor boxes and the other for regressing from anchor boxes to ground-truth object bounding boxes. RetinaNet also addressed the extreme foreground-background class imbalance problem, which is the main cause of the accuracy gap between one-stage and two-stage detectors. A novel loss function, referred to as *focal loss*, was proposed to address this class imbalance issue. Compared to common loss functions (e.g. balanced cross entropy), *focal loss* reduces the loss contribution from easy examples and extends the range in which an example receives low loss, and finally largely improves the accuracy of a one-stage detector.

3.2.4 Image matching

Image matching has also been extensively used in computer vision. Han et al. (2015) presented MatchNet, a deep convolutional approach based on Siamese networks for patch-based matching between two images I_1 and I_2. The MatchNet architecture consists of a feature network followed by a metric network. The former is a “two-tower" structure network which jointly processes two patches (one extracted from I_1, another from I_2) and maps them to a feature representation. The latter estimates the similarity between the paired features through fully-connected (FC) layers and a softmax layer to get a matching score. Zagoruyko and Komodakis (2015) made use of and explored CNN architectures to encode a general similarity function to quantify the correspondence between image patches. Amit et al. (2015) combined unsupervised segmentation and random-forest classification to detect candidate masses in CC and MLO views before estimating the correspondence between pairs of candidates in the two views. Ma et al. (2019) exploited the latent relation information between the corresponding mass regions of interest (ROI) from



(a) Siamese network (b) Pseudo-siamese network

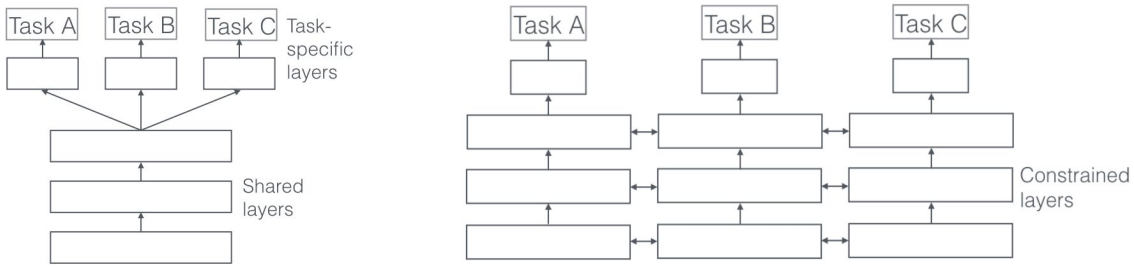
Figure 3.9 – Siamese and pseudo-siamese networks.

the two paired views using a cross-view relation region-based CNN for mass detection.

Siamese network. A Siamese model (Koch et al., 2015) includes two identical sub-networks with shared weights such that features from two different input images can be extracted simultaneously (Fig. 3.9 (a)). If the two sub-networks do not share weights, it is named pseudo-siamese network (Fig. 3.9 (b)). The purpose of the Siamese network is to measure how similar two inputs are. As is shown in Fig. 3.9, a Siamese network takes two inputs and feeds into two neural networks. These two sub-networks respectively convert both inputs into a "vector" which is mapped into a new representation space. Through the calculation of a predefined loss function, the similarity of the two inputs can be evaluated. Siamese networks prove effective in learning representation space by controlling the distance between pairs of similar and dissimilar instances (Alaverdyan et al., 2020). Siamese network is used to deal with the situation where two inputs are "similar", whereas pseudo-siamese network is suitable for dealing with the situation where there is a certain difference between the two inputs. In other words, it is necessary to determine which loss function and which structure should be used according to the specific application.

3.3 Learning strategies

Whether it is strongly supervised, weakly-supervised or unsupervised learning, deep learning algorithms are always data-driven. Sufficient and qualified data allow a deep model to learn enough required knowledge and achieve good performance. In other words, all knowledge needs to be obtained from training data. However, this also means that for each individual task, we must first prepare a certain scale of training data, and these training data need to be consistent with the distribution of the real data. However, in practice, especially for clinical applications, it is usually difficult to meet the above requirements. The



Hard parameter sharing Soft parameter sharing

Figure 3.10 – Hard parameter sharing versus soft parameter sharing for multi-task learning. Images are extracted from https://ruder.io/multi-task/.

data distribution of the training task and the target task are often inconsistent, or the training data may not be big enough. The application of deep learning may suffer from these limitations and the performance will be greatly limited. Moreover, on many occasions, it is often needed for a model to quickly adapt to new tasks. In order to cope with the above challenges, various learning strategies such as transfer learning, active learning, multi-task learning, semi-supervised learning or meta-learning have attracted widespread attention. In this section, we introduce three learning strategies that are adopted in this thesis.

3.3.1 Transfer learning

Transfer learning is a technique used to leverage a model trained on task A to another related task B. Given an insufficient dataset, training a model from scratch or using random weight initialization can not guarantee successful results. Therefore, for tasks that are difficult to obtain sufficient data, we can firstly train the model on other dataset collected for similar tasks, or use publicly available pre-trained weights, such as the ones based on ImageNet (Deng et al., 2009), a large visual object recognition database with more than 14 million images. Then, we can fine-tune the model on the small task-specific dataset. Transfer learning is widely adopted in analyzing medical images since it is difficult to collect enough images due to privacy concerns.

3.3.2 Multi-task learning

Multi-task learning refers to learning multiple related tasks at the same time, allowing these tasks to share knowledge in the learning process, and using the correlation between multiple tasks to improve the performance and generalization of the model on each task. Multi-task learning can be regarded as a kind of inductive transfer learning, which is to improve generalization ability by using the information contained in related tasks as inductive bias (Caruana, 1997). By sharing representations between related tasks, we can enable our model to generalize better on our original task.

Hard and soft parameter sharing are the two methods typically used for multi-task learning (Fig. 3.10).

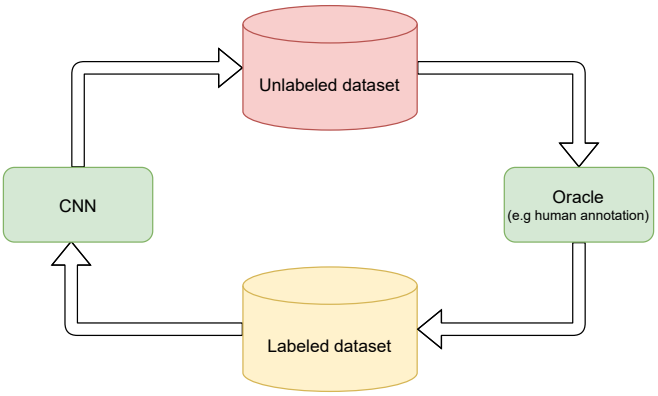


Figure 3.11 – The process of active learning.

Hard parameter sharing is generally applied by sharing the hidden layers (usually low-level) between all tasks, while keeping several private task-specific high-level layers for each task. Soft parameter sharing, on the other hand, does not explicitly set the shared modules. Each task has its own model with its own parameters, while each task can "steal" some information from other tasks to improve its own capabilities. A regularization process is then performed to make the model parameters similar.

3.3.3 Active learning

Supervised deep learning on medical imaging requires massive manual annotations, which are expertise-needed and time-consuming to perform. When only a few images can be labeled, it is possible to select the most relevant images for model training to be labeled by humans, which is the core idea of active learning (AL) (Sener & Savarese, 2017). Active learning aims at reducing human annotation efforts by adaptively selecting the most informative samples for labeling, as shown in Fig. 3.11. Acquisition functions can be designed to find diverse samples in the feature space, as extensively studied in various fields including language processing, anomaly detection or recommendation systems. AL has shown high potential in reducing the annotation cost (Budd et al., 2019).

3.4 Conclusion

This chapter mainly focuses on presenting the related deep learning concepts that are exploited later in this thesis. We first started by introducing the key concepts regarding deep learning and convolutional neural networks, then we listed the state-of-the-art CNN architectures for different image recognition tasks that we investigated in this work, including image classification, segmentation, detection and matching. Finally, different deep learning strategies such as transfer learning, multi-task learning and active learning have been briefly introduced. As explained later, we actively integrated these strategies into our work to improve learning performance.

CHAPTER 4

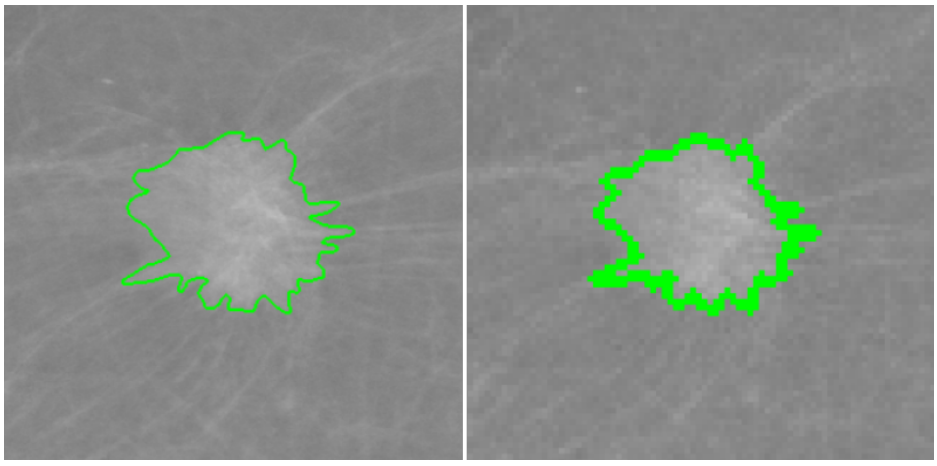
LESION SEGMENTATION FROM NATIVE HIGH-RESOLUTION IMAGES

Contents

4.1 Introduction

Manual breast mass detection and segmentation from whole mammograms remains a time-consuming and tedious process. Compared to surrounding healthy tissues, the variability combined with low signal-to-noise ratio make mass segmentation from high-resolution whole mammograms challenging for traditional

CAD systems. Most existing CAD tools focus on segmentation from low-resolution mammograms (Al-antari et al., 2018; Dhungel et al., 2017a) or from manually extracted suspicious areas (Byra et al., 2020; Caballo et al., 2020; H. Li et al., 2018; Singh et al., 2020; Zhu et al., 2018). Even if those solutions largely simplify the segmentation process, they come at the cost of overall robustness and applicability in clinical routine. First, mass patches are less representative than the entire image. Second, accurate pre-selected mass regions are not available in a real screening scenario. Fig. 4.1 displays the mass regions of interest (ROIs) extracted from the high-resolution mammogram as well as its downsampled version, respectively. We can observe that compared with the high-resolution ROI, the edge delineation of the lesion in the downsampled ROI is less precise and the texture is less clear. Therefore, achieving breast mass segmentation from native full X-ray mammograms is essential and challenging for the development of efficient automated mammogram analysis CAD systems. Using high-resolution mammograms directly as input for deep learning techniques showed an increase in performance in other recent studies (McKinney et al., 2020; Ribli et al., 2018; Tardy & Mateus, 2021).



mass ROI from 2048×1024 image mass ROI from 512×256 image

Figure 4.1 – Comparison of mass ROIs extracted from high-resolution (2048×1024) mammogram (left) and downsampled 512×256 mammogram (right). Green lines indicate ground truth delineations. Image is from the INbreast (Moreira et al., 2012) dataset.

In this chapter, two approaches are proposed to address the problem of high-resolution whole mammogram mass detection and segmentation. In Sect. 4.2, mass segmentation is achieved through an end-to-end pipeline with a multi-scale cascade of deep convolutional encoder-decoders without any pre-detection scheme. Multi-scale information is integrated using auto-context (Tu & Bai, 2010) to make long-range spatial context arising from lower scale impact training at higher resolution. This work was presented at the IEEE International Engineering in Medicine and Biology Conference (EMBC 2019, Yan et al. (2019a)). Other than continuing to explore the in-depth combination of multi-scale CED models, in Sect. 4.3, we present an alternative two-stage multi-scale framework combining a deep, coarse-scale mass detection with a new multi-scale fusion strategy and a fine-scale mass segmentation using dense and nested

skip connections, towards fully-automatic and highly precise mass segmentation from native resolution mammograms. This work was presented at the International Symposium on Biomedical Imaging (ISBI 2021, Yan et al. (2021a)) and was published in the journal of Biocybernetics and Biomedical Engineering (Yan et al. (2021d)).

4.2 Cascaded multi-scale convolutional encoder-decoder

4.2.1 Background and motivation

In the past few years, statistical models (Hizukuri et al., 2017) and machine learning techniques (Hmida et al., 2017; Y. Liu et al., 2020) have been mainly used in lesion segmentation tasks to assist clinicians for computer-assisted diagnosis of breast cancer. Some studies also focused on mammographic density characterization (Kanbayti et al., 2020; Oliver et al., 2015; Skarping et al., 2019) to target breast cancer management. In particular, Oliver et al. (2015) proposed a pixel-based support vector machine (SVM) classifier for breast density segmentation. Hizukuri et al. (2017) introduced a level set method which is based on an energy function defined with region, edge and regularizing terms to segment breast masses. Hmida et al. (2017) performed mass segmentation using a fuzzy active contour model obtained by combining fuzzy C-means and Chan-Vese models before classifying masses based on possibility theory. All these tasks are now routinely carried out in a purely data-driven fashion through convolutional neural networks (CNN). Deep CNN models have shown the most promising performance in recent breast cancer mammography-related competitions (Hamidinekoo et al., 2018). Specifically, many contributions have been proposed for breast imaging segmentation purposes, as it is an important and active research area.

The convolutional encoder-decoder (CED) paradigm has been widely adopted by most of the recent approaches designed for breast mass segmentation. Owing to large but highly similar contextual features of mammograms and unpredictable shapes and sizes of masses, most segmentation techniques focus on pre-segmented ROIs. H. Li et al. (2018) integrated the benefits of residual learning to improve the performance of U-Net to address gradient vanishing and exploding issues arising when increasing CNN depth. Dhungel et al. (2017a) combined deep belief networks, Gaussian mixture models with convolutional neural networks as potential functions into structured prediction models. Adversarial learning based on end-to-end FCN with position a-priori followed by conditional random fields (Zhu et al., 2018) have shown a better ability to handle small datasets while reducing over-fitting. Singh et al. (2020) advocated conditional GAN with mass ROI as conditioning inputs to make delineations more realistic. Caballo et al. (2020) also exploited GAN (Goodfellow et al., 2014) but as an augmentation strategy to generate synthetic breast images to simulate a larger dataset and therefore further improve deep segmentation. Byra et al. (2020) developed a selective kernel U-Net to adjust receptive fields through an attention mechanism and fused feature maps with dilated and conventional convolutions. However, these strategies focus on local segmentation of suspicious areas only, assuming that non-mass regions are previously removed either manually or using a mass candidate extractor, thus neglecting crucial contextual information.

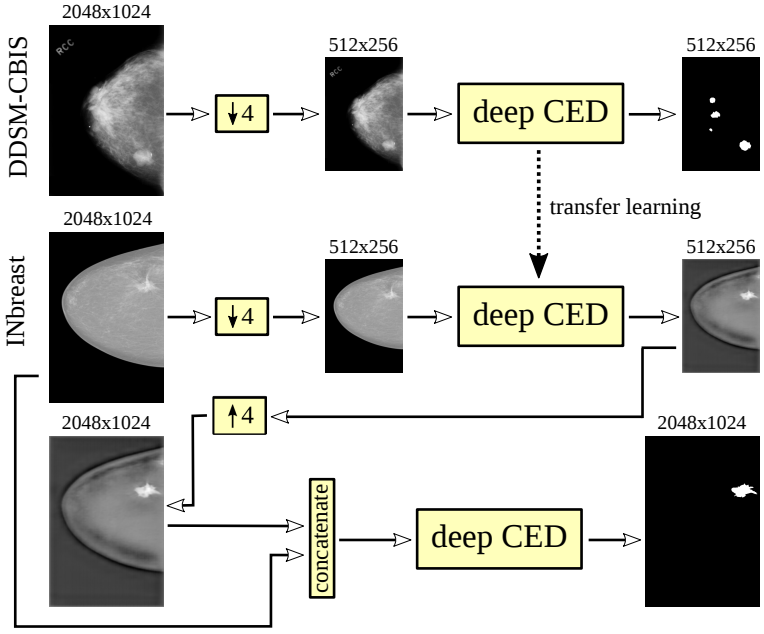


Figure 4.2 – Multi-scale cascade of deep convolutional encoder-decoders combining auto-context (Tu & Bai, 2010) and transfer learning for breast mass segmentation in high-resolution mammograms.

To address the aforementioned limitations, we have focused on an ideal CAD scenario where masses are segmented from native high-resolution mammograms to benefit from high-level information without any pre-detection scheme. To take this direction, we need to cope with a strong class imbalance issue, huge diversity of mass size, shape, texture and contour as well as the limited receptive field of CED models. In this context, we propose to exploit long-range spatial context arising from lower resolution through a multi-scale cascade of deep CEDs (Sect. 3.2.2) embedding auto-context (Tu & Bai, 2010) to fuse multi-level image information and various amounts of spatial context (Roth et al., 2018). The pipeline (Fig.4.2) is trained end-to-end to benefit from multi-scale segmentation refinements. It incorporates transfer learning from DDSM-CBIS (Lee et al., 2017) to INbreast (Moreira et al., 2012) datasets (Sect. 2.1.3) to further improve mass delineations.

4.2.2 Proposed model

4.2.2.1 Towards stacked convolutional encoder-decoders

Existing deep learning methods that used to segment breast masses remove healthy areas to process suspicious regions only. Exploiting entire high-resolution mammograms instead of small patches centered around mass candidates can better manage long-range spatial context while eliminating the risk of forgetting pathological areas. In this regard, increasing the network depth can help exploiting larger receptive fields. However, it can not be done *ad-infinitum* due to memory and computational issues.

Moreover, as the network goes deeper, we discard too many resolution details. In turn, we propose to process high-resolution mammograms using a cascade of two U-Net working at different scales (Roth et al., 2018) to exploit multi-level contextual information. In particular, we focus on how two U-Net resp. processing low (512×256) and high-resolution (2048×1024) images can be optimally combined.

The most common setup (`T1-A`) consists in training the low-resolution U-Net and to use the weights of the resulting model as initialization of the high-resolution U-Net through transfer learning (Sect. 3.3.1) and fine-tuning. Although this strategy can greatly speed up convergence, the ability of the high-resolution U-Net to extract long-range contextual features remains limited as for a single U-Net configuration (`F1`) processing high-resolution data only. The idea of stacking the two deep CEDs to further integrate multi-level information directly arises naturally. To our knowledge, no other study has recovered this concept for mammogram analysis. To deal with class imbalance, we rely on Dice instead of cross-entropy as loss function.

4.2.2.2 Multi-scale cascade with auto-context

The proposed alternative (Fig.4.2) consists in combining both U-Net with auto-context (Tu & Bai, 2010). The auto-context method combines low-level appearance features with high-level context such as the contour and implicit shape of an object, or various relationships between objects to integrate low-level and context information. In our context, the auto-context can be implemented using posterior probabilities resulting from the first U-Net as features for the second one (see Fig. 4.2) (Salehi et al., 2017). In practice, the low-resolution U-Net is trained to capture the largest amount of context based on downsampled images provided as inputs. After training, the sigmoid activation used in the last 1×1 convolution layer (Fig.3.7) to generate low-resolution binary segmentation masks is replaced by a linear function to get continuous output maps. These maps are normalized, upsampled from low to high-resolution and concatenated to high-resolution mammograms. Stacked images are given as inputs of the high-resolution U-Net which is trained from scratch (i.e. with random weights as initialization) to finally provide high-resolution binary segmentation masks.

By this way, long-range context arising from lower scale can thus have a strong impact at higher resolution. Making the first U-Net generating continuous instead of binary outputs allows propagating pixel-wise confidence information to the second U-Net. This postpones the final segmentation decision at the high-resolution level. Both models can be trained separately (`S1-A`) as in (Choi & Jin, 2016; Salehi et al., 2017) but this two-steps manner prevents refining the low-resolution model from the high-resolution one during back-propagation. Therefore, our pipeline (Fig.4.2) is trained end-to-end (`E1-A`) in order to exploit simultaneous multi-level segmentation refinement. The low-resolution U-Net is thus improved according to the analysis performed at high-resolution and *vice-versa*.

setup		resolution	architecture	CED #1	CED #2	dice	sens	spec
no DDSM transfer	F1	2048×1024	single U-Net	-	from scratch	43.66±5.7	61.15±7.3	98.00±1.1
	F4-A	512×256	single U-Net	from scratch	-	52.30±3.8	58.50±6.9	99.40±0.2
	T1-A	2048×1024	single U-Net	-	pre-train F4A	47.56±6.1	64.82±5.7	98.46±0.9
	S1-A	2048×1024	serial separately	F4A	from scratch	53.78±4.6	61.99±6.7	98.53±0.6
	E1-A	2048×1024	serial end-to-end	from scratch	from scratch	58.27±3.3	65.64±5.1	99.38±0.2
DDSM transfer	F4-B	512×256	single U-Net	pre-train DDSM	-	66.38±6.3	69.57±8.5	99.57±0.1
	T1-B	2048×1024	single U-Net	-	pre-train F4B	54.05±4.0	58.69±6.1	99.34±0.3
	S1-B	2048×1024	serial separately	F4B	from scratch	64.31±6.4	**72.31**±6.0	99.06±0.6
	E1-B	2048×1024	serial end-to-end	pre-train DDSM	from scratch	**70.04**±5.1	72.19±7.0	**99.61**±0.1

Table 4.1 – Assessment of various CED-based strategies, including our end-to-end multi-scale cascaded strategy with auto-context (E1-A/B). Cross-validation results are provided for 2048×1024 INbreast (Moreira et al., 2012) mammograms. Best results are in bold. Underlined scores highlight best results among schemes employed without DDSM-CBIS (Lee et al., 2017) transfer learning.

4.2.2.3 Integrating transfer learning

Improved model generalizability can be achieved by taking into account several datasets. Thus, our framework incorporates transfer learning from DDSM-CBIS (Lee et al., 2017) to INbreast (Moreira et al., 2012) (Sect. 2.1.3). Since training a large dataset at high-resolution is tedious, a model trained for downsampled DDSM-CBIS images is used to provide a relevant initialization to the low-resolution U-Net dedicated to downsampled INbreast images (Fig.4.2). This procedure is achieved through transfer learning and fine-tuning and concerns all previously described training schemes (T1-B, S1-B, E1-B).

4.2.3 Experiments and results

To assess the end-to-end multi-scale cascade (E1-A/B) comparatively to standard strategies (F1, T1-A/B, S1-A/B), experiments focus on mass segmentation from high-resolution 2048 × 1024 INbreast (Moreira et al., 2012) images. The performance reached using a single U-Net working at low-resolution (512 × 256) is also reported (F4-A/B) after up-sampling segmentation masks to high-resolution. Each setup is processed without (·-A) or with (·-B) transfer learning from DDSM-CBIS (Lee et al., 2017) whose training is performed at 512x256. Tab.4.1 gives an overview of all tested methodologies.

A ratio of 70% is employed to split INbreast into training and test subsets containing 74 and 33 images, respectively. Five random splits are performed to provide averaged results with cross-validation. Breast mass segmentation is quantified based on Dice ($\frac{2TP}{2TP+FP+FN}$), sensitivity ($\frac{TP}{TP+FN}$) and specificity ($\frac{TN}{TN+FP}$) scores where TP, FP, TN and FN are the number of true or false positive and negative pixels. Models are trained with 300 epochs, a batch size of 2 images (10 for F4-A/B), an *Adam* optimizer with 10^{-5} as learning rate and a fuzzy Dice loss function. Training undergoes data augmentation including random scaling, rotation, shearing and shifting. Once training is performed, predictions for high-resolution images take around 140ms only, which is suitable for clinical practice.

We present in Tab.4.1 a comparative assessment of all previously described methods. Comparisons

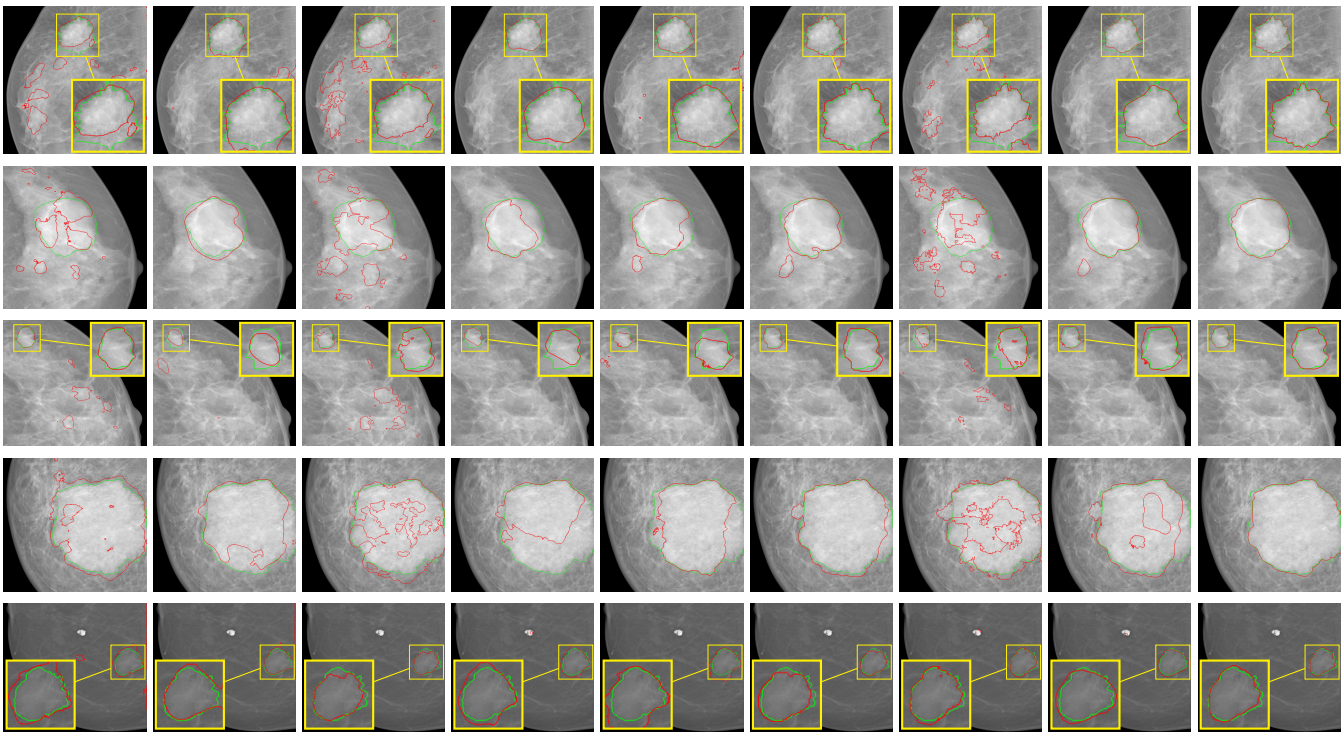


Figure 4.3 – Automatic mass segmentation for high-resolution INbreast (Moreira et al., 2012) images through CED-based strategies including our end-to-end multi-scale cascade with auto-context (E1-A/B). Ground truth and estimated delineations are respectively in green and red.

based on Dice measures between F1 and F4-A indicates that the same single U-Net provides better results with low than high-resolution images since the receptive field covers larger context in 512×256, compensating for the lack of details. Using the weights of the low-resolution model as initialization of the high-resolution one (T1-A) achieves gains for all metrics compared to F1 but does not outperform F4-A in Dice and specificity. Compared to T1-A, separately training stacked U-Net with auto-context (S1-A) appears better in Dice (from 47.56 to 53.78%) which demonstrates that long-range context arising from lower scale greatly impacts high-resolution data management. End-to-end training (E1-A) results in further improved performance. Despite similar spec, E1-A gives the best Dice (58.27%) and sens (65.64%) among all methods without DDSM-CBIS transfer thanks to the simultaneous multi-scale segmentation refinement. Variabilities between simulations explain the observed large standard deviations.

A significant gap is crossed when transfer learning from DDSM-CBIS images is considered. Gains concern all methods and vary from 6.5 (T1) to 14.1% (F4) in Dice. Except between F4 and S1, the same conclusions as without transfer learning arise: E1-B > F4-B > S1-B > T1-B > F1. The proposed multi-scale cascade with auto-context and transfer learning achieves the best scores in Dice and specificity with 70.04% and 99.61% against 66.38% (resp. 64.31%) and 99.57% (99.06%) for F4-B (S1-B). This reveals that our method efficiently takes advantage of both low-level broad context and high-level fine details.

Sensitivity for `E1-B` is slightly below `S1-B` but clearly outperforms `F4-B`.

Evaluation is supplemented by qualitative results given in Fig.4.3 for all methods. Provided examples report inconsistent shapes combined with false positive areas located far-away from ground truth mass locations for `F1` and `T1-A/B`. Despite better shape integrity, `F4-A/B`, `F1-A/B` and `E1-A` setups are prone to under-segmentation, especially without pre-training. Conversely, we notice a much more accurate boundary adherence and subtle contour delineation using `E1-B` for both small and large masses. This confirms that our contributions provide good model generalizability despite the class imbalance issue and large as well as mass appearance variability.

4.2.4 From one-stage to two-stage

In this section, mass segmentation was achieved through a multi-scale cascade of deep convolutional encoder-decoders without any pre-detection scheme. Multi-scale information was integrated using auto-context to make long-range spatial context arising from lower scale impact training at higher resolution. The pipeline was trained end-to-end to benefit from simultaneous segmentation refinement performed at each level. We incorporated transfer learning and fine-tuning from DDSM-CBIS to INbreast datasets to further improve mass delineations. The comprehensive evaluation provided for high-resolution INbreast images highlights promising model generalizability against standard encoder-decoder strategies. Further attempts of this strategy should involve cascading more CEDs to further refine the multi-scale information, which can be beneficial to the feature refinement.

Other than continuing to explore the in-depth combination of multi-scale CED models, we also investigated an alternative two-stage strategy. It is worth noting that, despite being complementary, localizing mass areas from mammograms and extracting precise boundaries for each mass are naturally two tasks with contradictory focuses: context-level semantic information for the former, resolution-level details for the latter. Addressing both challenges into one single network may lead to a sub-optimal trade-off and thus hinder precise full mammogram delineations. In this context, we came up with the idea of a two-stage method which is desired to imitate the realistic procedure in clinical scenarios, and we tried to automatize the candidate selection process using multi-scale fusion. Accordingly, in the following section, we present a two-stage solution for full mammogram segmentation, providing another option for accurate and automatic mass localization and segmentation CAD systems.

4.3 Two-stage breast mass detection and segmentation

It has been proven in Sect. 4.2 that the end-to-end training of a multi-scale cascaded CEDs model can achieve mass segmentation without any pre-detection scheme. However, the strong mass size variation (Sect. 2.1.3) is still a limitation factor to the performance, since the reception field of deep models tends to be limited to segment very large masses and very small masses at the same time. Alternatively from one-stage segmentation approaches (Singh et al., 2020; Yan et al., 2019a), we proposed a two-stage

pipeline where masses are firstly localized before being precisely delineated. The proposed framework (Fig.4.4) consists of two modules: image-based mass localization (Sect.4.3.2) followed by region-based mass segmentation (Sect.4.3.4). The former is based on a deep detection model extended based on a novel multi-scale fusion procedure (Sect.4.3.3) to alleviate wrong proposals and further improve detection accuracy. This stage performs coarse mass detection on entire mammograms and provides suspicious regions to the second stage. The latter conducts refined mass segmentation on extracted areas relying on a deep convolutional encoder-decoder architecture with nested and dense skip connections. An image reconstruction step is finally followed to visualize both mass location and segmentation results in high-resolution full mammograms.

4.3.1 Related works

Regarding breast mass detection, although many recently proposed object detection models (Dai et al., 2016; Girshick et al., 2015; Redmon & Farhadi, 2018; Ren et al., 2015) have achieved great success on common object detection tasks, automatic mass detection still remains a challenge due to the low signal-to-noise ratio and the unpredictable appearance of masses in X-ray mammograms. Agarwal et al. (2019) analyzed the performance of popular deep CNN architectures in terms of mass/non-mass classification. Alternatively, Jung et al. (2018) proposed a mass detector based on RetinaNet (T.-Y. Lin et al., 2017b) exploiting a feature pyramid network optimized through a focal loss. Yap et al. (2020) automated breast lesion detection using Faster-RCNN (Ren et al., 2015) with Inception-ResNet-v2 (Szegedy et al., 2017). However, these learning-based detectors may fail in identifying masses of any size, position or shape from the whole image. Existing detectors might therefore not produce sufficiently good proposals for further breast mass segmentation purposes.

Many studies focus on building multi-stage networks or integrating a series of steps together. Dhungel et al. (2017a) proposed a cascade of deep belief networks and Gaussian mixture models to provide mass candidates, followed by two cascades of CNN and random forest to refine detection results. Once suspicious areas are identified, they employ deep structured learning to perform mass segmentation. Al-antari et al. (2018) proposed an integrated mass detection, segmentation and classification pipeline from downsampled mammograms. Although their system could assist radiologists in multi-stage diagnosis, they still manually eliminated false localized candidate masses before the segmentation stage, which is impractical as an automatic CAD system. Apart from that, they exploited low-resolution mammograms. Image details are therefore lost during this process. In comparison, our approach aims at avoiding complex processing pipelines and human interventions, towards accurate and precise breast mass segmentation.

4.3.2 Image-level mass detection

Among existing deep detectors, YOLOv3 (Redmon & Farhadi, 2018) is adopted in this work for mass localization from full mammograms thanks to its good trade-off between accuracy and efficiency.

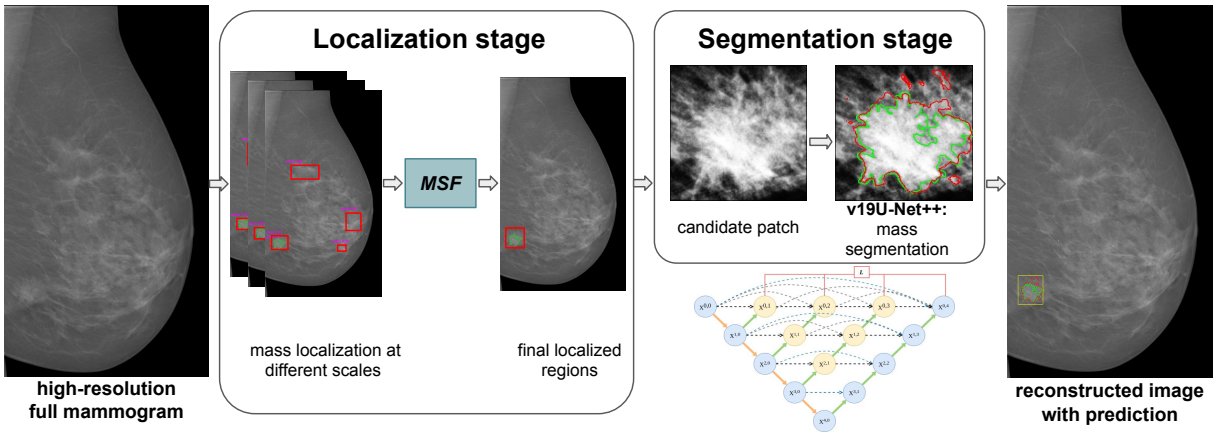


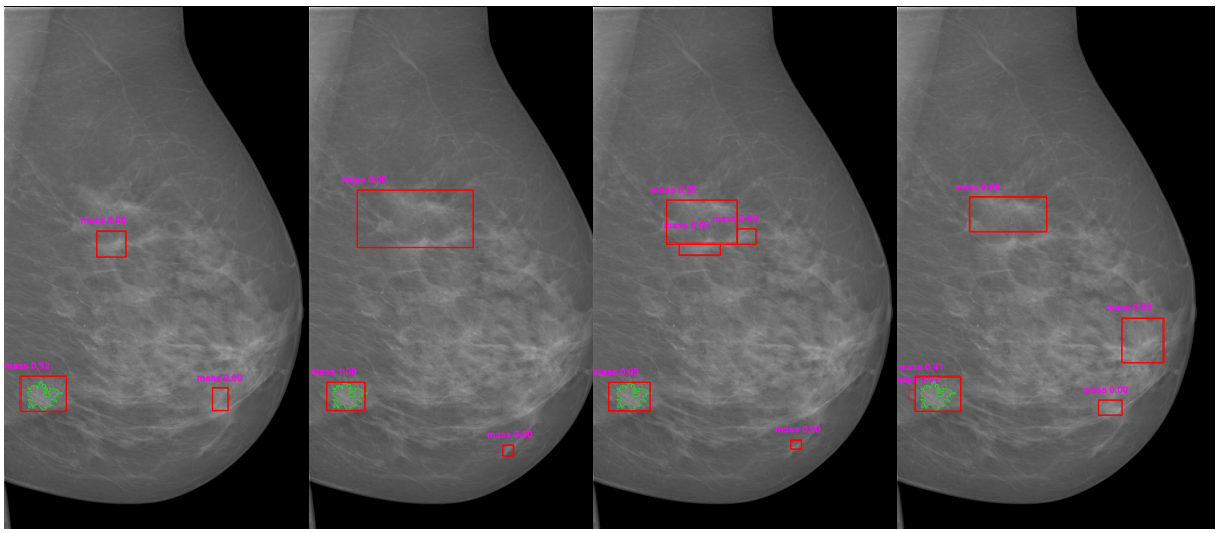
Figure 4.4 – Two-stage multi-scale pipeline for mass localization and segmentation from high-resolution X-ray mammograms. Red (green) lines indicate estimated (ground truth) delineations. MSF deals with the proposed multi-scale fusion strategy for automatic mass selection.

However, other detectors such as SSD (W. Liu et al., 2016), Faster R-CNN (Ren et al., 2015) or RetinaNet (T.-Y. Lin et al., 2017b) can also be applied as alternative detection schemes (Sect. 3.2.3).

The employed YOLO (You Only Look Once) implementation exploits the Darknet-53 backbone architecture consisting of 53 successive 3×3 and 1×1 convolutional layers as well as some shortcut connections. Feature maps from different scales are used to deal with huge mass size and aspect ratio variance, i.e., larger feature maps are assigned to detect smaller masses and vice versa. Following (Redmon & Farhadi, 2018), YOLOv3 uses anchor boxes to predict through regression the coordinates of bounding boxes. Different from Faster R-CNN (Ren et al., 2015) which uses manually selected boxes, k-means clustering is used to recompute the 9 anchor settings to adapt YOLOv3 to the target mammography datasets. For training, we use pre-trained weights arising from ImageNet (Deng et al., 2009) pre-training.

4.3.3 Extension using multi-scale fusion

Although recently proposed detection models (Dai et al., 2016; Girshick et al., 2015; Redmon & Farhadi, 2018; Ren et al., 2015) have achieved excellent results on public common object detection datasets such as Pascal VOC (Everingham et al., 2015) or Microsoft-COCO (T.-Y. Lin et al., 2014), they are not optimal to be applied directly to mammograms for two reasons. First, they are still struggling with object size variance. Typically, most object detectors have worse performance for small objects than for medium or large structures. Especially in our context, this problem becomes more serious as the size and aspect ratio of masses vary greatly. Second, mass detection is more difficult than common object detection since masses are visually less obvious and less contrasted with respect to surrounding healthy tissues, combined with a great diversity of shape and texture. Therefore, single-scale prediction might not provide sufficiently good proposals, leading to the failure of the next stage dedicated to mass segmentation.



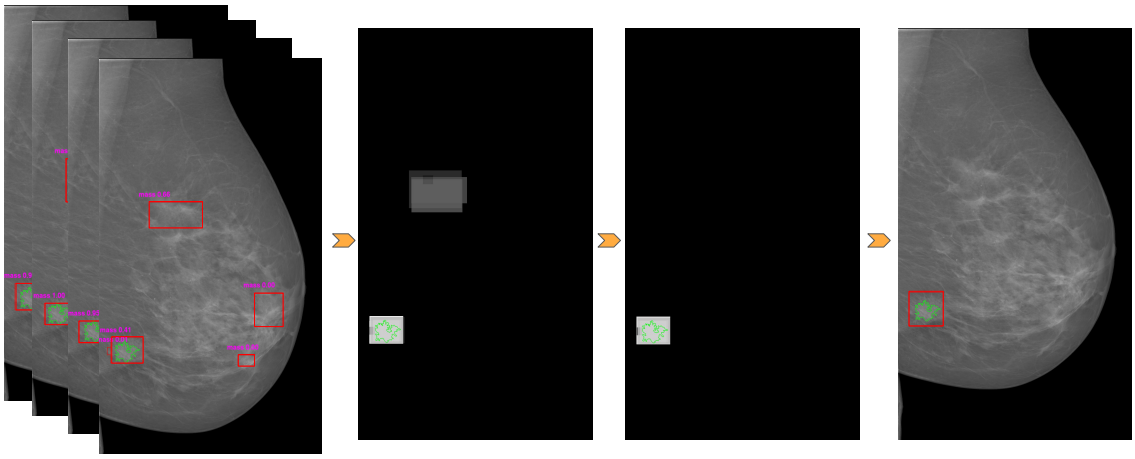
160 × 320 256 × 512 320 × 640 480 × 960

Figure 4.5 – YOLOv3 predictions performed at multiple scales for one given mammogram. Red boxes correspond to mass ROI candidates with associated probabilities in magenta. Green contours arise from ground truth annotations.

In addition, previous works including Al-antari et al. (2018) that also used YOLO as mass detection model tend to manually select candidate masses to avoid false-positive proposals before the segmentation stage. We argue, however, that such approaches assume that they already have box-level expert annotations during validation and test phases, which is less practical and not obvious. As a matter of fact, an automatic and fully-integrated CAD system should not require any expert annotations for clinical purposes.

To address the problem of unsuccessful single-scale detection and avoid manual selection, we propose a multi-scale fusion (MSF) strategy. Note that one of the important designs in YOLOv3 is the multi-scale training, for which input images are dynamically resized every 10 batches instead of fixing the input image resolution. Image resolutions are randomly chosen from multiples of 32 since the model is downsampled by a factor of 32. As a consequence, our MSF extension tends to fully exploit the multi-scale features extracted by YOLOv3 during training to further refine the generated candidates. Moreover, it allows us to be robust to the input size so that images with different resolutions can be processed without multiple training. In the same spirit as for training, we propose in the prediction stage to exploit results from different resolutions to make the network being more sensitive to masses with very small or large spatial extents. As shown in Fig.4.5, we are able to perform different predictions at different scales using the same network. Thus, for a given mammogram, we propose to fuse predictions arising from multiple scales.

The proposed MSF scheme consists of three main steps (Fig.4.6). For a given mammogram, detections are first carried out at different image scales (Fig.4.6a). Since larger resolution will exceed the memory limits while smaller resolution will reduce the accuracy, we use the following 5 image ratios: (160×320), (256×512), (320×640), (416×832), (480×960). Second, we collect all B coordinates of candidate



(a) (b) (c) (d)

Figure 4.6 – Proposed multi-scale fusion (MSF) applied to YOLOv3 predictions. The MSF strategy focuses on redundant information in multiple predictions. Red boxes correspond to mass ROI candidates. Green contours arise from ground truth annotations.

bounding boxes and the corresponding confidence score sets C provided in the previous step by YOLOv3. For each box B_i with $i \in N$, we create a confidence mask M_i where the value of the box area is the corresponding confidence score c_i. Let $(X, Y)_i$ be the set of coordinates from the bounding box B_i. For each pixel (x, y) inside B_i, we assign $M_i(x, y) = c_i$ with $c_i \in C$. Second, a single confidence mask M_s (Fig.4.5b) is created by fusing confidence masks $\{M_1, M_2, \ldots, M_N\}$ obtained at each prediction scale. M_s is computed and normalized as follows:

$$M_s = \frac{\sum_{i=1}^{N} M_i}{N \times \max(c_1, c_2, \ldots, c_N)} \tag{4.1}$$

Third, we consider an empirically selected threshold λ to implement majority voting (Fig.4.6c) to the fusion mask M_s by keeping areas where $M_s \geq \lambda$. All connected regions of M_s are assigned the same integer label. We measure the properties of labeled M_s and find bounding box(es) that describe the fusion mask most properly (Fig.4.6d), i.e. we find box tuples $(\min_x, \min_y, \max_x, \max_y)$ such that pixels of the same label belong to the same box in the half-open intervals $[\min_x; \max_x)$ and $[\min_y; \max_y)$.

Through the proposed MSF, we focus on redundant information that appears in multiple scales. From a statistical point of view, MSF allows to identify the most frequently detected regions in multiple predicted maps in order to limit false-positive predictions. Conversely, areas detected in few prediction maps or areas with low confidence scores are unlikely to be selected. Moreover, we analyze the effect of the empirical parameter λ in order to keep a high level of sensitivity while improving specificity. Accordingly, we are able to remove most of the uncertainty and find the most reliable predictions. Final detections are resized to 256×256 patches and fed into our second stage.

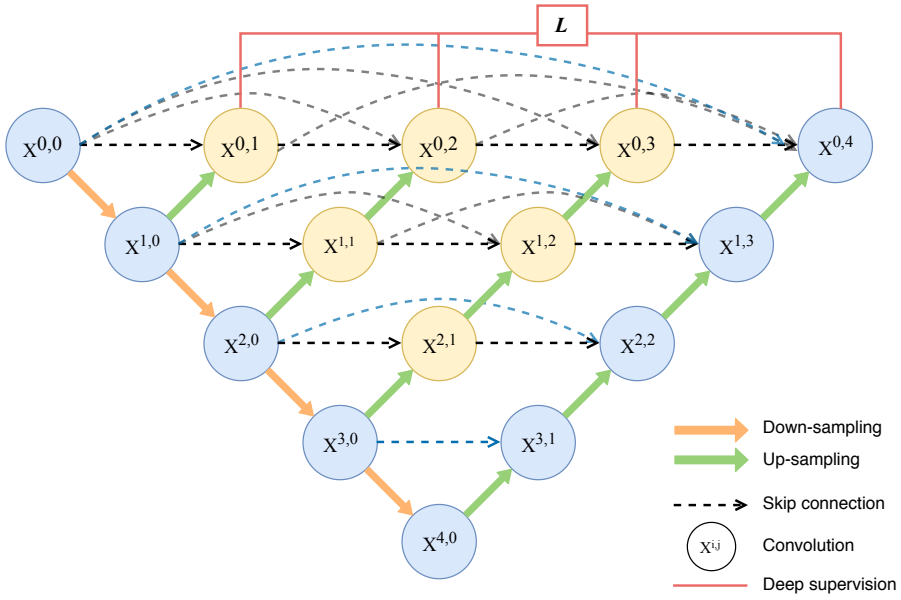


Figure 4.7 – Deep convolutional encoder-decoder architecture with nested and dense skip connections, following UNet++ (Z. Zhou et al., 2018).

4.3.4 Patch-level mass segmentation

After the image-based mass detection stage, we propose a region-based mass segmentation stage that performs refined mass delineation from candidate patches using a deep convolutional encoder-decoder. Among recent advances of segmentation approaches, we implemented a powerful deep architecture with nested and skip connections, following U-Net++ (Z. Zhou et al., 2018) (Sect. 3.2.2).

The architecture (Fig. 4.7) is derived from standard U-Net : we employ in practice the vgg19 network as backbone for the encoder, which consists of 16 convolutional layers with repeated 3×3 convolutions followed by ReLU activation function and 2×2 max-pooling (3 fully-connected layers are not included). The decoder is symmetrically designed. The proposed mass segmentation method is referred to as v19U-Net++. Since reaching a generic from scratch model without overfitting is difficult, we pre-train the encoder branch using ImageNet (Deng et al., 2009) following (Conze et al., 2020) to reduce the data scarcity issue while allowing faster convergence. We exhaustively implemented four segmentation models for comparison: U-Net (Ronneberger et al., 2015a), cGAN (Singh et al., 2020), cascaded U-Net (Sect. 4.2, Yan et al. (2019a)) as well as v19U-Net++ as suggested. Once we get segmentation results, we can reconstruct high-resolution full mammograms with mass identification and delineation for visualization purposes.

4.3.5 Experiments and Results

In what follows, Sect. 4.3.5.1 presents the data used in this work. we report experimental settings (Sect.4.3.5.2) and results for image-based localization (Sect.4.3.5.4) and segmentation (Sect.4.3.5.3) of breast masses. In particular, evaluations of final segmentation results are carried out both quantitatively and qualitatively. All experiments are implemented using Keras backend with a single Nvidia GeForce GTX 1080Ti GPU.

4.3.5.1 Data

We focus on mass detection and segmentation from 2048×1024 full mammograms arising from INbreast (Moreira et al., 2012) and DDSM-CBIS (Lee et al., 2017) presented in Chapter 2. In this work, 107 (1514) INbreast (DDSM-CBIS) images containing masses are employed. The DDSM-CBIS database is only employed in the detection stage. The INbreast dataset is too small to be representative if being divided into three subsets (train, validation and test sets). Therefore, we employ a ratio of 70% to split INbreast into train and test subsets containing respectively 74 and 33 images. In order to eliminate the bias error, we use 5 random splits (denoted as T1, T2, ..., T5) to provide averaged results with cross-validation.

4.3.5.2 Experimental setup

Image-based mass detection. Experiments of this stage focus on mass detection from 2048×1024 mammograms. Typically, training a detection model on an insufficient dataset such as INbreast does not guarantee precise results. Therefore, a transfer learning technique is used to leverage a deep learning model on one task to another related task (Sect. 3.3.1). In this work, we use convolutional weights pre-trained on ImageNet (Deng et al., 2009), then we conduct transfer learning from DDSM-CBIS to INbreast. All 1514 DDSM-CBIS images containing masses are employed to pre-train the YOLOv3 model for 60k iterations before fine-tuning on INbreast for 30k iterations with batch size 32. The initial learning rate is set to 0.001 and decreases by 0.1 after 10k and 20k iterations.

Mass segmentation. We perform extensive experiments on INbreast (Moreira et al., 2012) to validate the employed CED network with nested and dense skip connections (v19U-Net++, Sect.4.3.4). We compared it with the baseline U-Net (Ronneberger et al., 2015a) as well as two other recently published architectures: cGAN (Singh et al., 2020) and cascaded U-Net (Sect. 4.2, Yan et al. (2019a)). Experiments are carried out using the same train-test splits on INbreast examinations as in the previous stage. Training image crops are extracted around ground truth masses and resized to 256×256 pixels. Histogram equalization is then used to enhance the contrast. We train each model with a batch size of 4, Adam optimizer and Dice loss (the cGAN network loss is formulated by combining binary cross entropy and Dice losses) defined as $\frac{2TP}{2TP+FP+FN}$ where TP, FP, TN, and FN are the pixel level true positives, false

Metrics	T1	T2	T3	T4	T5	average
AP (%)	78.64	70.24	76.11	79.05	73.28	75.46±1.7

Table 4.2 – Performance of YOLOv3 (Redmon & Farhadi, 2018) on the INbreast (Moreira et al., 2012) dataset using average precision (AP) scores. T1 to T5 correspond to 5 experimental test sets.

positives, true negatives and false negatives. We use pre-trained weights from ImageNet (Deng et al., 2009) and then train models until convergence.

4.3.5.3 Mass localization

We evaluate the detection performance of YOLOv3 by calculating the average precision (AP) score for masses present in each test set. Fig.4.8 shows precision-recall curves for each test set using an intersection over union ≥ 0.5. Precision-recall curves summarize the trade-off between the true positive rate and the positive predictive value using different probability thresholds. Then, we compute the average precision scores which summarize the weighted increase in precision with each change in recall for the thresholds in the precision-recall curve. From Fig.4.8, we can clearly see that the precision-recall curves are fairly consistent between different test sets, which demonstrates the consistency of YOLOv3. Tab.4.2 displays the corresponding AP scores of each curve. YOLOv3 yields an averaged AP of 75.46% with a standard error of 1.7. For comparison, most state-of-the-art methods achieve a mean AP of 80% on PASCAL VOC and 60% on MS-COCO, which reveals very reasonable precision given the complexity of the mass detection task.

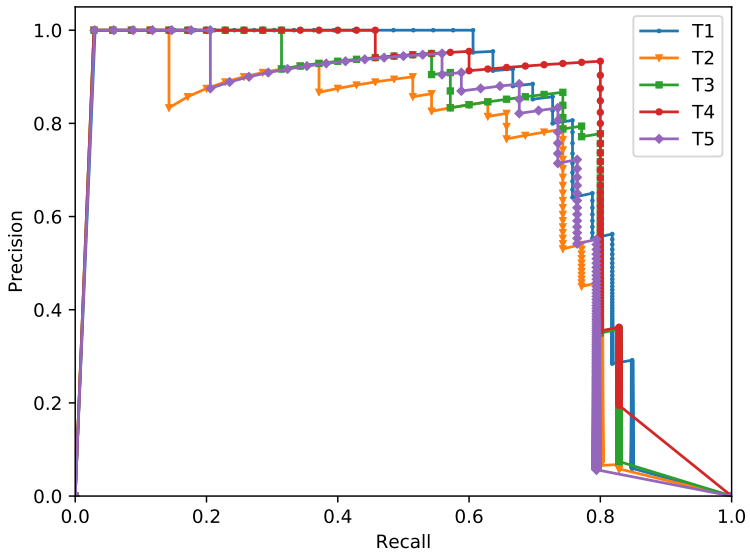


Figure 4.8 – Precision-recall curves of the YOLOv3 (Redmon & Farhadi, 2018) detection results on 5 test sets (from T1 to T5) extracted from the INbreast (Moreira et al., 2012) dataset.

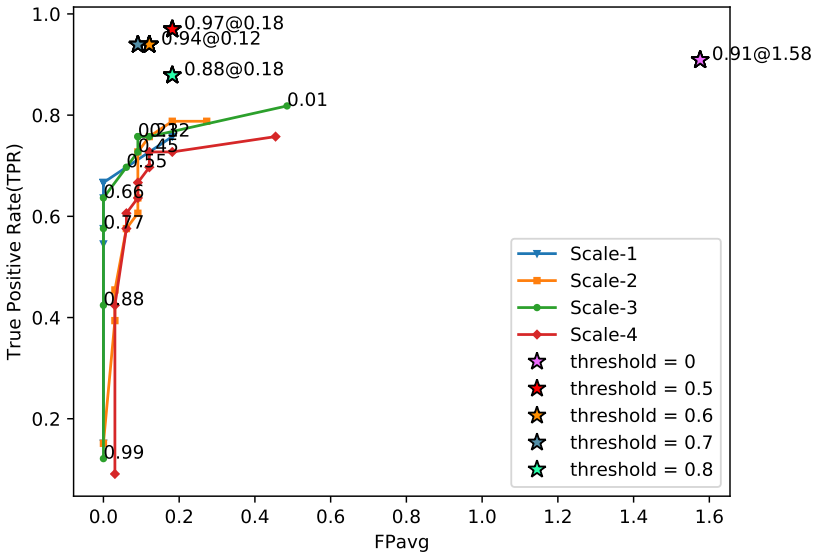


Figure 4.9 – Free response operating characteristic (FROC) curves of detection results on INbreast (Moreira et al., 2012), representing true positive rate (TPR) and average false positive per image (FPavg). Curves from Scale-1 to Scale-4 display results of single-scale predictions at 160×320, 256×512, 320×640 and 480×960. Stars show TPR@FPavg of the final decision at fixed thresholds.

We fuse prediction results obtained at resolutions 160×320, 256×512, 320×640, 416×832 and 480×960 for multi-scale fusion (Sect.4.3.3). We use free-response receiver operating characteristic (FROC) as evaluation criterion. Fig.4.9 illustrates the performance of MSF for test set T1 as an example. The FROC curve is created by plotting the true positive rate (TPR) against the average false positive per image (FPavg) using various thresholds. Since MSF uses an empirical threshold λ to make final decisions, we tested a set of thresholds $\lambda \in \{0, 0.5, 0.6, 0.7\}$ to get different TPR@FPavg scores. $\lambda = 0$ means that we keep all the detections of YOLO, while $\lambda = 0.5$ means that we keep the part of mask $M_s \geq 0.5$ (Eq.4.1) and so on. Apart from displaying the FROC curve of each scale, we also use stars (Fig.4.9) to show the final detection TPR@FPavg scores of MSF under different thresholds λ. Fig.4.9 indicates that TPR@FPavg scores of MSF are all located in the upper left corner of FROC space, showing that our MSF strategy largely boosts the accuracy of mass localization compared to single-scale detections, with a more reliable TPR and less FP proposals. Additionally, the TPR@FPavg scores shown in Tab.4.3 highlights the influence of λ. With a higher threshold, the false positives tend to be reduced while the TPR reaches the peak levels at around $\lambda = 0.5 \sim 0.6$. We finally chose $\lambda = 0.6$ considering the trade-off between true-positives and false-positives proposals.

We also compare the image-based mass detection with respect to state-of-the-art using TPR@FPavg (Tab.4.4). Even if results are only for reference since datasets used for training and testing are not identical, it highlights that MSF (0.94@0.22) significantly outperforms (Agarwal et al., 2019; Ribli et al., 2018; Sapate et al., 2020) in both TPR and FPavg and shows consistent TPR with respect to Dhungel et al.

λ	TPR@FPavg				
	T1	T2	T3	T4	T5
$\lambda = 0$	0.91@1.58	0.97@1.39	0.89@1.30	0.91@1.55	1.0@0.87
$\lambda = 0.5$	0.97@0.18	0.94@0.27	0.91@0.18	0.92@0.36	0.97@0.12
$\lambda = 0.6$	0.94@0.12	0.94@0.24	0.91@0.18	0.92@0.30	0.97@0.06
$\lambda = 0.7$	0.94@0.09	0.89@0.27	0.91@0.15	0.89@0.18	0.97@0.06

Table 4.3 – Performance of the proposed MSF method on INbreast (Moreira et al., 2012) using TPR@FPavg scores with different λ. T1 to T5 correspond to the 5 experimental test sets.

(Dhungel et al., 2017a) (0.95@5) while providing less FP.

Methods	TPR@FPavg	dataset	images
Sapate et al. (2020)	0.88 @1.51	DDSM	148
Ribli et al. (2018)	0.90 @ 0.3	INbreast	107
Dhungel et al. (2017a)	0.95 @ 5	INbreast	410
Agarwal et al. (2019)	0.92 @ 0.5	INbreast	410
YOLOv3+MSF (ours, $\lambda = 0.5$)	0.94 @ 0.22	INbreast	107

Table 4.4 – Detection performance comparisons between the proposed MSF and state-of-the-art. Our provided TPR@FPavg score is the average of T1 to T5 test sets at $\lambda = 0.5$.

4.3.5.4 Mass segmentation

To assess the final segmentation performance, we compute Dice scores over each test set on full mammograms for each different methodology (Tab.4.5). Compared to U-Net (Ronneberger et al., 2015a) (89.20±0.5), results of cascaded U-Net (89.49±0.3) are slightly better since it employs a multi-scale cascade of U-Net combing auto-context. The gain is relatively low considering that cascaded U-Net (Sect. 4.2, Yan et al. (2019b)) has been designed to tackle mass segmentation from entire mammograms. cGAN (Singh et al., 2020) also brings slight benefits (90.02±0.2) to the original U-Net but less than v19U-Net++ (Z. Zhou et al., 2018) which yields the best results on all test sets with 90.86% as average Dice score.

To assess the final segmentation performance of the proposed two-stage system (Fig.4.4), we compare the overall Dice on full mammograms from different methods. As a proof of concept, we test the second stage (Sect.4.3.4) using the candidate patches arising from the first stage (Sect.4.3.3), which are resized to 256×256 pixels before feeding into segmentation models. Tab.4.6 presents comparative evaluations for each model: one-stage segmentation, two-stage segmentation without and with the proposed MSF on high-resolution full mammograms. In particular, in the two-stage without MSF setup, mass candidates are provided by a simple single-scale prediction of YOLOv3.

Comparisons between models indicate that v19U-Net++ yields better segmentation results for two-stage segmentation, with an average Dice score of 70.96% without MSF and 80.44% with MSF. Compared

Methods	T1	T2	T3	T4	T5	average (%)
U-Net (Ronneberger et al., 2015a)	90.47	89.76	88.16	87.97	89.66	89.20±0.5
cGAN (Singh et al., 2020)	90.30	90.53	89.70	89.33	90.22	90.02±0.2
cascaded U-Net (Yan et al., 2019a)	89.20	90.40	88.83	89.18	89.82	89.49±0.3
v19U-Net++ (Z. Zhou et al., 2018)	**90.94**	**91.42**	**90.56**	**90.23**	**91.13**	**90.86±0.2**

Table 4.5 – Average Dice score (%) of different patch-based deep segmentation methods on INbreast (Moreira et al., 2012) mass patches centered around ground truth masses. Best scores are in bold.

Method	**Setup**	T1	T2	T3	T4	T5	average (%)
U-Net (Ronneberger et al., 2015a)	one-stage	43.66	44.12	45.93	40.79	47.36	44.37±1.1
	two-stage w/o MSF	70.59	68.46	70.56	74.66	66.06	70.07±2.8
	two-stage with MSF	**77.40**	**83.07**	**75.45**	**77.80**	**82.47**	**79.24±1.5**
cGAN (Singh et al., 2020)	one-stage	25.27	30.91	24.74	23.21	40.45	28.92±3.2
	two-stage w/o MSF	70.28	66.93	70.22	74.93	63.73	69.22±3.7
	two-stage with MSF	**75.66**	**81.66**	**76.70**	**77.44**	**83.45**	**78.98±1.5**
cascaded U-Net (Yan et al., 2019a)	one-stage	64.37	61.56	65.63	65.35	70.55	65.49±1.5
	two-stage w/o MSF	70.89	67.78	70.01	73.35	65.02	69.81±3.4
	two-stage with MSF	**75.76**	**82.51**	**76.78**	**77.69**	**83.16**	**79.18±1.5**
v19U-Net++ (Z. Zhou et al., 2018)	one-stage	53.38	49.38	47.44	48.85	61.80	52.17±2.6
	two-stage w/o MSF	72.18	68.55	72.27	76.10	65.69	70.96±3.6
	two-stage with MSF	**77.51**	**84.38**	**77.39**	**78.80**	**84.12**	**80.44±1.6**

Table 4.6 – Average Dice score (%) obtained on final delineations from 2048×1024 full INbreast (Moreira et al., 2012) mammograms. Best scores are in bold.

with one-stage segmentation, a significant gap is crossed when using a two-stage scheme, demonstrating the effectiveness of our two-stage localization-segmentation design. MSF brings Dice improvements to the two-stage scheme from 9.17% with U-Net to 9.76% with cGAN (9.48% with v19U-Net++), showing that adding the MSF strategy into the pipeline can further greatly improve performance. We also observe that one-stage segmentation methods reach various levels of robustness (Sect. 4.2, Yan et al. (2019a)) when applied to high-resolution mammograms: from 28.92% (cGAN) to 65.49% (cascaded U-Net). Conversely, our two-stage scheme provides more stable and reliable results, which suggests that it could be very effective in clinical practice.

Evaluation is supplemented with qualitative results. Fig.4.10 shows full mammogram detection and segmentation results using the proposed two-stage with MSF compared to two-stage without MSF. We observe that by using the MSF strategy, we have considerable improvements in both mass localization accuracy and mass delineation precision. It also shows that we can successfully detect multiple masses in a single mammogram. In addition, we compare in Fig.4.11 the proposed method with cascaded U-Net (Yan et al., 2019a) which also addresses full mammogram segmentation. Our method obtains more accurate detections and boundary adherence, while almost all false-positive proposals are eliminated. Moreover, the method is robust in dealing with masses of any size, shape or texture. This confirms that our methodology is very generalizable in handling the problem of strong class imbalance and tumor

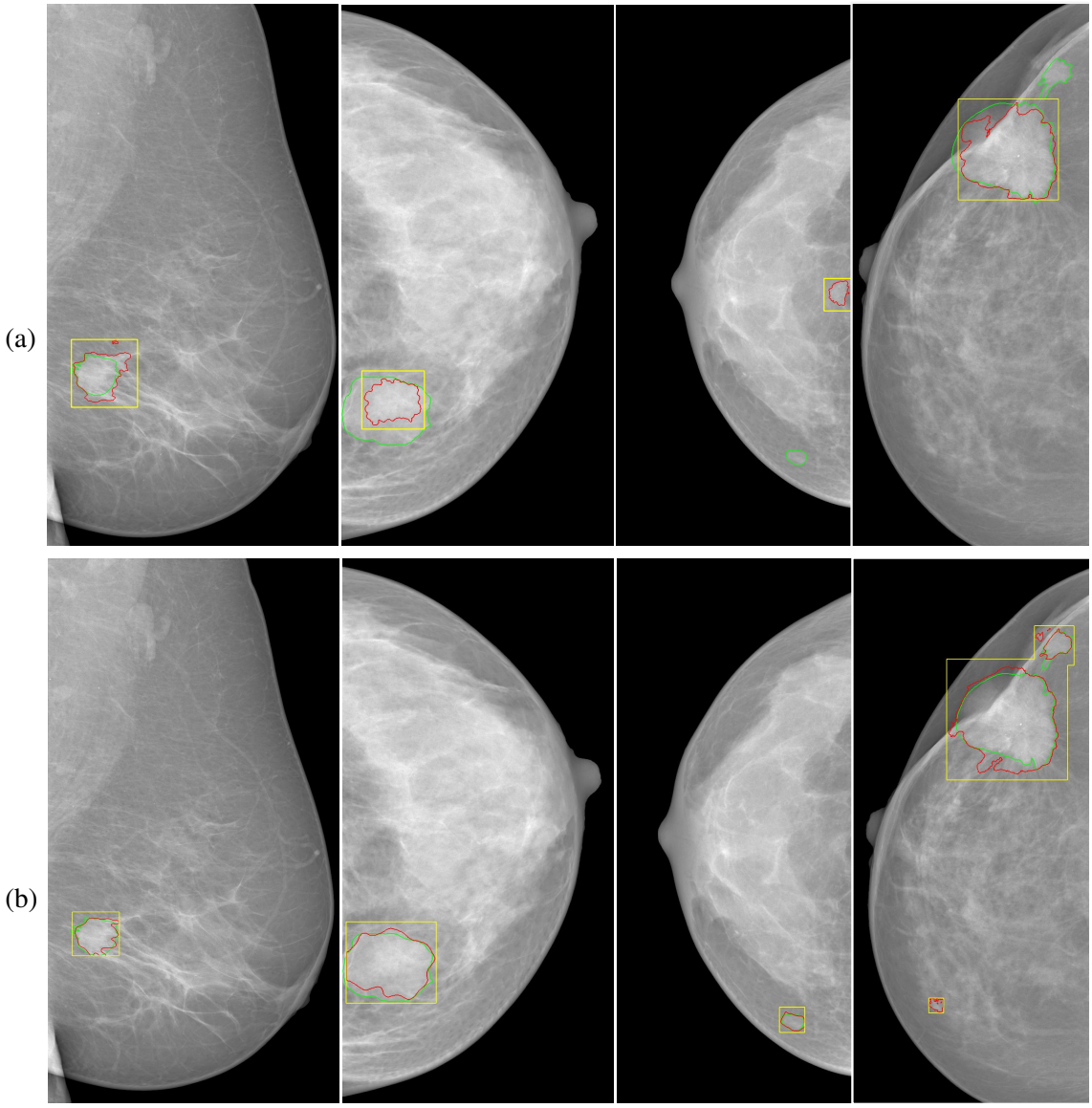


Figure 4.10 – Mass segmentation using our two-stage method without (a) and with (b) multi-scale fusion (MSF). Yellow, red and green stand for final detection, segmentation and ground truth.

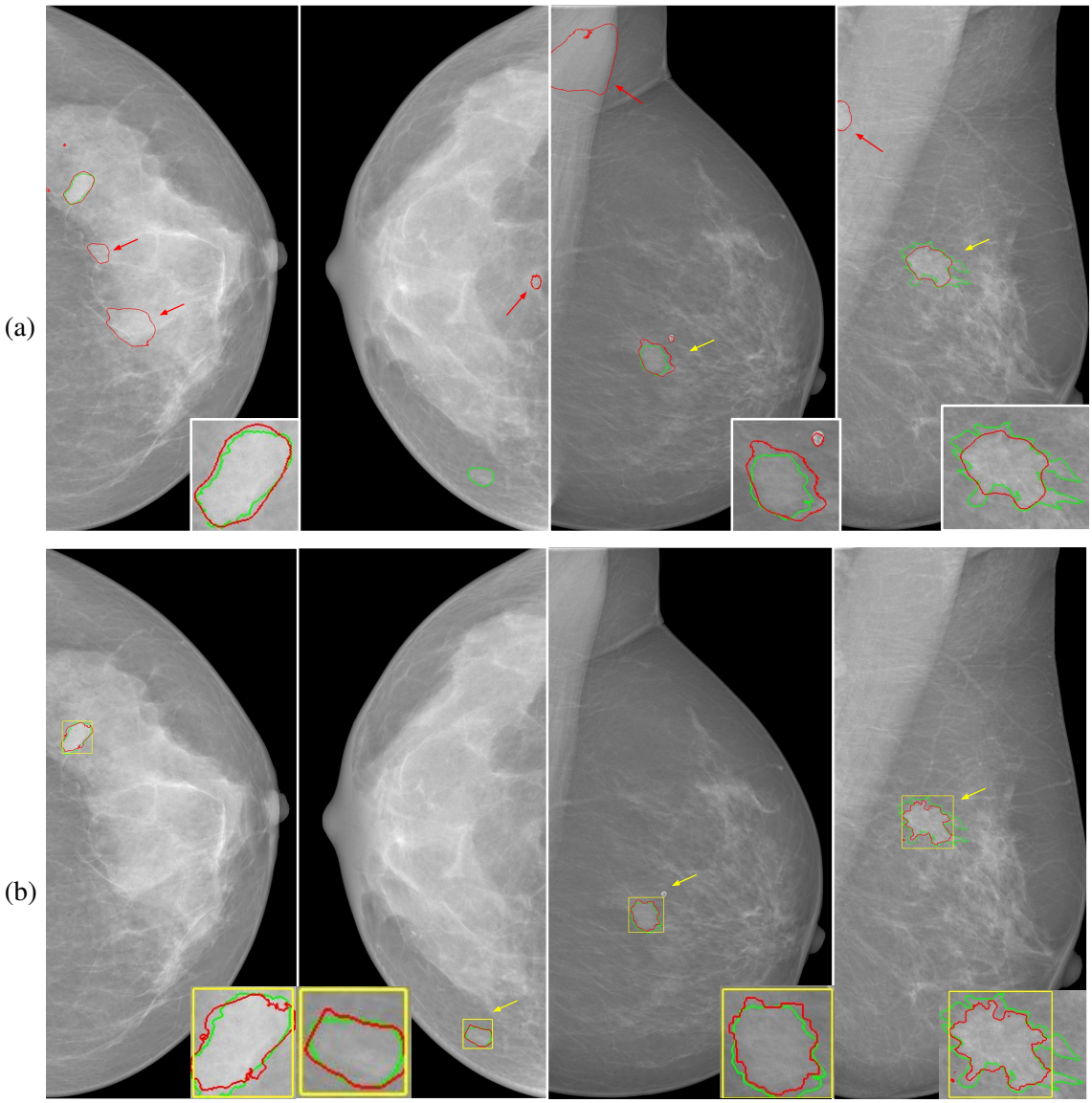


Figure 4.11 – Mass segmentation using cascaded U-Net (Yan et al., 2019a) (a) and our two-stage method with MSF (b) on INbreast (Moreira et al., 2012) images. Yellow, red and green lines stand for final detection, segmentation and ground truth contours. Yellow (red) arrows highlight true-positive (false-positive) cases.

appearance variability.

4.3.6 Discussion

In this section, we proposed a two-stage framework combining a deep, coarse-scale mass detection with a new multi-scale fusion strategy and a fine-scale mass segmentation using dense and nested skip connections. Our system achieves an overall average Dice of 80.44% on INbreast test images, which sets the state-of-the-art performance in mass segmentation on the publicly available INbreast dataset, outperforming one-stage segmentation schemes such as cGAN (Singh et al., 2020) (28.92%), U-Net (Ronneberger et al., 2015a) (44.37%) or cascaded U-Net (Yan et al., 2019a)) (65.49%). The newly designed MSF brings Dice improvements to the two-stage scheme from 9.17% (U-Net) to 9.76% (cGAN). By fusing predictions performed at multiple scales, we avoided manual selection and drastically reduced the number of unsuccessful pre-detections while allowing a variable number of candidate regions to be automatically selected for segmentation. These results confirmed the model robustness and generalizability of the proposed pipeline, leading to more reliable full-mammogram mass segmentation without any user intervention, and thereby pushing forward the implementation of realistic CAD systems.

4.4 Conclusion

In this chapter, we studied the problem of automated mass segmentation from high-resolution full mammograms, towards realistic CAD systems that deal with lesion segmentation from native high-resolution medical images. To cope with the ensuing problems such as strong class imbalance, huge diversity of lesion size, shape, texture and contour as well as limited receptive field, we put forward two approaches.

In Sect. 4.2, we extended standard segmentation pipelines to multi-scale cascades of deep convolutional encoder-decoders. Contextual information extracted at each level was combined using auto-context. End-to-end training was followed to benefit from simultaneous multi-scale training. Results on INbreast showed promising model generalizability, especially when transfer learning is employed from DDSM-CBIS. This proof-of-concept using U-Net suggests that embedding robust residual or adversarial models in such cascaded setup could achieve a further step forward for better mammogram analysis.

In Sect. 4.3, we proposed a two-stage multi-scale framework, which works as an accurate and automatic mass localization and segmentation CAD system. First, the deep network roughly localizes masses of any size, position and shape from the whole image by fusing predictions at multiple scales. Second, we performed an effective patch-based deep segmentation method with nested and dense shortcuts to obtain the accurate delineation of mass contours. Our system showed promising accuracy as an automatic full-image mass segmentation system. Extensive experiments revealed robustness against the diversity of size, shape and appearance of breast masses, towards better interaction-free computer-aided diagnosis.

The proposed approaches can be easily integrated into clinical routine and is able to help diagnosis

by acting as a relevant fully-automated second opinion. Future research should consider the potential effects of fusing multi-view and contralateral symmetry information to increase the robustness of breast lesion detection and delineation and therefore improve clinical guidance. Furthermore, our framework is generic enough to be extended to other medical imaging modalities for both anatomical and pathological structure segmentation.

CHAPTER 5

MULTI-VIEW INFORMATION FUSION

Contents

5.1 Introduction

Mammography screening involves two standard views acquired for left and right breasts: craniocaudal (CC), extracted from top-down, and mediolateral-oblique (MLO), an oblique view taken under 45°.

These two views comprise routine screening mammography. The information presented in the paired CC/MLO views is highly complementary and could serve as a second source of decision (Jouirou et al., 2019). Compared to single-view screening, examining the correspondence between suspicious findings in multiples views enables radiologists to reduce false-positive cases, improve clinical interpretations and subsequent decisions (Vijayarajan & Jaganathan, 2014), thus improving cancer detection rates (Warren et al., 1996). Therefore, the dual-view analysis is considered an effective way to reduce the morbidity and mortality associated with breast cancer (Jørgensen & Bewley, 2015) and is key to make decisions in clinical routine. However, due to breast deformation and different acquisition conditions combined with the lack of 3D information, multi-view fusion for dual-view mammogram analysis is challenging. Therefore, only a few deep methods for breast screening consider learning jointly effective features from both views. The use of multi-view context is a known weakness of current CAD technology. Thus, there is a huge potential to improve the performance of CAD tools by integrating information from paired views.

The concept of multi-view information fusion was recently introduced to improve the performance of detection, classification or content-based mammogram retrieval tasks (Jouirou et al., 2019). An increasing number of works focus on multi-view mammography analysis. Vijayarajan and Jaganathan (2014) extracted 2D features from whole mammograms, obtained the component location value from CC and MLO views and merged this information to get a 3D view of masses in the mammogram image. Carneiro et al. (2015) trained a separate CNN model for each view and finally applied a CNN classifier that estimates the BI-RADS score using features learned from unregistered CC and MLO mammograms, as well as respective mass delineations. Geras et al. (2017) proposed to apply a CNN model separately to each view to obtain view-specific representations for further classification purposes. All the above studies are designed based on whole mammograms. However, there may be multiple different benign or malignant masses in a given examination. In order to simplify the complex analysis of whole mammograms, some studies assign a unique label (benign, malignant or normal) to the whole image. The drawback is that it avoids conducting a comprehensive analysis of each mammogram, comprising lesion types and locations.

In this chapter, we introduce two multi-view information fusion methods. Sect. 5.2 presents a dual-view multi-tasking combined network for breast mass matching, classification and segmentation. Sect. 5.3 proposes an active learning based dual-view mammogram analysis approach where the dual-view prediction consistency is integrated as selection criterion. The work described in Sect. 5.2 has been published in the journal of Medical Image Analysis (Yan et al., 2021b) whereas Sect. 5.3 has been presented at the Machine Learning in Medical Imaging (MLMI 2021) MICCAI workshop (Yan et al., 2021c) .

5.2 Dual-view mammogram matching for improved breast mass detection

5.2.1 From single-image to dual-view mammogram analysis

In recent years, CAD systems that employ deep learning have demonstrated stronger robustness in clinical implementation than traditional methodologies. Nevertheless, breast mass detection, segmentation and classification are still open issues due to the strong variations in mass appearance. Some studies (L. Shen et al., 2019; X. Zhang et al., 2018; Zhu et al., 2017) focus on whole mammograms which simplify such complex problem by providing a unique image-level label (normal, benign or malignant), without conducting a comprehensive analysis comprising lesion types and locations. Other works are mostly region-based methods (Arevalo et al., 2015; Choukroun et al., 2017; Lévy & Jain, 2016; H. Wang et al., 2018; H. Zhou et al., 2017), where images are first decomposed into regions to further distinguish normal from abnormal tissues. However, most of the above methods use single-view mammograms only, thus neglecting the rich information that can be extracted from multi-view images.

To address the limitation of single-view processing, we aim at taking advantage of information arising from CC and MLO mammograms, as do clinicians when making decisions in clinical practice (Vijayarajan & Jaganathan, 2014). Several multi-view fusion schemes learn full images from each view separately and concatenate respective features afterwards. Geras et al. (2017) proposed to apply CNN models to each view separately to obtain view-specific representations for further classification purposes. Nevertheless, such late-fusion schemes only exploit image-level view-specific representations. Alternatively, we propose a novel multi-tasking Siamese deep model that combines CC and MLO mammograms to improve breast mass detection. Our contributions are two-folds. First, we propose a new deep learning algorithm that capitalizes on multi-view fusion and multi-task learning to improve breast mass detection. To the best of our knowledge, our framework is the first that exploits multi-tasking abilities of deep learning models to improve mass detection using multi-view matching. Second, we conduct a comprehensive evaluation of various networks towards multi-task learning on public datasets. Both quantitative and visual results prove the effectiveness of the proposed strategy.

Based on MatchNet (Han et al., 2015), Perek et al. (2018) proposed a dual-view Siamese network (Koch et al., 2015) (Sect. 3.2.4) that learns patch representations and similarity for lesion matching. This suggests a potential added value of multi-view matching to improve breast mass detection, with respect to single-view detection strategies. However, these are single-task studies dedicated to mass detection (Ma et al., 2019) or mass matching (Perek et al., 2018) only.

To design a more comprehensive and efficient CAD system, we aim at exploiting the multi-tasking properties of deep CNN. Multi-task learning processes multiple tasks jointly with many advantages such as saving computation time and resources as well as improving robustness against overfitting (Ruder, 2017) (Sect. 3.3.2). The network parameters from feature extraction layers are updated through the optimization of a combined loss dealing with both mass/non-mass classification and matching. Contrary to (Ma et al., 2019; Perek et al., 2018), our method can provide both classification and matching results.

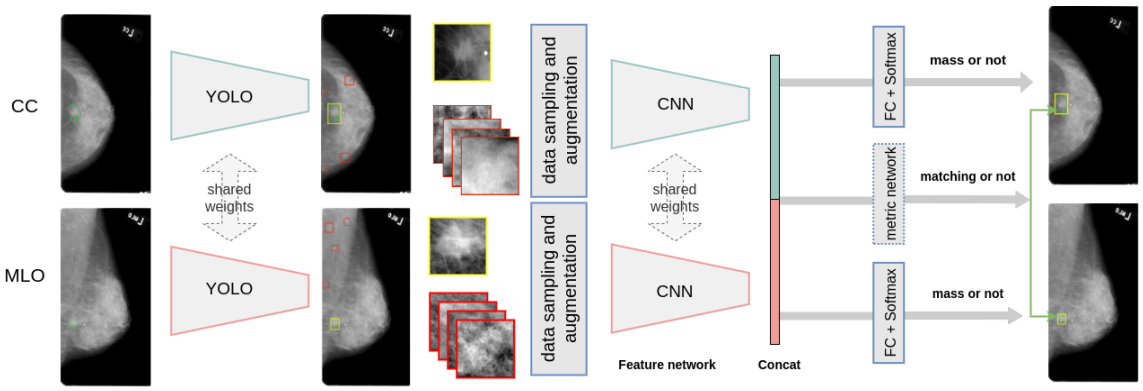


Figure 5.1 – Proposed multi-tasking deep pipeline. In images, green contours indicate ground truth delineations, red and yellow boxes respectively indicate false and true detections.

5.2.2 Methods

In this section, we propose a novel multi-tasking Siamese deep model that combines CC and MLO mammograms to improve breast mass detection. We first formally define the problem settings and provide an overview of the proposed unified framework for mass classification and matching in Sect.5.2.2.1. Multi-view mass matching combining Siamese networks (Sect. 3.2.4) and contrastive learning is described in Sect.5.2.2.2. Multi-task learning (Sect.5.2.2.3) is followed to obtain better predictive breast mass classification performance, towards improved mass detection than traditional single-task learning schemes.This methodology is then extended to address classification, matching and segmentation simultaneously (Sect.5.2.2.4) as a supplementary test.

5.2.2.1 Overview

Our multi-tasking framework (Fig.5.1) takes unregistered CC/MLO view pairs as inputs and provides as output accurate mass detections along with correspondences between mass regions in both views. Among existing deep detectors including Faster R-CNN (Ren et al., 2015) or SSD (W. Liu et al., 2016), YOLOv3 (Redmon & Farhadi, 2018) is adopted for candidate patch generation and selection from full mammograms, since it offers a good trade-off between accuracy and efficiency (Sect. 3.2.3).

Given a pair of mammograms $\{I_{CC}, I_{MLO}\}$, YOLO predicts two sets of candidate mass patches $P_{CC} = \{p^1_{CC}, ..., p^N_{1\ CC}\}$ and $P_{MLO} = \{p^1_{MLO}, ..., p^N_{2\ MLO}\}$. Although recent deep learning-based detectors have yielded impressive accuracy for object detection in natural images, it still remains difficult to reach the same level of performance when applied to medical images, especially mammograms. The following reasons arise. First, object size variance may affect the performance. In our context, mass sizes and aspect ratios may strongly vary (Yan et al., 2019b). Second, mass detection is generally more difficult than common object detection since masses are visually less obvious and less contrasted with their surrounding healthy tissues, combined with a great diversity of shape and texture. On top of that,

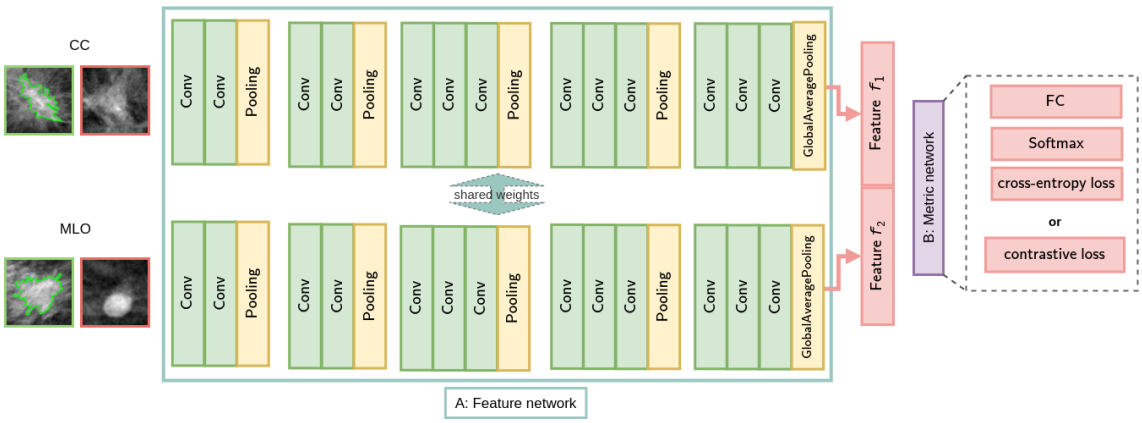


Figure 5.2 – Matching Siamese network. A: Two-branch feature network which takes as input both positive (green patch) and negative (red patch) patch samples of CC and MLO views separately to compute features. Resulting features f_1 and f_2 are concatenated for patch comparison. B: Metric network. Green contours indicate ground truth delineations.

we should also struggle with the barrier between true and false masses to retain as much as possible true positives while reducing false positives.

To further finely select mass candidates and discover the latent relation between CC and MLO views, we design a combined model through a Siamese network that jointly deals with patch-level mass/non-mass classification and matching (Fig.5.1). We sample candidate mass patches P_{CC} and P_{MLO} to the same size via a data sampler, while performing data augmentation to prevent from overfitting. These samples are then fed into our combined network. Based on robust generic feature extraction, the result of our model is whether each patch of the two views contains mass as well as the correspondence between two patches arising from each of the two views. Subsequently, we can visualize final detection results on both views to further guide clinicians in their mammogram interpretation task.

5.2.2.2 Dual-view mammogram matching

Inspired by (Perek et al., 2018) and (Han et al., 2015), we employ a Siamese framework to identify correspondences between masses in both CC/MLO views. The deep architecture for multi-view mammogram matching is shown in Fig.5.2. Patch pairs from CC and MLO views are fed separately to the two branches of the network. The feature network A is a Siamese model in which two fully convolutional networks with shared weights are employed for feature extraction. For illustration (Fig.5.2), we use a VGG16 architecture (Simonyan & Zisserman, 2014) with repeated 3×3 convolutions followed by an activation function (ReLU) and 2×2 max pooling. To reduce the number of parameters while avoiding overfitting, we apply a global average pooling layer before subsequent FC layers (He et al., 2016b; Szegedy et al., 2015). Particularly, different widely used deep convolutional models such as VGG16 (Simonyan & Zisserman, 2014), ResNet50, ResNet101 (He et al., 2016b), InceptionV3 (Szegedy et al., 2016) and

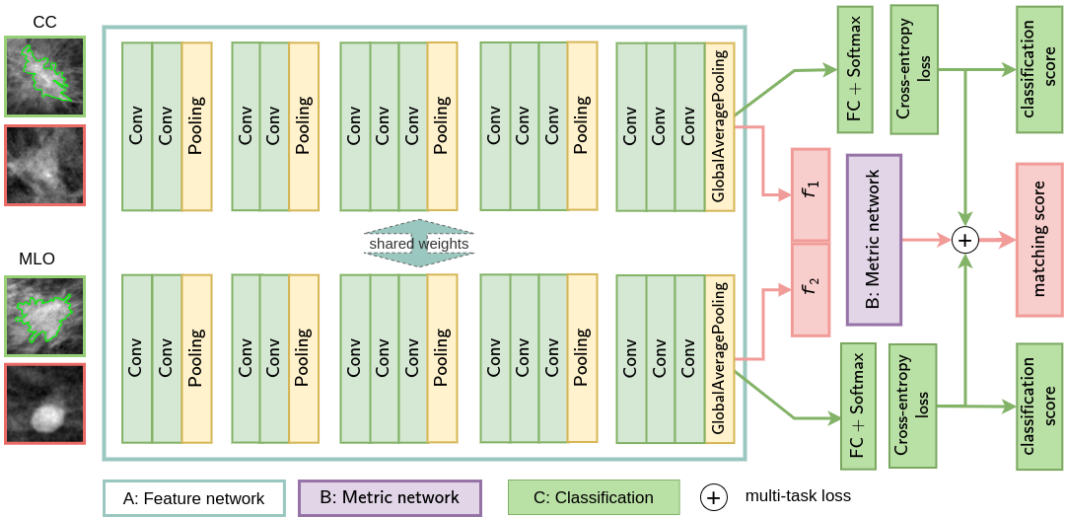


Figure 5.3 – The proposed Combined Matching and Classification Network (CMCNet). Green (red) patches correspond to positive (negative) samples. Green contours indicate ground truth delineations.

EfficientNet (Tan & Le, 2019) can be exploited for feature extraction purposes. For feature comparison, two manners are explored based on different loss functions. First, one can use a metric network as in (Perek et al., 2018) and (Han et al., 2015) consisting of several FC layers and softmax layers, trained with a cross-entropy loss. Alternatively, we can rather employ a contrastive loss (Hadsell et al., 2006) to improve the representation ability of network A to extract discriminative features.

Contrastive loss for matching. Contrastive learning, whose labels are used to guide the choice of positive and negative pairs, is employed to learn powerful feature representations. The contrastive loss is usually exploited for image retrieval tasks, along with Siamese networks to learn paired data relationships. During training, an image pair is fed into the model with their ground truth relationship Y. The loss function is defined as follows:

$$L_{mat}\left(Y, X_1, X_2\right) = \frac{1}{2N}\sum_{n=1}^{N} Y D_W\left(X_1, X_2\right)^2 + (1-Y)\max\left(m - D_W\left(X_1, X_2\right), 0\right)^2 \tag{5.1}$$

where $D_W\left(X_1, X_2\right) = \|f_1 - f_2\|_2$ represents the Euclidean distance between two sample features f_1 and f_2 estimated from X_1 and X_2. Y is the label of whether the two samples match: $Y = 1$ if the two samples are similar (correspond to the same anatomical location) and 0 otherwise. $m > 0$ is a margin that defines a radius: dissimilar pairs contribute to the loss only if their distance is within this radius. N is the number of samples. Unlike conventional learning systems where the loss function is a sum over samples, the contrastive loss runs over pairs of feature vectors $\{f_1, f_2\}$ such that there is no more need for FC and

softmax layers. Moreover, compared to cross-entropy which learns the patch "match" or "not match" in an inexplicable manner, the contrastive loss optimizes the mass matching task by manipulating the distance between pairs in feature space. Therefore, the contrastive loss is more in line with matching requirements than binary sample classification. The loss function L_{mat} (Eq.5.1) is minimized using stochastic gradient descent (SGD).

5.2.2.3 Combined classification and matching

Mass classification and dual-view matching are two tasks of a very different nature. The challenge is thus to learn generic features for both tasks. We propose to exploit Siamese networks towards simultaneous deep patch-level matching and classification. In this direction, we design a multi-tasking learning model (Fig.5.3) referred to as Combined Matching and Classification Network (CMCNet). Positive and negative patch samples of CC/MLO views arising from YOLOv3 detector are fed into the two-branch feature network (Fig.5.3A) to compute robust patch representations. Apart from the matching network (Sect.5.2.2.2, Fig.5.3B), we incorporate into the pipeline two branches (Fig.5.3C) for CC/MLO mass classification purposes. Each of these branches has its own FC layers. We not only jointly learn representations from the two views but also simultaneously learn matching and classification tasks to exploit the potential relationship between view-points.

The combined learning of classification and matching refers to the idea of multi-task learning which has been proven to improve learning efficiency and generalization performance of task-specific models. We expect thus that the dual-view matching task can improve the robustness of mass classification, towards better predictive results than classification-only strategies. The designed loss L is the sum of three loss terms to optimize the entire CMCNet parameters through SGD:

$$L = \alpha L_{cls,CC} + \beta L_{cls,MLO} + \gamma L_{mat} \tag{5.2}$$

where $L_{cls,CC}$ and $L_{cls,MLO}$ represent the classification losses (cross-entropy) for CC and MLO view respectively. L_{mat} is the matching loss which can be cross-entropy or contrastive loss (Eq.5.1). α, β and γ are coefficients balancing the loss terms.

5.2.2.4 Combined classification, matching and segmentation

Once the combined classification and matching step has been completed, our second goal is to integrate the mass segmentation task into the multi-tasking network, i.e., to build a final network which simultaneously performs three tasks: classification, matching and segmentation of breast masses.

Apart from mass classification and matching, mass segmentation is another task that plays an essential role in mammogram analysis. In contrast to these previous tasks which are performed at the patch-level, the segmentation task deals with pixel-level classification, that is, the dense classification of each pixel to identify whether it is part of a mass. Deep methods for breast mass segmentation are

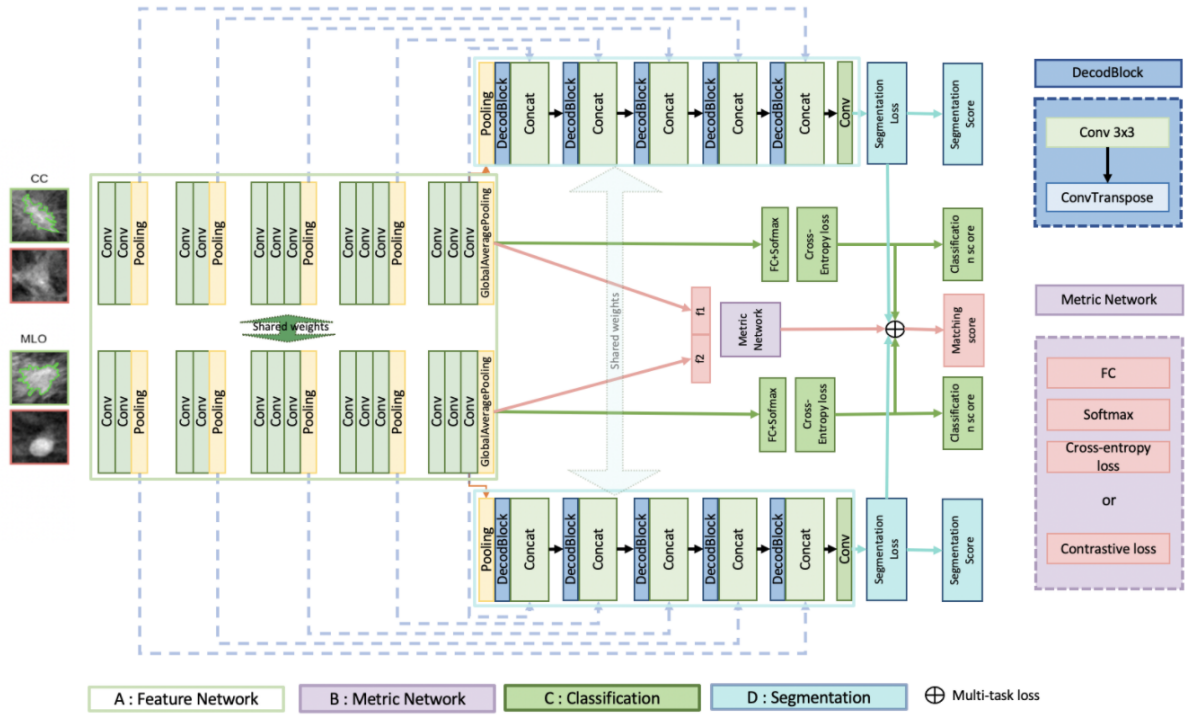


Figure 5.4 – Extension of CMCNet from two to three tasks: mass matching, classification and segmentation.

mainly based on convolutional encoder-decoder (CED) architectures, which is usually composed of two symmetrical branches: a contracting path (encoder) that gradually reduces the spatial dimension for feature extraction, and an expanding path (decoder) that progressively recover details to the initial resolution. Skip connections are usually designed to combine corresponding encoder and decoder feature maps to better recover high-level details.

As illustrated in Fig. 5.4, in the basis of the aforementioned CMCNet (Fig. 5.3), we added a decoder after each feature extraction branch of the Siamese network network. We used skip connections between each encoder-decoder as done in many other deep CED models (Ronneberger et al., 2015a; Z. Zhou et al., 2018). The final network is therefore composed of two feature extraction branches followed by two symmetrical decoder branches sharing the same weights. We expect thus that the dual-view matching task can improve the robustness of both mass classification and segmentation tasks. The segmentation task is supervised by the combination of binary cross-entropy (L_{bce}) and Dice (L_{dice}) losses following $L_{seg} = L_{dice} + \lambda_1 L_{bce}$ with:

$$L_{dice} = 1 - \frac{2|p \circ y|}{|p| + |y|} \tag{5.3}$$

	DDSM-CBIS		INbreast	
	training	validation	test	full-pipeline
classif.	4690/4690	1170/1170	700/700	125/225
matching	2345/4690	585/1170	350/700	125/225

Table 5.1 – Data distribution setting for experiments. Each cell has the following format: number of positive samples / number of negative samples.

$$L_{bce} = \begin{cases} -\log(p) & \text{if } y = 1 \\ -\log(1-p) & \text{otherwise} \end{cases} \tag{5.4}$$

where p and y represent the prediction mask and the ground truth mask respectively, $|.|$ and $\circ$ the pixel-wise sum and multiplication operations. The empirical factor λ_1 is a coefficient to balance the loss terms. The designed combined loss L is therefore the weighted sum of the losses of all tasks:

$$L = \alpha L_{cls,CC} + \beta L_{cls,MLO} + \gamma L_{mat} + \delta L_{seg,CC} + \eta L_{seg,MLO} \tag{5.5}$$

where $L_{cls,CC}$ and $L_{cls,MLO}$ represent the classification loss (cross-entropy) for CC and MLO view respectively. $L_{seg,CC}$ and $L_{seg,MLO}$ denote the segmentation loss for the two views. L_{mat} is the contrastive loss (Eq.5.1) employed for matching purposes. α, β, γ, δ and η are coefficients balancing the loss terms.

5.2.3 Experiments and results

5.2.3.1 Experimental setup

In this section, we evaluate the proposed approaches both quantitatively and qualitatively. Deep models for classification, matching and segmentation are implemented using pytorch. Experiments are performed with a single Nvidia GeForce GTX 1080Ti GPU (11GB/s).

Data selection. In this work, we selected 35 CC/MLO pairs (70 images) out of 107 INbreast (Moreira et al., 2012) (Sect.2.1.3.1) examinations with mass, and 586 CC/MLO pairs (1172 images) out of 2,500 DDSM-CBIS (Lee et al., 2017) (Sect.2.1.3.2) mammograms. During training, the principle followed for CC/MLO pair selection is that both views should contain only one mass. The reason is that if the image contains two or more masses, multi-view correspondences become unknown without the help of an expert. Accordingly, the existence of a single pair is the only data selection criterion for training. However, this is not a limitation for inference or clinical use. Among the 586 DDSM-CBIS pairs, 80% are used for training and the remaining 20% for validation. The 35 INbreast pairs are only used in the testing stage since it is too small to be representative as training data.

Network	learning rate	batch size	margin (m)	γ
VGG16	0.0005	128	15	0.1
ResNet50	0.0005	128	15	0.1
ResNet101	0.001	64	10	0.1
InceptionV3	0.001	32	10	0.1
EfficientNet-B3	0.001	128	10	0.1

Table 5.2 – Optimal hyper-parameters employed for each backbone.

Data sampling and augmentation. In our multi-task framework, sampling in training is crucial. However, regions of healthy tissues in a whole mammogram are much larger than the mass areas, leading to inevitable false positive YOLOv3 proposals. Similarly, sample imbalance can make the deep classification model very biased. To minimize these effects, data sampling is conducted as follows. For classification training, positive samples are extracted according to provided ground truth segmentation masks, while negative patches are generated by YOLOv3 (Sect.5.2.2.1). In particular, we randomly generate K patches per image with an intersection over union (IoU) with the ground truth box larger than 0.5. In practice, K=5 (respectively 10) for DDSM-CBIS (INbreast) since the INbreast dataset is much smaller. Likewise, we choose K negative patches from false YOLO predictions. We thus use a very small threshold ($<10^{-4}$) on detection probabilities to retain as many predictions as possible and select the K false candidates achieving the highest probabilities. All patches are resized to 64×64 pixels, as in Han et al. (2015) and Perek et al. (2018). Random rotations of 25 degree, random horizontal flips and random resized crops are applied for data augmentation. For matching, we consider a pair of positive patches of the same mass from the two views as a matching sample. If one of the patches is labeled negative, they are considered as a negative match. The detailed data distribution is shown in Tab.5.1 for both DDSM-CBIS and INbreast datasets.

Training patch-level classification and matching. As a proof of concept, we conduct experiments using various model backbones for the feature network: VGG16 (Simonyan & Zisserman, 2014), ResNet50, ResNet101 (He et al., 2016b), InceptionV3 (Szegedy et al., 2016) and EfficientNet (Tan & Le, 2019). The feature size varies depending on the model used. Let M be the number of feature map channels and B denote the batch size. The FC layers of each classification branch will turn the input vector (B, M) into $(B, 2)$ and pass it to the Softmax layer to transfer logits into probabilistic predictions. The input of the metric network (Fig.5.3B) is the concatenation of two feature vectors. For VGG16, $M = 512$. For ResNet50 and ResNet101, $M = 2048$. For EfficientNet, it has 8 pre-trained models from EfficientNet-B0 to B7 where M is respectively $\{1280, 1280, 1408, 1536, 1792, 2048, 2304, 2560\}$. All deep models are initialized using pre-trained weights (Litjens et al., 2017) from the ImageNet dataset (Russakovsky et al., 2015) and trained using the SGD optimizer. Optimal hyper-parameters vary depending on the network. For the global loss function, we choose $\alpha = \beta = 1$ as well as $\gamma = 1$ for cross-entropy and $\gamma = 0.1$ and margin $m \in \{5, 10, 15\}$ for contrastive loss (Eq.5.1). The detailed hyper parameters used are shown in Tab.5.2.

Network	matching	matching loss	CC acc	MLO acc	overall acc	p-value
VGG16 (Simonyan & Zisserman, 2014)	×	-	0.8558	0.8857	0.8699	-
	√	cross-entropy	0.8796	**0.9163**	0.8958	$< 1e^{-6}$
	√	contrastive	**0.9061**	0.9156	**0.9084**	$< 1e^{-6}$
ResNet50 (He et al., 2016b)	×	-	0.8517	0.9034	0.8734	-
	√	cross-entropy	0.8958	0.9116	0.9014	$4e^{-4}$
	√	contrastive	**0.9010**	**0.9122**	**0.9049**	$< 1e^{-6}$
ResNet101 (He et al., 2016b)	×	-	0.8680	0.9097	0.8823	-
	√	cross-entropy	**0.8980**	**0.9265**	**0.9098**	0.007
	√	contrastive	0.8891	0.9252	0.9049	0.003
InceptionV3 (Szegedy et al., 2016)	×	-	0.8238	0.8980	0.8601	-
	√	cross-entropy	0.8776	**0.9184**	0.8972	$< 1e^{-6}$
	√	contrastive	**0.8946**	0.9095	**0.9000**	$4e^{-4}$
EfficientNet-B3 (Tan & Le, 2019)	×	-	0.8701	0.8803	0.8741	-
	√	cross-entropy	0.8748	0.9116	0.8923	0.065
	√	contrastive	**0.8830**	**0.9163**	**0.8979**	$< 1e^{-6}$

Table 5.3 – CMCNet (with cross-entropy and contrastive losses) versus classification-only. Results include CC, MLO and overall classification accuracy (acc) as well as statistical significance p-values with respect to the classification-only baseline. Best results per network are in bold.

Training combined classification, matching and segmentation. The proposed model is trained and evaluated using the VGG16 (Simonyan & Zisserman, 2014) and ResNet50 (He et al., 2016a) model. All experiments are initialized using pre-trained weights from ImageNet (Russakovsky et al., 2015) and trained using SGD as optimizer. To study the impact of adding the segmentation task to the combined network, we trained this model with and without taking into account the other two tasks (classification, matching) in parallel. We assessed different variations of the empirical loss coefficient $\lambda_1 \in \{0, 0.5, 1\}$ to study the influence and complementarity of binary cross entropy and Dice loss ($\lambda_1 = 0$ means Dice loss only while $\lambda_1 = 1$ means that the two loss functions account in the same proportion). For the combined loss (Eq. 5.5), we choose $\alpha = \beta = \delta = \eta = 1$ and $\gamma = 1$ for cross-entropy and $\gamma = 0.1$ and margin $m = 10$ for contrastive loss.

5.2.3.2 Results

Multi-task learning versus classification-only. Classification performances are measured using classification accuracy (acc). We calculate the accuracy of each view separately (CC acc, MLO acc) and collectively (overall acc). The statistical significance of the multi-tasking model with respect to the classification-only baseline is estimated using Student's t-tests (Tab.5.3). Overall, we observe better classification results on MLO than on CC views. In most cases, multi-tasking models that combine classification and matching are better than classification-only from 2% to 4% in accuracy with statistical significance ($p < 0.05$), which reflects the benefits of dual-view matching. Except for ResNet101, we obtain slight gains with the contrastive loss compared to cross-entropy. However, the difference between networks is not obvious. ResNet101 achieves the best overall accuracy with statistical significance (acc = 0.9098, p = 0.007 compared to baseline). Improvements obtained by VGG16 using the contrastive loss are also

significant (acc = 0.9084, p < $1e^{-6}$), followed by ResNet50 (acc = 0.9049), InceptionV3 (acc = 0.90) and EfficientNet-B3 (acc = 0.8979), showing that using deeper networks is not necessary to reach better performance.

Multi-task learning versus segmentation-only. We compare in Tab. 5.4 the segmentation performance in dice score and the classification performance in accuracy (acc) to assess the proposed combined classification, matching and segmentation framework. Since negative patches have no segmentation mask, the dice score is calculated only on patches that contain mass. Among above-mentioned deep backbones (e.g. Tab. 5.3), we evaluated this framework on VGG16 and ResNet50 as preliminary tests. We noticed from Tab. 5.4 that combining classification and multi-view matching did not bring robust and significant improvements to the segmentation task. Best dice score (0.7386) is achieved by segmentation-only with VGG16, $\lambda_1 = 0.5$. For the classification task, results combining classification and matching tasks without the segmentation task (Tab. 5.5, underlined results) is still slightly better, with acc = 0.9084 compared to 0.8916 for VGG16, and acc = 0.9049 compared to 0.9007 for ResNet50, showing that the segmentation task could not bring further improvement to the classification performance. The lack of improvement may be due to differences in the nature of the segmentation task with respect to classification/matching tasks, i.e., pixel-level classification for the former, image-level classification based on global context for the latter. Moreover, general segmentation tasks normally include patches containing a mass, while the multi-tasking initiative introduces negative patches (i.e. patches with no mass), resulting in adding more negative samples (pixels) to the segmentation task. This undoubtedly increases the bias of the segmentation model. However, it is still worth noting that the model achieves better performance when the loss coefficient $\lambda_1 = 0.5$ for both multi-tasking (dice = 0.7119 for VGG16, dice = 0.7369 for ResNet50) and segmentation-only (dice = 0.7386 for VGG16, dice = 0.7349 for ResNet50). For the ResNet50 setting with $\lambda_1 = 0.5$, the multi-task learning achieves comparable dice (0.7369) with respect to the best dice (0.7386) with negligible calculation increase.

Full detection pipeline. To further prove the effectiveness of our method, we conduct experiments with a full detection pipeline. Here we employed the CMCNet without including the segmentation task as it brings no improvement to the other tasks. Instead of extracting positive candidates using ground truth mass delineations while using YOLO as a negative patch generator, we use YOLO to generate all candidate patches.

Specifically, coarse mass YOLO detections (Yan et al., 2021d) are performed on INbreast images to generate testing samples. YOLO is pre-trained on ImageNet and fine-tuned on 1514 DDSM-CBIS images. Thereafter, we use a small threshold (10^{-4}) on detection probabilities to ensure that predictions with high and low confidence are both selected. The averaged inference time per image is 78.7ms. We finally obtained 350 candidates, labeled as positive (125 cases) or negative (225 cases) according to IoU ($\geq$ or $<$ than 0.5) between RoIs and ground truth. All 350 candidate patches arising from INbreast are

Model	Segmentation loss	classification	matching	segmentation	dice	acc
VGG16	-	✓	✓	-	-	<u>0.9084</u>
	λ_1=0	✓	✓	✓	0.6450	0.8916
		-	-	✓	0.6650	-
	λ_1=0.5	✓	✓	✓	0.7119	0.8951
		-	-	✓	**0.7386**	-
	λ_1=1	✓	✓	✓	0.7085	0.8979
		-	-	✓	0.6068	-
ResNet50	-	✓	✓	-	-	<u>0.9049</u>
	λ_1=0	✓	✓	✓	0.6593	**0.9007**
		-	-	✓	0.6883	-
	λ_1=0.5	✓	✓	✓	0.7369	0.8972
		-	-	✓	0.7349	-
	λ_1=1	✓	✓	✓	0.6896	0.8944
		-	-	✓	0.6513	-

Table 5.4 – Combined classification, matching and segmentation versus segmentation-only. Underlined scores highlight the results without the segmentation task (i.e. the proposed CMCNet) using VGG16 and ResNet50 backbones for easy comparison. Results include the average segmentation dice score (calculated on patches containing mass) as well as overall classification accuracy (acc). Best results are in bold.

for the test set and all combinations are evaluated. The performance of each setting (classification-only, CMCNet with cross-entropy and CMCNet with contrastive loss using different backbones) is measured using the AUC (Area Under the receiver operating characteristics Curve).

Results for full-pipeline experiments (Tab.5.5) show that the classification performance is highly improved over the baseline models by combining classification with dual-view matching. In terms of AUC, the performance of VGG16 (resp. ResNet101) increases from 90.47% (71.46%) to 94.78% (92.82%), which corresponds to a gain of 4.31% (21.36%). These results prove the appropriateness of our contributions. The best AUC score (94.78%, $p = 0.001$) is obtained using the VGG16 model trained with contrastive loss, with an overall accuracy of 0.8791. Results using the contrastive loss are slightly better than cross-entropy in most cases, except for ResNet50. InceptionV3 using cross-entropy and EfficientNet using both losses improve moderately without statistical significance ($p > 0.05$). Compared to Tab.5.3, the advantages of combining classification and matching are more highlighted with full-pipeline experiments. Higher AUC indicates that we can significantly reduce false positive proposals resulting from YOLO. We also compute the inference time per image to compare computing time costs of each method (Tab.5.5). This includes testing all possible pairs. The inference time of the CMCNet varies from 2.7 (VGG16) to 25.4ms (EfficientNet-B3). Since no significant improvement arises when using deeper models, models with low time complexity (VGG, ResNet) are more appropriate. Multi-tasking methods do not cost more time than classification-only schemes. Computing time increases significantly with model complexity, whereas no significant improvement arises. The time increase with respect to the YOLO detector (78.7ms per image) is almost negligible.

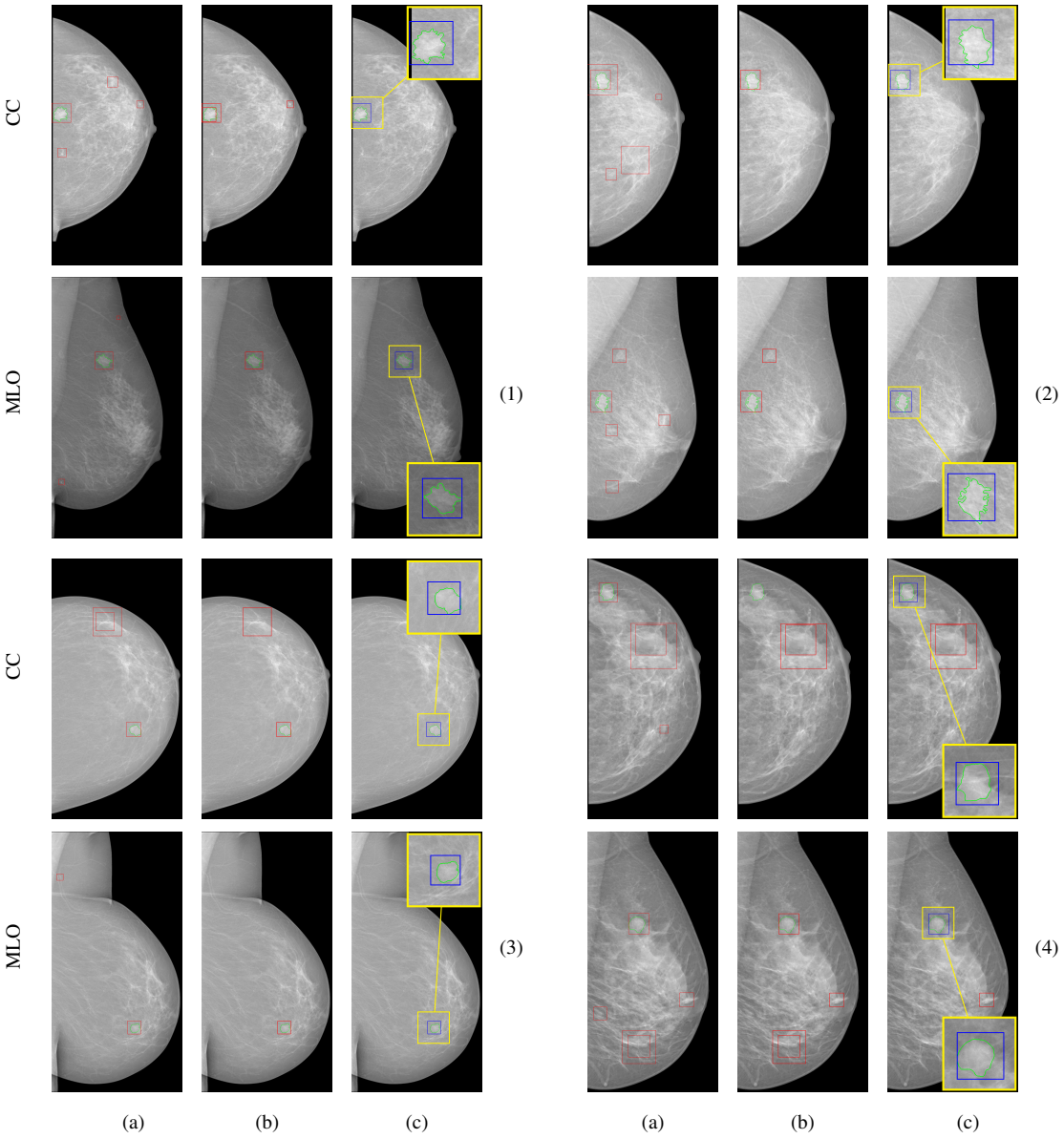


Figure 5.5 – Full-pipeline mass detection: (a) YOLO detection only, (b) YOLO followed by a classification-only model, (c) YOLO followed by the proposed CMCNet (with VGG16 and contrastive loss). Red and blue boxes are detected mass bounding boxes. Green labels represent ground truth annotations. Blue boxes show the matching pair selected through dual-view matching. Visual examples are labeled from (1) to (4).

Network	matching	matching loss	overall acc	AUC (%)	AUC p-value	runtime (ms)
VGG16 (Simonyan & Zisserman, 2014)	×	-	0.8260	90.47	-	2.7
	√	cross-entropy	0.8761	94.17	$2e^{-5}$	2.7
	√	contrastive	**0.8791**	**94.78**	0.001	2.7
Resnet50 (He et al., 2016b)	×	-	0.6814	70.03	-	8.4
	√	cross-entropy	**0.8555**	**91.98**	$< 1e^{-6}$	8.3
	√	contrastive	0.8496	90.30	$< 1e^{-6}$	8.4
Resnet101 (He et al., 2016b)	×	-	0.7080	71.46	-	16.2
	√	cross-entropy	0.8555	91.74	$< 1e^{-6}$	15.8
	√	contrastive	**0.8584**	**92.82**	$< 1e^{-6}$	16.3
InceptionV3 (Szegedy et al., 2016)	×	-	0.8112	89.75	-	17.2
	√	cross-entropy	0.8201	89.86	0.9142	16.7
	√	contrastive	**0.8702**	**93.61**	0.009	16.8
EfficientNet-B3 (Tan & Le, 2019)	×	-	0.8142	87.97	-	25.2
	√	cross-entropy	0.8378	**89.80**	0.1795	25.4
	√	contrastive	**0.8466**	88.91	0.5735	24.7

Table 5.5 – Full detection pipeline results including overall classification accuracy (acc), AUC scores, statistical significance p-values of AUC scores with respect to the classification-only baseline, as well as inference times per image. Best results per network are in bold.

Additionally, we provide mass matching performances during inference. As shown in Tab.5.6, mass matching performance is measured using accuracy (*acc*) and AUC. We compare mass matching using our multi-task learning (with cross-entropy and contrastive losses) versus matching-only. The matching-only scheme refers to the matching Siamese network illustrated in Fig.5.2. Results show that the proposed multi-task learning brings gain from 0.62% to 8.64% in AUC and from 0.9% to 11.12% in acc. Best results are achieved by ResNet50 (AUC = 94.30%, acc = 89.49). On the basis of the experimental results, we can draw the conclusion that not only matching can improve classification, classification can also improve matching, proving that the multi-tasking properties and the multi-view learning can help towards better breast cancer diagnosis and management.

Using the INbreast dataset, we also compare the overall mass detection performance using the true positive rate (TPR) at the average false positive per image (FPavg) with state-of-the-art methods (Tab.5.6). Since there is no official split of INbreast, each study has its own split between training, testing and validation subsets. Results shown in the top part of Tab.5.6 give an idea of the overall detection performance without giving a relevant comparison with these studies. For a fair comparison with the state-of-the-art, we re-implemented the recently published method of Agarwal et al. (2019) and conducted experiments using the same data as used in our work (80% DDSM-CBIS for training, 20% DDSM-CBIS for validation and 70 INbreast images for testing) to obtain the Free Response Operating Characteristic (FROC) curve of final detections. The bottom part of Tab.5.7 includes results obtained on the same testing data. In particular, it displays the best TPR@FPavg score achieved using Agarwal et al. (2019): 0.74@0.99. The best TPR@FPavg score (0.96@0.23) is reached by the proposed framework (CMCNet with VGG16 and contrastive loss). It outperforms the classification-only model (0.89@0.29) and shows consistent performance with respect to existing approaches such as Yan et al. (2021d) and Dhungel et al.

Network	matching	classification	matching loss	matching AUC(%)	matching acc
Perek et al. (Perek et al., 2018)	√	×	cross-entropy	79.92	0.7504
VGG16 (Simonyan & Zisserman, 2014)	√	×	cross-entropy	91.05	0.8523
	√	√	cross-entropy	91.49	0.8671
	√	√	contrastive	**92.97**	**0.8714**
ResNet50 (He et al., 2016b)	√	×	cross-entropy	92.77	0.8693
	√	√	cross-entropy	92.46	0.8775
	√	√	contrastive	**94.30**	**0.8949**
ResNet101 (He et al., 2016b)	√	×	cross-entropy	90.01	0.8345
	√	√	cross-entropy	**92.31**	0.8716
	√	√	contrastive	91.72	**0.8758**
InceptionV3 (Szegedy et al., 2016)	√	×	cross-entropy	89.50	0.8405
	√	√	cross-entropy	**90.64**	0.8536
	√	√	contrastive	90.01	**0.8588**
EfficientNet-B3 (Tan & Le, 2019)	√	×	cross-entropy	82.99	0.7391
	√	√	cross-entropy	89.50	0.8379
	√	√	contrastive	**91.63**	**0.8503**

Table 5.6 – Mass matching AUC with the proposed CMCNet model (with cross-entropy and contrastive losses) versus matching-only schemes including (Perek et al., 2018). Best results per network are in bold.

Methods	TPR@FPavg	dataset (INbreast)
Kozegar et al. (2013)	0.87 @ 3.67	107
Akselrod-Ballin et al. (2017)	0.93 @ 0.56	100
Ribli et al. (2018)	0.90 @ 0.3	107
Dhungel et al. (2017a)	0.95 @ 5	410
Agarwal et al. (2019)	0.92 @ 0.5	107
Yan et al. (2021d)	0.94 @ 0.22	107
Agarwal et al. (2019)	0.74 @ 0.99	70 (inference only)
YOLOv3 only	0.86 @ 1.41	70 (inference only)
classification-only VGG16	0.89 @ 0.29	70 (inference only)
CMCNet VGG16 (ours)	0.96 @ 0.23	70 (inference only)

Table 5.7 – Final detection performance comparisons on INbreast (Moreira et al., 2012) between the proposed method (CMCNet with VGG16 and contrastive loss) and state-of-the-art approaches.

(2017a) obtaining respectively 0.94@0.22 and 0.95@5, while additionally providing accurate dual-view mass correspondences.

Evaluation is supplemented with qualitative results on full mammograms (Fig.5.4). The additional classification stage (b) helps in eliminating most of false YOLO detections (a). The improvement reached by the combined model (c) compared to the classification-only scheme (b) is highlighted with further wrong proposal removals. For instance, in Fig.5.5 (2), the number of false positive detections decreased from 7 to 1 from (a) to (b) and further decreased to 0 without any false negatives. In addition, the CMCNet (c) also successfully identifies the matching patches in both views, which can provide clinicians with reference to further rule out false positives that are difficult to detect, as in Fig.5.5 (4). Fig.5.5 also demonstrates that variable mass sizes and shapes can be correctly managed. All these findings suggest that exploiting multi-view relationships and multi-tasking learning can greatly guide mammogram interpretation, towards better breast cancer diagnosis and management.

Closer to clinical screening conditions. To evaluate the incidence of false positives under closer to clinical breast screening conditions, a test set of normal mammograms (without masses) has been considered to evaluate our method. Among the 410 INbreast mammograms, 60 CC/MLO image pairs that contain no mass were found. Coarse mass detections are firstly performed on these normal images to generate candidate patches. We use YOLOv3 pre-trained on ImageNet and fine-tuned on 1514 DDSM-CBIS images. Then, a small threshold (10^{-3}) is applied on detection probabilities to ensure that enough predictions from YOLOv3 are selected. We finally obtain 646 candidate patches (374 from CC view, 410 from MLO view), labeled as negative. Then, candidate patches from two views of the same patient are given as inputs of the two branches of our Siamese model for mass/non-mass classification and matching.

We finally obtained 56 false positive predictions whereas 590 true negatives were detected. Accordingly, the obtained specificity was 0.9133. Concerning the 56 false positive detections, only 6 pairs were considered as matched pairs. Thus, the other 44 patches can be further eliminated because there is no corresponding detection in the other view. These results confirm that our contributions can provide reliable detection results in a setup more similar to a screening process which could be conducted in real life.

5.2.4 Conclusion

To conclude, we proposed a novel multi-tasking approach that combines breast mass/non-mass classification with dual-view mass matching between complementary CC/MLO mammograms. We prove the effectiveness of integrating multi-view information within the breast mass detection pipeline by extensive experiments on public datasets. Based on Siamese networks and contrastive learning, our method generalizes well using different deep networks and shows impressive results as an integrated CAD system. We can thus easily address the problem of false detections without struggling with difficult whole-image detection schemes. We also extend this framework by associating detection and matching with segmentation techniques to further guide clinicians in multi-task interpretations. More globally, the proposed contributions pave the way for robust automatic second opinions in breast cancer diagnosis.

Even if multiple masses can still be detected using the classification network, dealing with more than one mass with respect to matching purposes should deserve further investigation. Furthermore, the integration of the segmentation task still needs further study to deal with the limitation of negative patches, so as to obtain a more complete and complex multi-tasking CAD system. Last but not least, it is essential to push further data fusion by extracting and integrating both multi-view and longitudinal information.

In the following section, we go further with the multi-view information fusion by exploiting the dual-view consistency as criterion for a novel deep active learning approach which addresses the common lack of labeled data, thereby reducing the labeling work of clinicians.

5.3 Deep active learning for dual-view mammogram analysis

Based on supervised learning using convolutional neural networks (CNN), recent studies have achieved impressive performance regarding mass segmentation (Singh et al., 2020; Yan et al., 2019b; Yan et al., 2021d) or detection (Agarwal et al., 2019; Dhungel et al., 2017a; Kooi et al., 2017; Ribli et al., 2018; Yan et al., 2021d). Despite such success, supervised deep learning still faces obstacles, including data acquisition and high-quality manual annotations which are expertise-needed and time-consuming. The current rise of deep learning made the analysis of mammograms more automatic and accurate thanks to effective training methods, advances in hardware, and most importantly, large amounts of annotated training data (Kooi et al., 2017). Computational analysis of dual-view mammograms (Gu et al., 2018; Ma et al., 2019; Perek et al., 2018; Yan et al., 2020) has been validated as an effective way to reduce false-positive cases and improves screening performance. Nevertheless, the labeling workload of radiologists is further increased. Therefore, it is greatly needed to develop an effective annotation suggestion algorithm to alleviate this issue. In this section, we propose a novel approach of deep active learning (AL) for dual-view mammogram analysis including breast mass segmentation and detection, where the dual-view prediction consistency is integrated as selection criterion. Second, two task-specific neural networks are carefully designed for more effective mammogram mass segmentation and detection. Third, extensive experiments are conducted to reveal the relationship between dual-view consistency and mammogram informativeness.

5.3.1 Active learning

Extensively studied in various fields including language processing, anomaly detection or recommendation systems, active learning (AL) aims at reducing human annotation efforts by adaptively selecting the most informative samples for labeling. As for medical imaging, AL has shown high potential in reducing the annotation cost (Budd et al., 2019).

Recent studies (H. Li & Yin, 2020; H. Shen et al., 2020) propose AL frameworks for breast cancer segmentation, respectively on immunohistochemistry and biomedical images. However, AL methods have not been widely exploited in X-ray mammography analysis. Zhao et al. (2019) first introduced AL into a mammography classification system based on a support vector machine (SVM) classifier. R. Shen et al. (2019) proposed a mass detection framework that incorporates AL and self-paced learning (SPL) to improve the model generalization ability. These studies demonstrate the great potential of AL in mammogram analysis.

Muslea et al. (2006) first proposed a co-testing algorithm for active learning in a multi-view setting. They focused on the unlabeled examples on which different views predict different labels, then these examples are picked for human annotation. W. Wang and Zhou (2008) first theoretically analyzed the sample complexity of multi-view active learning using the same paradigm as Muslea et al. (2006) and proved the exponential improvement. Different from existing studies that focus on querying the class

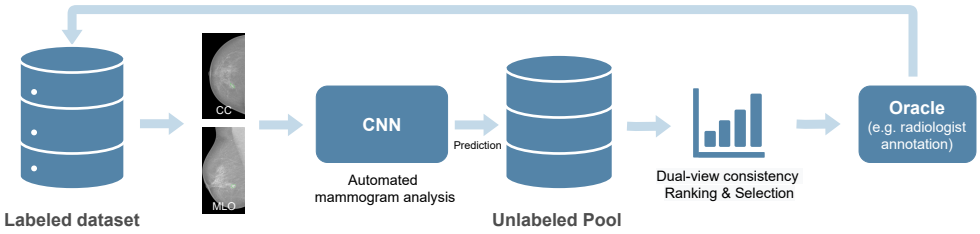


Figure 5.6 – Proposed deep active learning workflow.

labels, Cai et al. (2019) proposed a novel multi-view active learning framework which actively queries the missing view of selected examples for video recommendation. However, the existing active learning approaches have rarely focused on exploiting the multi-view properties of medical imaging. Contrary to existing studies that are based on uncertainties and diversities of a single view of mammogram, our goal is to score the dual-view mammograms according to their prediction consistency. Our work can be seen as a complement to existing methods and proves that combining inter-view information can bring further improvements.

5.3.2 Network architectures for mass segmentation and detection

To reduce the labeling efforts dealing with breast masses in mammograms, we propose a novel approach of deep active learning for dual-view mammogram analysis. Specifically, we consider two scenarios: mass detection and segmentation. The key insight of our method is to use the consistency of mass detection or segmentation results arising from CC/MLO view-points as active learning criterion.

The proposed AL process starts by pre-training the model on a small labeled subset D_l. Then, we perform model inference on the unlabeled dataset D_u to select the most informative mammogram pairs according to the calculated dual-view prediction consistency. These selected pairs are then sent to an oracle (i.e. the radiologists) for annotation and appended to D_l, where the model is consequently fine-tuned on. Such AL cycle (Fig. 5.6) is repeated several times to gradually improve the model performance, until the annotation budget is exhausted. The key feature of AL is the query algorithm for the informativeness ranking of unlabeled images, which in our work is the scoring function related to the dual-view prediction consistency.

Breast mass segmentation and detection are two main tasks in mammogram analysis. We take inspiration from recent advances of deep neural networks (He et al., 2016a; T.-Y. Lin et al., 2017b; Ronneberger et al., 2015b) and design simple and efficient networks for each of these tasks (Fig.5.7).

5.3.2.1 Mass segmentation network

The architecture is composed of an encoder network for feature extraction, a decoder network for spatial detail reconstruction, and several skip-connections between both branches to recover spatial

information. Instead of using a standard symmetric encoder-decoder architecture (Conze et al., 2021; Ronneberger et al., 2015b), we apply an alternative asymmetric architecture where residual blocks are integrated into the encoder and 1×1 convolution layers are part of the decoder (Fig.5.7). This design enables the encoder to extract features from inputs, while the decoder maintains the same performance of recovering context information, while the network complexity is greatly reduced. The optimization is supervised by the combination of binary cross-entropy (L_{bce}) and Dice (L_{dice}) losses following $L_{seg} = L_{dice} + \lambda_1 L_{bce}$ (Eq. 5.3, Eq. 5.4), where the empirical factor λ_1 is set to 0.5 to prevent the combined loss from degenerating into L_{bce}.

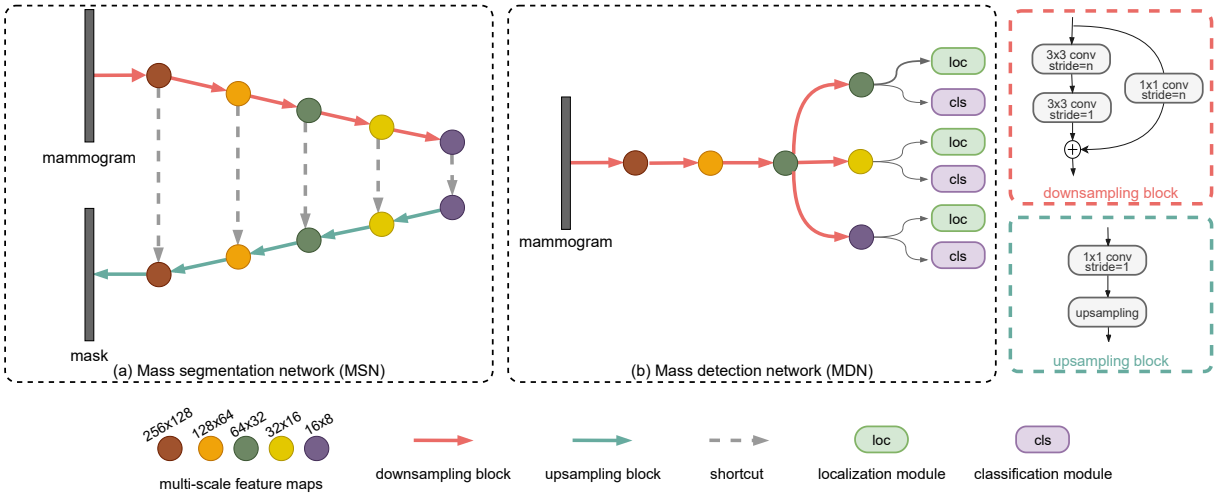


Figure 5.7 – Proposed network architectures for mass segmentation (a) and detection (b). A downsampling (upsampling) block is applied in each red (green) arrow.

5.3.2.2 Mass detection network

We designed a single-stage mass detection network, where a multi-scale prediction strategy is applied to detect masses of different scales. Three detection branches with different scales $\{64\times32, 32\times16, 16\times8\}$ are attached to a regular feature extraction network (Fig.5.7(b)) consisting of 3 residual blocks. Larger scale detection branch aims at detecting smaller mass and vise-versa. Each branch consists of a localization module and a classification module, where the former is in charge of regressing the spatial transformation (4 coordinates offset) from predefined anchor boxes to ground truth boxes, and the latter predicts the probability of mass presence for each anchor box. Following Pang et al. (2019), the classification and localization tasks are simultaneously under the guidance of a multi-task loss to ensure a better detection performance, which is the combination of the focal loss (L_{focal}) to supervise classification modules and the balanced L1 loss (L_{bl1}) to supervise localization modules. The combined loss is thus defined as:

$$L_{det} = L_{focal} + \lambda_2 L_{bl1} \tag{5.6}$$

$$L_{focal} = \begin{cases} -\alpha_1(1-p)^{\gamma_1}\log(p) & \text{if } y = 1 \\ -(1-\alpha_1)p^{\gamma_1}\log(1-p) & \text{otherwise} \end{cases} \tag{5.7}$$

$$L_{bl1} = \begin{cases} \frac{\alpha_2}{\beta}(\beta|x|+1)\ln(\beta|x|+1) - \alpha_2|x| & \text{if } |x| < 1 \\ \gamma_2|x| + C & \text{otherwise} \end{cases} \tag{5.8}$$

We use the default parameters of L_{focal} and L_{bl1} as respectively introduced in T.-Y. Lin et al. (2017b) and Pang et al. (2019): $\alpha_1 = 0.25, \gamma_1 = 2.0$ for L_{focal}, $\alpha_2 = 0.5, \gamma_2 = 1.5, \beta = 1.0$ for L_{bl1}. The final detection loss is the combination of L_{focal} and L_{bl1} with $\lambda_2 = 1$.

5.3.3 Dual-view consistency

At the selection stage of each AL cycle, we aim at filtering the most informative mammograms in D_u through the analysis of dual-view consistency. Theoretically, given a pair of mammograms $\{I_{CC}, I_{MLO}\}$ from the same breast, the analysis results should be consistent. Many latent relationships can potentially be exploited as query factors, such as the number of masses detected on both views or the mass size, position, shape, texture... In our work, we consider the first two factors as consistency criteria since their correlation is more obvious. In particular, the number of identified masses from both views $\{N_{CC}, N_{MLO}\}$ should be identical and their sizes $\{S_{CC}, S_{MLO}\}$ (i.e. number of pixels) should be similar. We define two scores (S_{num} and S_{size})to measure the following factors:

$$S_{num} = \frac{min(N_{CC}, N_{MLO})}{max(N_{CC}, N_{MLO})}, S_{size} = \frac{min(S_{CC}, S_{MLO})}{max(S_{CC}, S_{MLO})} \tag{5.9}$$

where S_{num} and S_{size} varies from 0 (low consistency) and 1 (high consistency). Correct predictions should meet the above two conditions simultaneously, thus the final combined score is calculated as the minimum of S_{num} and S_{size}:

$$S = min(S_{num}, S_{size}) \tag{5.10}$$

The proposed consistency score S provides a rough estimation of the mass segmentation/detection prediction quality: mammogram pairs with higher S values are regarded as easy samples, while pairs with lower scores are considered as hard samples. Fig.5.8 shows mammogram pairs with different S values for both segmentation and detection tasks. When the S value is low, the prediction on at least one mammogram appears inaccurate.

5.3.4 Active learning strategies

The key of AL is to select the most informative samples to optimize a learnable model. However, the definition of informativeness is still an open question. In the common practice of AL, one considers

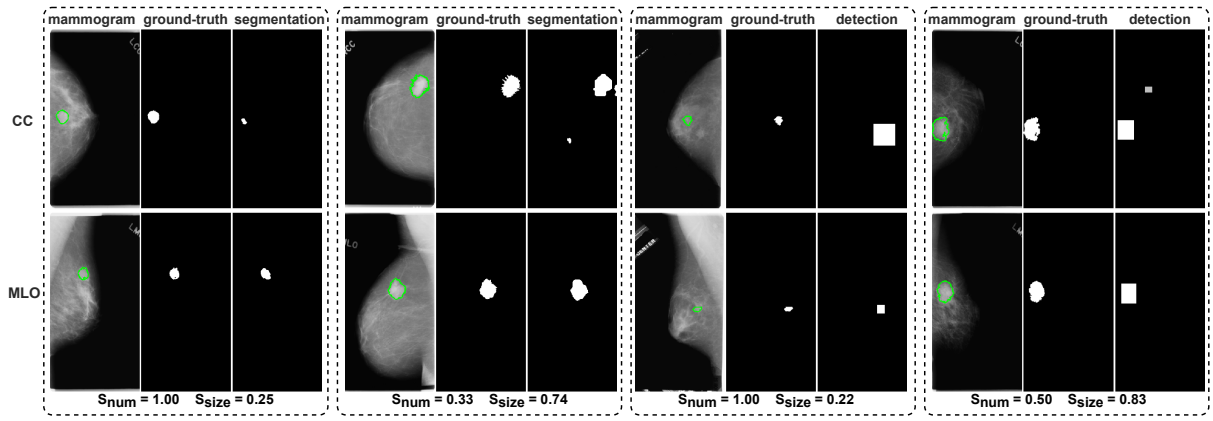


Figure 5.8 – Examples of mass segmentation (left half) and mass detection (right half) for CC/MLO pairs from DDSM-CBIS and corresponding dual-view consistency. S_{num} and S_{size} are respectively consistency scores of mass numbers and mass sizes, higher score for higher consistency. Green delineations represent ground truth mass annotations.

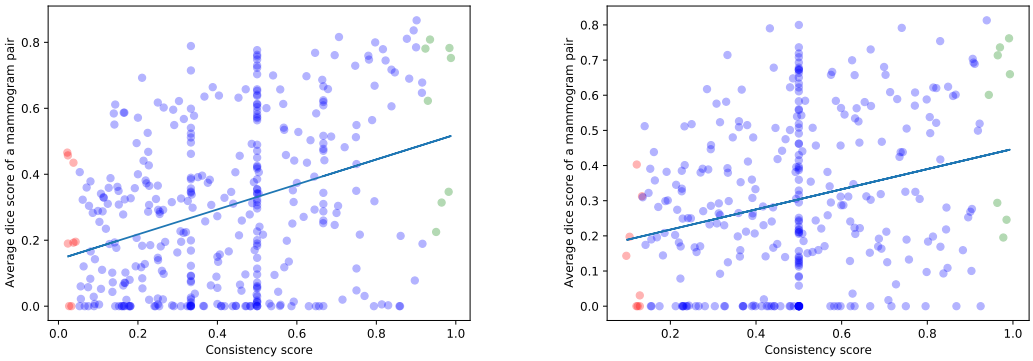
examples with the most uncertainty or examples that are most likely to be wrong as informative examples. However, we need to check if this paradigm remains valid in the field of medical imaging and especially for mammogram analysis. To this end, we implement three AL strategies: random (`rand`), best consistency (`bestC`) and worst consistency (`worstC`) selections. For each AL cycle, `rand` strategy randomly selects b mammogram pairs from unlabeled dataset D_u, while `bestC` (`worstC`) selects b pairs with the highest (lowest) consistency score S (Eq. 5.10).

We visualize in Fig.5.9 the mammogram pairs selected by different AL strategies. Specifically, each point represents a CC/MLO mammogram pair, and red (green) points are b pairs selected by `worstC` (`bestC`) strategy. We also estimate the linear regression between consistency score S with mass segmentation / detection accuracy. It can be observed that the consistency score is a reasonable reference of the prediction quality.

5.3.5 Experiments and results

5.3.5.1 Implementation details

We use two publicly-available datasets for our experiments: DDSM-CBIS (Lee et al., 2017) and INbreast (Moreira et al., 2012), with respectively 1514 and 107 cases containing ground truth mass delineations. Specially, DDSM-CBIS is employed as the training set and is divided into a small labeled subset D_l and a simulated unlabeled pool D_u for training AL cycles. Since we need the dual-view information consistency for AL, only 586 CC/MLO mammogram pairs from DDSM-CBIS are used. For INbreast, all 107 images are employed as the test set since pair-wise data is not mandatory during inference. The original mammograms have a resolution of 4084×3328 or 3328×2560, which is computationally expensive. Therefore, we resize input images to 512×256 for all experiments. Mammograms are normalized



(a) Mass segmentation (b) Mass detection

Figure 5.9 – Visualization of mammogram pairs selected by different AL strategies for mammogram segmentation (a) and detection (b) tasks by plotting the average dice score of a mammogram pair against the consistency score. Here, the dice scores for mass segmentation (detection) are calculated between the predicted masks (bounding boxes) with respect to the ground truth masks. Red (green) points are picked by `worstC` (`bestC`) strategy. The straight line estimates the linear regression.

according to the dataset mean and standard deviation before feeding into neural networks. Several data augmentation strategies are applied during the training phase, including random image rotation, cropping, padding and flipping operations.

The proposed framework was implemented using PyTorch. We use a SGD optimizer with a learning rate of 0.1 combined with a cosine annealing schedule. The proposed MSN (MDN) has 45,705 (80,202) learnable parameters in total. With a batch size set to 32 and an input image resolution of 512×256, the training process takes 2,217 (2,199) MiB of GPU memory for the segmentation (detection) task. Each experiment is repeated 5 times using randomly initialized labeled subset D_l, and we report their average performance and the standard error. We adopt the Dice coefficient to evaluate the segmentation performance and the Average Precision (AP) score to evaluate the detection performance. Dice coefficient is defined as 1-L_{dice} (Eq.5.3). AP score is calculated by taking the area under the precision-recall curve.

For each AL experiment, we start by training an initial model on a random labeled subset D_l containing b pairs. During each AL cycle, we adaptively select the next b pairs from DDSM-CBIS using three different AL strategies (`rand`, `bestC` or `worstC`) from the unlabeled dataset D_u. These images are assigned with annotations and appended to D_l for fine-tuning at the next AL cycle. We fix an annotation budget B to end AL cycles. Concretely, we set b to 8 (16 images) for all experiments. Noting that the annotation cost for segmentation is much higher than for detection, we set B to 40 (80 images) for the mass segmentation task and 56 (112 images) for the detection task. In other words, we implement 4 (6) active cycles for segmentation (detection). Each cycle for segmentation (detection) adds 1.37% of labeled data, and the whole segmentation (detection) AL process takes 6.83% (9.56%) of labeled data in the training set.

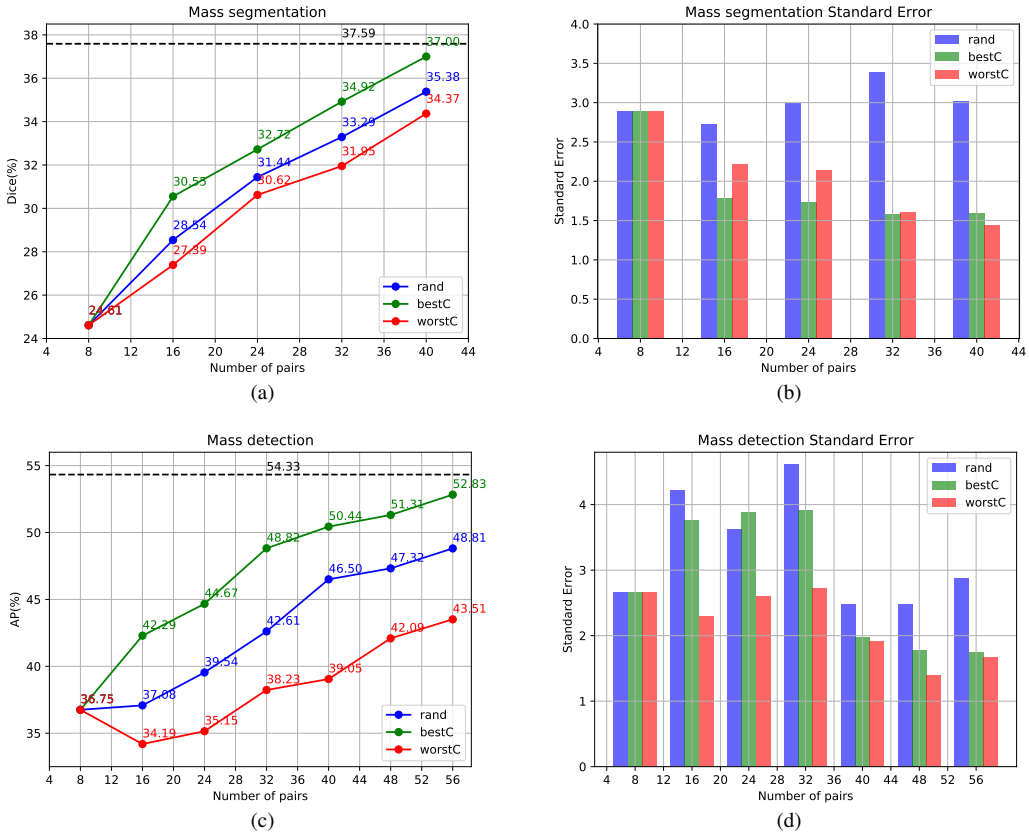


Figure 5.10 – Mass segmentation and detection performance with `rand` (blue), `bestC` (green) and `worstC` (red) AL strategies. Black dashed lines indicate results using the complete training set. We report average dice score of mass segmentation (a), dice score standard error (b), average AP score of mass detection (c) and AP standard error (d).

5.3.5.2 Results

We conducted extensive experiments to evaluate the performance of `rand`, `bestC` and `worstC` AL strategies. Averaged results are shown in Fig.5.10. It can be seen that the model performance is improved progressively cycle by cycle, and that `bestC` (dice=37.00%, AP=52.83%) is consistently better than the other strategies. `bestC` presents 1.62% dice improvement and 4.02% AP gains with respect to the `rand` baseline. Conversely, `worstC` (dice=34.37%, AP=43.51%) does not outperform the baseline. From Fig.5.10 (b) and (d) we observe that the standard errors of `rand` for dice and AP remain both relatively high, while both `bestC` and `worstC` reduce the performance instability of `rand` strategy to a certain extent. In particular, with only 6.83% (9.56%) labeling budget for mass segmentation (detection), `bestC` achieves comparable performance with respect to fully-supervised models (37.00 vs 37.59% for segmentation, 52.83 vs 54.33% for detection), showing the great potential of our method in alleviating the annotation burden. Besides, we observe greater performance gaps for detection than segmentation.

Since detection annotations only provide sparse box-level supervision, the detection task is more critical in terms of the amount of training images.

In the common practice of traditional AL, examples with high consistency scores provide better prediction quality, and could be seen as well-learned examples, which are normally not included into AL cycles. Our results seem to contradict this common practice since pairs with higher consistency seem more useful than those with lower consistency. For this finding, we propose some explanations: mammography analysis is actually more difficult than general natural image analysis tasks since it is difficult for humans without clinical knowledge to distinguish masses from surrounding healthy tissues. Medical imaging datasets can also be very biased due to different acquisition conditions. Learning with a small amount of medical images is challenging, especially for the first few AL cycles. For detection, Fig.5.10 (c) shows an AP drop for the first AL cycle of `worstC`, indicating that not all labeled data are beneficial when the model does not yet have a full understanding of what masses are. Picking examples with good prediction results helps to consolidate what has been learned while avoiding corner cases.

5.3.6 Discussion

We propose a label-efficient deep learning approach that explores the prediction consistency arising from dual-view mammograms. The main novelty is the combination between multi-view mammogram analysis and active learning, which has not been studied in the field of medical imaging to our knowledge. Our contributions significantly alleviate the burden of manual labeling in breast mass segmentation and detection tasks, which is beneficial to the development of CAD tools. As a future work, more complex query factors of the multi-view consistency can potentially be exploited. Another future possible extension to this work is to integrate existing single-view criteria into our current framework, towards a unified active learning system.

5.4 Conclusion

In clinical routine, radiologists usually confirm the diagnosis through cross information arising from both views. In this chapter, we studied the potential of multi-view information fusion to the improvement of CAD systems. We investigated two different perspectives by presenting (1) a multi-view multi-tasking framework that improves breast mass detection by exploiting the dual-view mass matching and (2) a label-efficient deep active learning approach that explores the dual-view mammogram consistency. Extensive experiments of both studies reveal great effectiveness and promising robustness of multi-view information fusion in various aspects of mammogram analysis, including mass detection, classification, segmentation, matching and data labeling problems. The great potential of information fusion in the field of medical imaging is greatly highlighted through these studies. In the following chapter, we will bring this topic to another extent: longitudinal information fusion for the prediction of severity grade changes.

CHAPTER 6

LONGITUDINAL PREDICTION OF SEVERITY GRADE CHANGES

Contents

6.1 Introduction

Diabetic retinopathy (DR) recognition has been an active research area over the last few decades and has been exploited in many aspects over the years. Early detection and adapted treatment, especially in the mild to moderate stage of non-proliferative DR (NPDR), helps to slow down the progression of DR, thereby preventing the occurrence of diabetes-related visual impairment and blindness.

Recently, deep learning (DL) has been widely adopted in various tasks of retinal image analysis. For the prediction of the severity of lesions, many studies have focused on the DR grading classification at image-level, as severity labels can be easily extracted from radiology reports. Gulshan et al. (2016)

applied InceptionV3 (Szegedy et al., 2016) architecture for automated detection of DR in retinal fundus photographs. Quellec et al. (2017) proposed a multiple-instance learning framework that is supervised using only image-level labels for both automatic prediction of DR scale and DR-related. They further developed an instant automatic diagnosis system of DR (Quellec et al., 2019) which incorporates multiple CNN models and targets three classification tasks: laterality identification, referable DR detection and DR severity assessment. Gharaibeh et al. (2018) proposed an effective and automatic screening system for DR detection through a series of processes: image pre-processing, optic disc detection and removal, blood vessel segmentation and removal, elimination of fovea, feature extraction, selection and classification. More recently, Shankar et al. (2020) applied a synergic deep learning (SDL) model incorporating histogram-based region-of-interest segmentation for DR classification. H. Liu et al. (2020) trained three hybrid models using an improved loss function to improve the performance of basic DR classification models, including EfficientNetB4, EfficientNetB5, NASNetLarge, Xception, and InceptionResNetV2. Sikder et al. (2021) dealt with DR severity classification from noisy retinal images using an ensemble learning technique named Extreme Gradient Boosting (XGBoost) based on the gray-level intensity and texture features extracted from fundus images. However, instead of more extensive DR grading classification, existing methods focus more on DR/non-DR detection or classification of high-level DR (severe NPDR or PDR), and only use a single study without considering previous studies.

Previous studies have demonstrated the potential of fusion methods in medical imaging, such as multi-view (Geras et al., 2017; Perek et al., 2018; Yan et al., 2020) or bilateral (Geras et al., 2017) fusion. In clinical routine, radiologists often compare the current screening to one or more prior studies from the same patient. Several studies attempted the automatic analysis of longitudinal medical images. Santeramo et al. (2018) analyzed the evolution of longitudinal chest X-rays using long short-term memory (LSTM) networks. Perek et al. (2019) developed and compared four deep learning fusion methods for longitudinal mammography studies: early fusion of input images, feature fusion incorporating CNN and gradient boosting trees, feature fusion based on LSTM cells and late fusion of prediction scores. However, to our knowledge, the number of existing deep learning methods that analyze longitudinal screenings to assist in DR classification is still limited. Narasimha-Iyer et al. (2007) presented methods for unified analysis of both vascular and non-vascular changes that are observed in longitudinal time-series of color fundus photographs (CFP). Bernardes et al. (2009) use a microaneurysm-tracker to evaluate DR progression in a follow-up study based on computer-assisted earmarking of microaneurysms. Adal et al. (2017) presented a robust and flexible multistage approach for tracking retinal changes due to small red DR lesions such as microaneurysms and dot hemorrhages in longitudinal fundus images. They measure the absolute difference between the extremes of the multiscale blobness responses of fundus images from two time points, then identify the DR related changes based on several intensity and shape features by a support vector machine classifier.

In this regard, we aim to integrate longitudinal information of CFP images to help detect the referable DR severity change. Specifically, we target the change detection between no DR/mild NPDR and more

code	DR severity grade
1	unknown
2	no DR
3	mild DR
4	moderate DR
5	severe DR
6	PDR
7	high-risk PDR

Table 6.1 – Codification of DR severity grade in OPHDIAT database.

severe DR through two consecutive longitudinal follow-ups of DR. To this end, we explore four image fusion methods that incorporate current and prior studies: (1) early fusion of input images; (2) intermediate fusion of feature layers using spatial transformer network (STN); (3) late fusion of feature vectors; (4) late fusion using a squeeze-and-excitation network (SENet) comprising an attention mechanism. We conduct a comprehensive evaluation of each fusion network on a large dataset from the OphDiaT telemedical network (Massin et al., 2008) for the comparison of performance. This work has been accepted in the 8th Ophthalmic Medical Image Analysis (OMIA8) MICCAI workshop (Yan et al., 2021e).

6.2 Data

The proposed models are trained and evaluated using the OPHDIAT dataset (Sect. 2.2.3). Tab.6.1 shows the codification of DR severity grade. In this work, we study the grade change from normal/mild NPDR (grade = 2 or 3) to more severe DR (grade $\geq$ 4). In Fig. 6.1, we show some examples of retina fundu images arising from the same patient, captured from left (L) and right (R) eyes and from different viewpoints from a series of times.

6.2.1 Data for longitudinal fusion

From the 763,848 images of 101,383 patients in the entire OPHDIAT database, we first select patients with up to two-year follow-up screenings and whose severity grade changes from grade = 2 or 3 to grade $\geq$ 4. Afterwards, to train our longitudinal fusion frameworks, the input image pairs $\{I_{t-1}, I_t\}$ should meet the following conditions: (1) arising from the same patient; (2) captured from the same viewpoint of the ipsilateral eye: (3) coming from two different screening times $\{t-1, t\}$. The prediction of severity grade change will be the binary output.

Image pair selection. Image pairing and registration are fundamental pre-processing steps for longitudinal analysis (Saha et al., 2019). In order to avoid the influence of position shifts, image scales or other factors related to the heterogeneity of retinal images from the OPHDIAT dataset, we first need to select image pairs captured from almost the same viewpoint from two consecutive images series $\{E_{t-1}, E_t\}$.

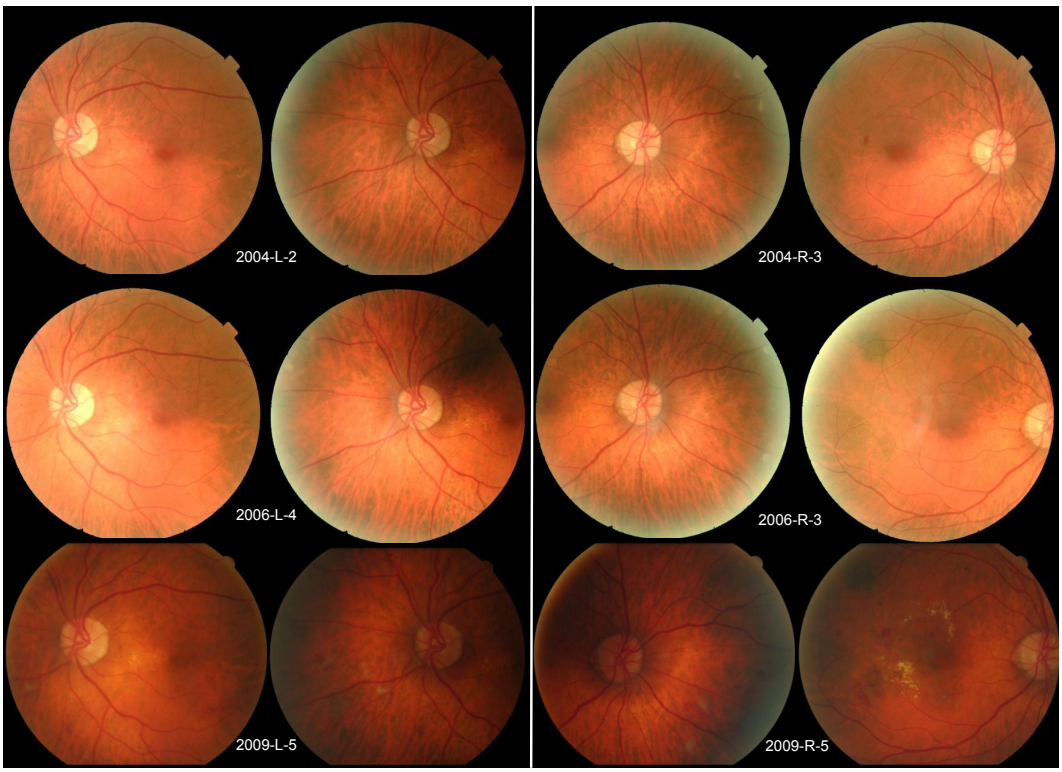
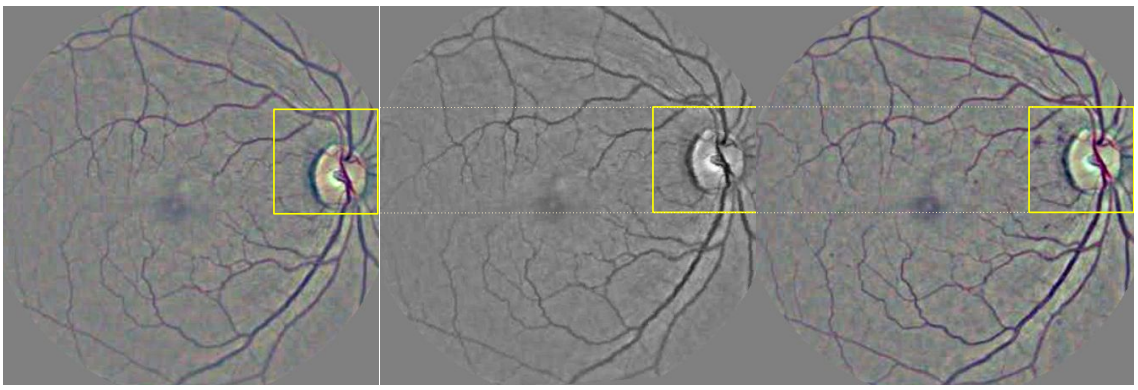


Figure 6.1 – Examples of retina fundu images arising from the same patient, captured from left (L) and right (R) eyes, from different viewpoints from a series of times. The notations in the figure indicate *screening year-laterality-severity grade*.

The calculation process is as follows: for each image I from E_t and each image J from E_{t-1}, we use an affine transformation to align J to I and obtain J_{warp} (Fig. 6.2). This transformation could not be done inversely (i.e. from I to J) because lesions may appear between time t-1 (image J) and time t (image I), such as the example illustrated in Fig. 6.2. Instead of warping images with underlying lesions, we hope to preserve as much original characteristics of lesions as possible, so that allowing the network to focus more on lesion areas. Then, we calculate a mean square error (MSE) between $\{I, J_{warp}\}$ (i.e. the sum of the squared difference between images I and J_{warp}). The image J that minimizes MSE (I, J_{warp}) is considered as a correspondence of image I. The image pairing is necessary for all proposed fusion schemes, while in particular, only the early fusion scheme requires the registered image J_{warp} as the input, as the registration allows the network to focus more on the tissue modification area, where is likely to be lesions.

Following the above process, we finally obtained 25,843 pairs of images of 2668 patients as data for longitudinal fusion purposes. This dataset is further randomly divided into a training set (60%), a validation set (20%) and a test set (20%). The number of pairs with and without grade change of each subset is shown in Tab. 6.2.



J J_{warp} I

Figure 6.2 – Example of image registration between image at time t-1 (J) and image at time t (I). An affine transformation is applied to align J to I to obtain J_{warp}.

subset	change	non-change	total pairs
train (60%)	4839	10666	15505
validation (20%)	1613	3556	5169
test (20%)	1626	3543	5169

Table 6.2 – Distribution of pairs with change/non-change in each subset.

6.2.2 Data for pre-training

Among the 101,383 patients from the OPHDIAT database, about 70% have no follow-up. Moreover, the proportion of normal cases exceeds 79%. This means that most of the data are not used in the longitudinal study. Nevertheless, the remaining data can be used for pre-training purposes. The effectiveness of pre-training on a large dataset and then fine-tuning on a specific small subset has been widely demonstrated. Accordingly, we finally use 649,365 images for pre-training by excluding images with grade = 1 (status unknown), without annotations, and the images used for the longitudinal study (Sect. 6.2.1). We randomly choose 80% as training set and 20% as validation set.

6.3 Methods

In this section, we first introduce in Sect. 6.3.1 the DL models on which the proposed methods are based. Then, we present three pre-training strategies in Sect. 6.3.2. Finally the proposed longitudinal fusion schemes for early-grade DR severity change detection are described in detail in Sect. 6.3.3.

6.3.1 Deep learning models

Two backbone networks are investigated in this study: VGG16 (Simonyan & Zisserman, 2014) and InceptionV4 (Szegedy et al., 2017). These DL architectures have been proven effective in various image

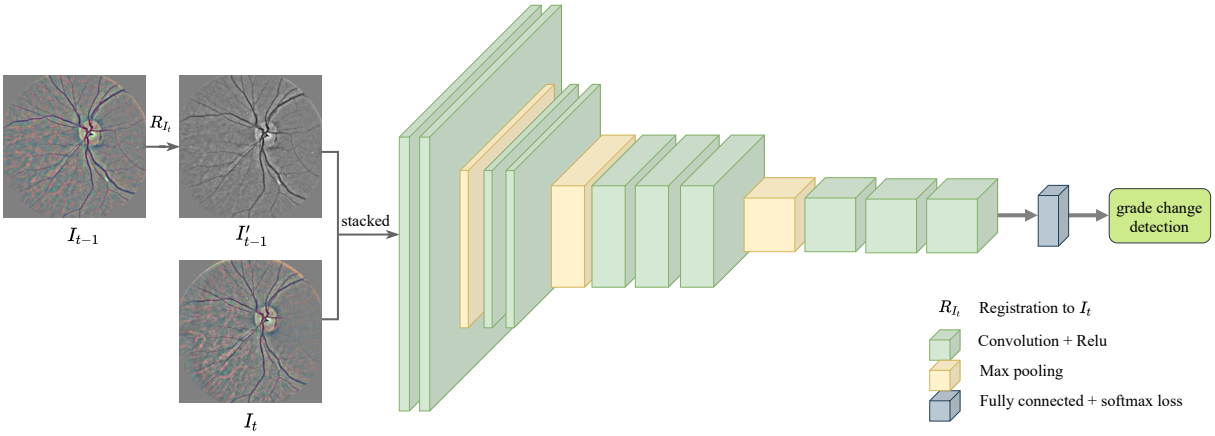


Figure 6.3 – Early fusion network architecture

recognition tasks. Note that in diagrams, we use a simplified CNN architecture for sake of clarity.

6.3.2 Pre-training strategies

Three pre-training strategies are proposed:

(1) **ImageNet**: using pre-trained weights from the ImageNet dataset (Russakovsky et al., 2015).

(2) **K-label classification model**: based on (1), training a K-label classification model trained with cross-entropy loss. The output of the softmax layer is K scores of DR grade (Tab. 6.1). We set K = 5 representing five classes: grade = 2, 3, 4, 5 and grade $\geq$ 6.

(3) **K-logistic multi-classifier model**: based on (1), training a K-logistic multi-classifier model trained with BCEWithLogits [9] loss. In this setting, K = 4 represents four binary classifiers, which respectively correspond to grade $\geq$ 3, grade $\geq$ 4, grade $\geq$ 5 and grade $\geq$ 6.

6.3.3 Longitudinal fusion schemes

6.3.3.1 Early fusion

As shown in Fig. 6.3, given a pair of consecutive images $\{I_{t-1}, I_t\}$, we firstly perform registration from I_{t-1} to I_t using affine transformation and obtain I'_{t-1}. Afterwards, we concatenate $\{I'_{t-1}, I_t\}$ as an input tensor with a dimension of 6. Accordingly, the first convolutional layer of different models are adjusted, while the other layers remain unchanged with respect to standard VGG16 or InceptionV4 architectures. The output of the network is the confidence score of whether there is a grade change between timepoints $t-1$ and t.

9. https://pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html

6.3.3.2 Intermediate fusion

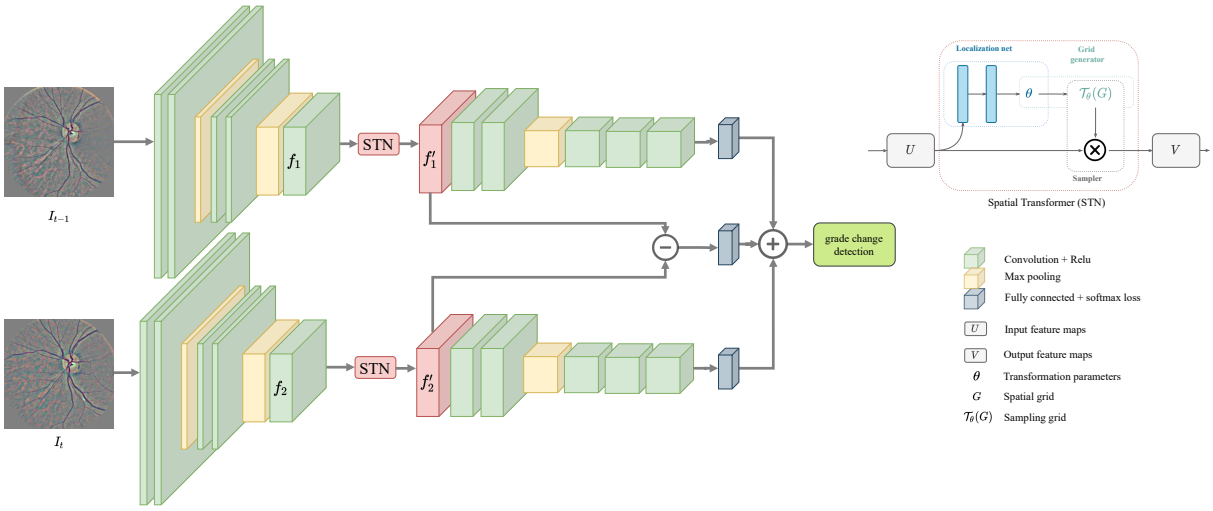


Figure 6.4 – Intermediate fusion network architecture

Different from early fusion, the intermediate fusion strategy, also known as feature-level fusion, implements fusion at convolutional stage, i.e. at an intermediate level within the model. Fig. 6.4 shows the architecture of intermediate fusion. We employ a Siamese network combined with a specific fusion module, a spatial transformer network (STN), that actively transforms feature maps without any extra supervision.

Spatial transformer network (STN). The STN module was first proposed by Jaderberg et al. (2015). Traditional CNNs are not invariant to spatial transformations (large-scale translation, scale, rotation, warping etc.) The STN module can be inserted into existing convolutional architectures, resulting in better model invariance to spatial translation, rotation and scaling. An STN network consists of three parts: (1) a localization network to calculate parameters θ of the spatial transformation of the feature map U; (2) a grid generator to create a parameterized sampling grid $\mathcal{T}_\theta(G)$; (3) a differentiable image sampler to produce the output feature map V based on the estimated spatial transformation.Ordinarily, STNs are used to modify the feature volumes of a one-stream network. In our case, we adaptively modify the use of STN in order to adjust the feature maps from the two branches of the Siamese network.

Normally, a Siamese network includes two identical branches with shared weights, in such a way that features from two different input images are extracted simultaneously and trained jointly. Instead, in our proposed intermediate fusion scheme, two branches are trained without sharing weights because if the fusion operation is desired to be done in the STN modules, the two branches are not supposed to be commutative. As shown in Fig. 6.4, we provide the Siamese network with non-registered image pairs $\{I_{t-1}, I_t\}$. Each image is processed independently up to a given convolutional layer L_{break} (before dense layers), outputting two sets of feature volumes f_1 and f_2. Then, we add two independent STN

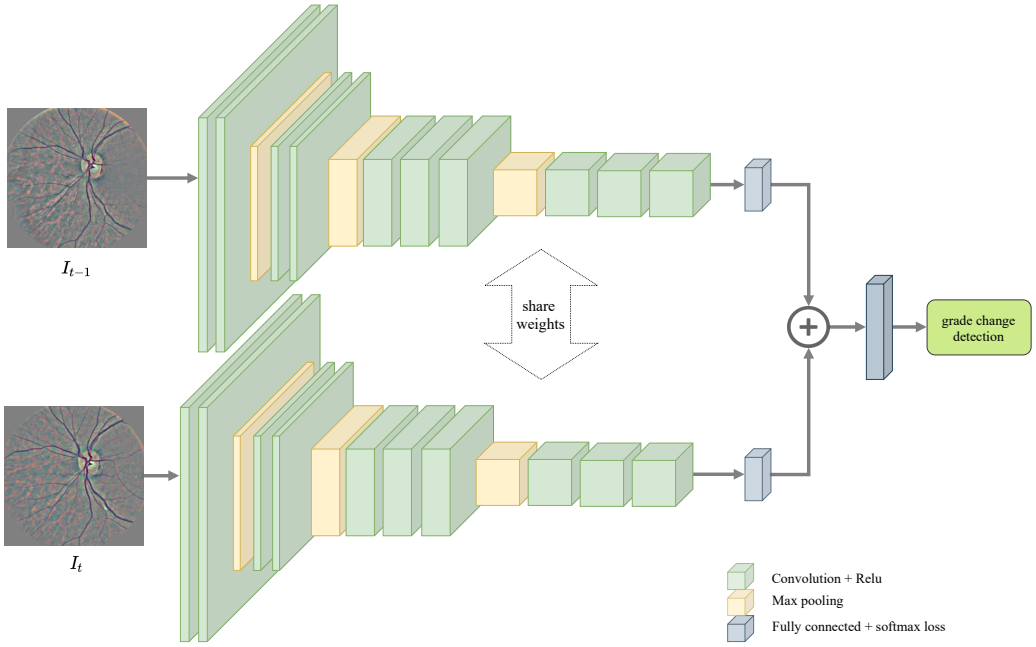


Figure 6.5 – Late fusion of feature vectors

modules after the given convolutional layer L_{break} to obtain two sets of transformed feature volumes f_1' and f_2'. Specifically, the STN modules are inserted before the 8th convolutional layer for VGG16 and before the first Inception-C module for InceptionV4. Thereafter, a fusion operation is applied to the two sets of feature maps, which can allow the back-propagation of loss which minimizes the difference in feature maps. Hence, a mean square error loss (MSELoss) is used as the fusion operator. Finally, the transformed feature maps f_1' and f_2' are processed with the remaining layers of the network. In practice, we use cross-entropy loss to optimize both branches of the Siamese network. The final loss function is defined as: $L_{interfusion} = L_{bce} + \lambda \cdot L_{MSE}$, where $\lambda = 100$ to balance the loss terms.

6.3.3.3 Late fusion

Similar to the intermediate fusion scheme, the late fusion is also performed using a Siamese network (Fig. 6.5). We provide the Siamese network with non-registered image pairs $\{I_{t-1}, I_t\}$. The reason why we do not need a registration of $\{I_{t-1}, I_t\}$ is that the fusion operation is in the feature vector level, which is invariant to spatial transformation of inputs. Two identical branches with shared weights are trained simultaneously, so that the training parameters and weights can be largely reduced. We concatenated the extracted feature vectors of size $1 \times Dim$ into a $1 \times 2 \cdot Dim$ vector (Dim=512 and 1536 resp. for VGG16 and InceptionV4), which is used for the final classification of severity grade change between I_{t-1} and I_t .

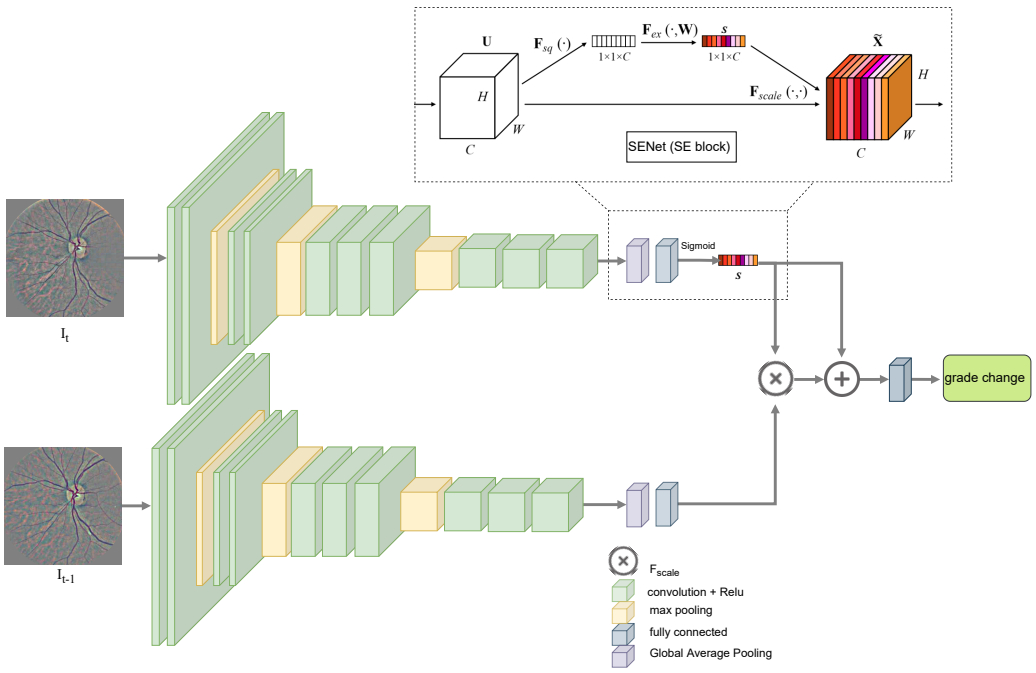


Figure 6.6 – Late fusion with an attention mechanism

6.3.3.4 Late fusion with an attention mechanism

On the basis of the simple late fusion of feature vectors, we propose to investigate an inter-attention mechanism. Our motivation is to enable the information from the prior image I_{t-1} to contribute to the DR severity grade classification of the current image I_t. Theoretically, not all spatial information of the image is equally important in contribution to the task. Only the information related to the task should be of concern. The key of the attention mechanism is to find useful information related to the task while neglecting other information (spatial transformation, noise...). In view of this, we add a squeeze-and-excitation network (SENet) (Hu et al., 2018) in the Siamese network as attention module (Fig. 6.6).

Squeeze-and-Excitation Network (SENet). The SENet is a channel-wise attention block which models the interdependencies between channels and adaptively recalibrates the contribution of each feature channel, thereby improving or removing different channels based on different tasks, and strengthening the representation performance of CNNs. The SE block first performs a squeeze operation $F_{sq}(\cdot)$, which compresses the spatial dimensions $(H \times W)$ of feature map U to obtain an embedding of the global distribution of the channel feature (i.e. each two-dimensional feature map becomes a real number, which is equivalent to a pooling operation). The number of channels remains unchanged. Subsequently, an excitation operation $F_{ex}(\cdot)$ produces a collection of channel-wise weights S on the global features. In the case of self-attention, these weights are applied to the feature maps U using a channel-wise multiplication $F_{scale}(\cdot)$.

In this work, we opt for a late fusion using an attention mechanism SENet. The processing steps are as follows: the non-registered image pairs $\{I_{t-1}, I_t\}$ are sent to a Siamese network with no weights sharing between the two branches. The current image I_t is used for attention extraction using an SE attention block. The output features of each branch are denoted as U_{t-1} and U_t:

$$U_t = F_{fea}(X_t)$$

$$U_{t-1} = F'_{fea}(X_{t-1})$$

where X_t, X_{t-1} are the model inputs, F_{fea} and F'_{fea} represent the feature extraction layers of each branch. Then, following the work of SENet (Hu et al., 2018), we opt for the global average pooling (GAP) as the squeeze operation $F_{sq}(\cdot)$, and a fully connected layer (denoted as $\mathcal{F}$) with sigmoid activation (σ) function as excitation operation $F_{ex}(\cdot)$. We model the weights S_t of each feature channel of I_t according to the DR severity grade change detection task:

$$Z_t = F_{sq}(U_t) = GAP(U_t)$$

$$S_t = F_{ex}(Z_t) = \sigma(\mathcal{F}(Z_t))$$

Then, we apply channel-wise multiplication $F_{scale}(\cdot)$ (denoted as $\otimes$) to the feature vectors generating from I_{t-1} to increase or decrease the weight of each feature channel. We sum up the weighted features with S_t to provide a bias to the final descriptor:

$$\widetilde{X} = F_{scale}(U_{t-1}, S_t) + bias = U_{t-1} \otimes S_t + S_t$$

6.4 Experiments and Results

6.4.1 Data pre-processing

Primary pre-processing of the dataset has been introduced in Sect. 2.2.3.2. In this study, images are further adjusted to various sizes depending on the model used. The InceptionV4 network receives as input images with size $299 \times 299 \times 3$. For the VGG16 model, we fix the input size to $224 \times 224 \times 3$ for all tests. Afterwards, random resized crops (scale range: [0.96, 1.0], aspect ratio range: [0.95, 1.05]) are applied for data augmentation. Note that random rotation and flip are not applied because the image pairs should keep aligned during the entire process.

6.4.2 Experiments and results

The various longitudinal fusion models for DR severity change detection are implemented using pytorch. Experiments are performed on an Nvidia GeForce GTX 1080Ti GPU (11GB/s) and trained using the SGD optimizer. We list in Tab. 6.3 the hyper-parameters used for VGG16 and InceptionV4

Network	imsize	learning rate	batch size	iteration
VGG16	224	0.005	32	20k
InceptionV4	299	0.005	16	20k

Table 6.3 – Hyper-parameters used for each deep network

Fusion	**Pre-training**			**VGG16**		**InceptionV4**	
	ImageNet	K-label	K-logistic	acc	AUC	acc	AUC
no fusion (only I_t)	✓			0.8594	0.9143	0.8510	0.9087
		✓		0.8603	0.9261	0.8692	0.9206
			✓	0.8555	0.9209	0.8632	0.9148
early fusion	✓			0.7684	0.8034	0.8187	0.8742
		✓		0.8140	0.8618	0.8392	0.8995
			✓	0.8179	0.8771	0.8383	0.8965
intermediate fusion	✓			0.7855	0.8513	0.7934	0.8451
		✓		0.8483	0.9032	0.8623	0.9091
			✓	0.8551	0.9151	0.8619	0.9088
late fusion	✓			0.8580	0.9216	0.8392	0.8993
		✓		**0.8696**	0.9289	**0.8756**	**0.9293**
			✓	0.8684	**0.9296**	0.8696	0.9168

Table 6.4 – Quantitative results using VGG16 (Simonyan & Zisserman, 2014) and InceptionV4 (Szegedy et al., 2017) backbones.

networks. The performance of each method is measured using the classification accuracy (acc) and area under the receiver operating characteristics curve (AUC) (Sect 3.2). The statistical significance was estimated using DeLong's t-test (Robin et al., 2011) to analyze and compare ROC curves.

Fusion results comparison. To explore the performance of the proposed longitudinal image fusion, we perform comparative experiments between different fusion schemes on two CNN architectures: VGG16 and InceptionV4. Three pre-training strategies are investigated for each model and each fusion scheme. In order to fairly compare these methods, we list in Tab. 6.4 their classification acc and AUC for VGG16 and InceptionV4, respectively. The baseline of each longitudinal fusion scheme is to train a CNN classifier using a single image I_t, without involving prior images.

According to our experimental results from Tab. 6.4, we can tell that the late fusion achieved the best performance for both models (acc = 0.8696, AUC = 0.9296, p = 0.007 for VGG16, acc = 0.8756, AUC = 0.9293, p = 0.006 for InceptionV4). The incorporation of the attention mechanism did not improve the classification performance. Surprisingly, for both models, incorporating the fusion of longitudinal studies in prior to the network (early fusion) or in the middle of the network (intermediate fusion) showed a considerable decrease compared to the no-fusion baseline. Nevertheless, it is noteworthy that the late fusion scheme remains a relevant strategy for both models, with better performance than other fusion schemes. In particular, the late fusion brings 0.2% - 0.9% AUC improvements to the baseline, with

statistical significance ($p < 0.05$).

Pre-train comparison. We compare three pre-training strategies (Sect. 6.3.2) for all fusion schemes. Apparently, regardless of the fusion scheme, pre-training on the OPHDIAT dataset largely boosts the classification performance, from 0.6% AUC (no fusion case) to 7.37% (intermediate fusion case with VGG16). The K-label pre-train brings average AUC improvements of 2.86% and 3.94% for VGG16 and InceptionV4, respectively, while the K-logistic pre-train brings about 3.36% and 3.58% gains. The proposed K-label model is slightly better than the K-logistic model in most cases.

Visual Explanations using Grad-CAM. The late fusion scheme achieves the best performance, but only slightly better than the no fusion benchmark. In order to find visual explanations for this conclusion, we visualize the activation maps using the Gradient-weighted Class Activation Mapping (Grad-CAM) algorithm (Selvaraju et al., 2017) and show some representative examples in Fig. 6.7.

The Grad-CAM algorithm is used to obtain a class activation map (similar to a heatmap), which can be used to locate sensitive areas related to the desired class in classification tasks. Given an input image and a class of interest, the image propagates through CNN with task-specific calculation to get a category score. Then, gradients are set to zero for all classes except the desired class. This signal is then back-propagated to the rectified convolutional feature maps, which will be combined to calculate the Grad-CAM (the blue heat-map in Fig. 6.7) which represents where the model pays attention to make particular decisions.

As shown in Fig. 6.7, three groups of Grad-CAM heat-map with the corresponding guided back-propagation feature map for three pairs of fundu images are presented. We compared the case of single image classification (classification individually of each image) and the late fusion scheme. Through observation, we can find the fact that when the model predict a normal case (RD grade = 2), the sensitive areas of the network are mostly concentrated in a centralized connected area (Fig. 6.7: a-1, a-3, a-4, a-6, b-2, b-4, b-6). On the contrary, when the prediction is "pathological" or "grade change", the sensitive areas are mostly scattered around the image (Fig. 6.7: a-5, b-1, b-3, b-5). Using late fusion, the difference of heat-maps (sensitive areas) between a pair of images is more obvious than no fusion. Nevertheless, we were also surprised to find that the sensitive areas of pathological image of the network are not exactly lesion areas, which is not what we expected. In view of this fact, we believe that more in-depth future work is needed to explore the interpretability of the network.

6.5 Discussion

In this study, we address the early-grade DR severity change detection by analyzing the fusion of two consecutive longitudinal follow-up images. Deep learning based DR classification that incorporates prior screening has not been exploited in existing studies, while the comparison with prior screening is an

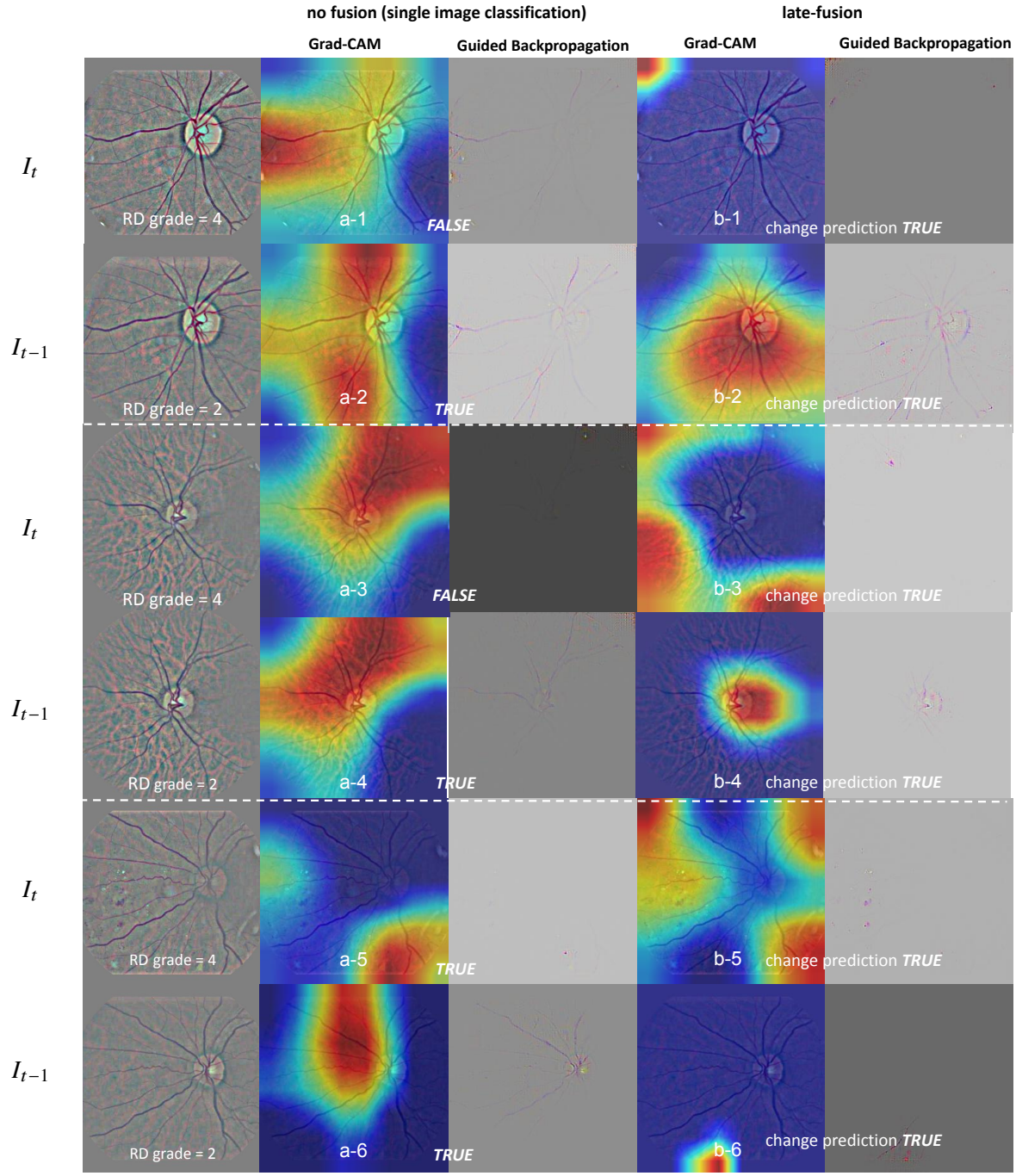


Figure 6.7 – Grad-CAM results illustration. From left to right are respectively the input images, heat-maps and guided back-propagation maps in the case of training with single images, heat-maps and guided back-propagation maps in the case of training using the late fusion scheme. Prediction results are labeled as *TRUE* or *FALSE* in the figure.

important step for clinicians towards the decision-making of the current study. Specifically, we studied the impact of the position of the fusion operations on network performance, and we additionally studied the effectiveness of the attention mechanism. Nevertheless, extensive experiments have demonstrated that the early and intermediate fusion can not bring further performance improvement, while a simple late fusion has shown stable performance gain. Our explanation for this is as follows: the experimental results of the no-fusion baseline have shown that the network can classify each image of the longitudinal pair with fairly good performance (AUC > 90%), indicating that the network has the capacity of extracting effective features from single-image DR severity classification purposes. However, as for the early or intermediate fusion, we hope that the network will focus on the lesion evolution at image-level or feature-map-level. This requires high-quality registration of the images to make sure that the lesion areas are well aligned. Moreover, due to the diversity of DR lesions and the subtlety of early lesions, it is more difficult for the network to target the lesion evolution. For the late fusion, the network firstly extract their respective effective features, followed by a Global Average Pooling layer, then the fusion operation is performed to the subsequent feature vectors that contain no spatial information. Accordingly, the mis-alignment will not affect the fusion results.

The main limitation of this work is the image registration of the input pair. As a pre-processing step, it requires higher registration quality; as a feature-level step, it is difficult to achieve automatic alignment in the middle of the network. From the current point of view, late fusion is still the simplest and most efficient method of image fusion. Experimental results validate that incorporating prior DR studies can improve the early-grade DR severity classification performance. In particular, the late fusion brings 0.2% - 0.9% AUC improvements to the baseline, with statistical significance ($p < 0.05$). This conclusion can also be extended to other medical imaging classification tasks. In the future, our method can be further investigated using multiple previous studies, or for other longitudinal pathology analysis, towards more accurate early diagnosis CAD systems.

CHAPTER 7

CONCLUSIONS AND FUTURE WORKS

Contents

Conclusions

Computer-aided medical imaging analysis has become an indispensable part of disease diagnosis, screening and treatment. To cope with the manual analysis of voluminous medical images which is inefficient, error-prone and highly expert-dependent, the application of deep learning (DL) based computer-aided diagnosis (CAD) is key. Automatic image processing techniques for disease diagnosis and pathological follow-up would be beneficial to the adaptive screening and management of each patient. In this thesis, we addressed the current limitations of traditional CAD systems by providing efficient and fully-automated DL methods, towards better interaction-free and more personalized medical care.

In this work, we investigated three main challenges associated with computer-assisted medical image analysis: (1) identification and segmentation of lesions from high resolution images, (2) multi-view information fusion for improved diagnosis, and (3) longitudinal prediction of severity grade changes. Specifically, we provided solutions through a comprehensive study of two clinical applications in screening dealing with the diagnosis of breast cancer and diabetic retinopathy (DR). In this thesis, we firstly presented the related clinical context in Chapter 1 and 2 and deep learning background in Chapter 3. Then, in each of the subsequent chapters, we carefully introduced our motivations, methodologies and experiments to each proposed approach. We elaborated and discussed the results obtained at the end of each chapter.

To deal with the first challenge, we studied automated mass segmentation from high-resolution full mammograms. To this end, two solutions from different perspectives were proposed. We first proposed to use a multi-scale cascade of convolutional encoder-decoders (CEDs) for segmentation without any pre-detection step. Multi-scale information was integrated using auto-context to make long-range spatial context arising from lower scale impact training at higher resolution. Our second contribution was to use a fully automated two-stage framework comprising a coarse-scale mass detection and a fine-scale mass segmentation, which are combined through a newly proposed multi-scale fusion strategy to eliminate false detections. By optimizing the performance of each stage, we achieved a good robustness against

the diversity of size, shape and appearance of breast masses. Both solutions are capable of identifying and segmenting lesions from high resolution images and have their own pros and cons: the first one-stage model is an end-to-end pipeline so that segmentation refinement was performed at each level simultaneously. Nevertheless, multiple deep CEDs need to be cascaded in order to make better use of the different context levels, which is less flexible and requires further research. The two-stage solution is more robust to lesions of any size, eliminating a large amount of false-positives proposals, but is substantially more complex to apply due to its multiple steps.

As we studied the second challenge, we attempted to take advantage of information arising from craniocaudal (CC) and mediolateral-oblique (MLO) multi-view mammograms to provide better diagnosis. Two methods were proposed for this purpose. First, a novel approach based on multi-view and multi-task learning was introduced. Specifically, we combined mass/non-mass classification with dual-view mass matching between complementary CC/MLO mammograms. Based on Siamese networks and contrastive learning, the integration of multi-view information has proved to be effective. By integrating mass detection, classification and matching, our method showed encouraging abilities to generalize to different deep models. Then, a second contribution based on active learning was subsequently proposed. As part of the multi-view information fusion, we applied a deep active learning approach that exploits dual-view consistency to mitigate the lack of labeled data, thereby reducing the workload of clinicians. Our contribution in this part was to combine the multi-view mammogram analysis with active learning, which to our knowledge have never been addressed before. Based on this method, it is possible to alleviate the burden of manual labeling in other multi-view medical image analysis scenarios, thereby contributing to the development of CAD tools. Extensive experiments of both studies reveal great effectiveness and promising robustness of multi-view information fusion in various aspects, thus highlighting the great potential of information fusion in medical imaging.

Regarding the third challenge, we intended to integrate longitudinal information of images to help analyze the lesion evolution. Deep learning based diabetic retinopathy (DR) classification that incorporates prior screening was exploited to address the referable DR severity change detection. Extensive experiments revealed that the early and intermediate fusion perform poorly in predicting severity change, while a simple late fusion have shown stable performance improvement. From an experimental point of view, our contribution lies in the comprehensive analysis of how fusion operations affect network performance, as well as the benefice of pre-training.

Future works

This thesis demonstrated that medical image analysis with deep learning is a powerful tool for clinical guidance in the fields of mammography and ophthalmology. We believe that it is a successful proof of concept for the development of more efficient and automated CAD systems. However, deep learning applied to medical images and pathology diagnosis is still a very large and untapped area, in which

plenty of new technologies, network architectures or training strategies need to be further investigated and deeper explored.

Here are some perspectives that we would like to address in future works:

1. Regarding mass segmentation using multi-scale cascaded convolutional encoder-decoders (CEDs), it could be interesting to involve cascading more CEDs to further refine the multi-scale information, which can be beneficial to the feature refinement. For the two-stage approach, instead of performing detection and segmentation separately, it would be possible to investigate an end-to-end pipeline that integrates both tasks.
2. Concerning the mammography study described in Sect. 5.1, future research should also take into account the potential impact of fusion of contralateral symmetry information to increase the robustness of breast lesion detection and description. Another potential subject is to deal with mass matching of multiple masses, which could be challenging but significant. Furthermore, the integration of the segmentation task within the pipeline dealing with combined classification and matching should deserve further investigation in dealing with the presence of negative patches, towards a more complete and complex multi-tasking CAD system.
3. As for the active learning-based approach (Sect. 5.2), further attempts of such strategy could focus on exploiting more complex query factors dealing with the multi-view consistency. Another future possible extension to this work is to integrate existing single-view criteria into our current framework, towards a unified active learning system.
4. Although our results in Chapter 6 demonstrated the potential of integrating a pair of longitudinal images, future research could continue to explore longitudinal fusion using multiple previous studies, instead of only pairs of consecutive examinations. We believe that future research on this topic might extend our explanations to the obtained results.
5. Last but not least, since the three challenges addressed in the thesis usually appear in many other clinical applications, this study provides a good starting point for further research for both identification and localization of other anatomical or pathological structures. More importantly, it is desirable for future works to extend the proposed methods from single-image processing towards the fusion of multi-view / temporal or multi-modality images, towards more generalized approaches regardless of the targeted clinical fields.

APPENDIX A

PUBLICATIONS

Here is the list of papers that have been published or submitted during my PhD.

Journal publications:

- (Yan et al., 2021b) **Yutong Yan**, Pierre-Henri Conze, Mathieu Lamard, Gwenolé Quellec, Béatrice Cochener and Gouenou Coatrieux. Towards improved breast mass detection using dual-view mammogram matching. Medical image analysis, 2021, vol. 71.
- (Yan et al., 2021d) **Yutong Yan**, Pierre-Henri Conze, Gwenolé Quellec, Mathieu Lamard, Beatrice Cochener and Gouenou Coatrieux. Two-stage multi-scale breast mass segmentation for full mammogram analysis without user intervention. Biocybernetics and Biomedical Engineering, 2021, vol. 41, no. 2, p. 746-757.

Conference and workshop papers:

- (Yan et al., 2021a) **Yutong Yan**, Pierre-Henri Conze, Gwenolé Quellec, Mathieu Lamard, Beatrice Cochener and Gouenou Coatrieux. Two-stage multi-scale mass segmentation from full mammograms. In : IEEE International Symposium on Biomedical Imaging (ISBI), 2021. p. 1628-1631.
- (Yan et al., 2020) **Yutong Yan**, Pierre-Henri Conze, Mathieu Lamard, Gwenolé Quellec, Béatrice Cochener, Gouenou Coatrieux. Multi-tasking Siamese networks for breast mass detection using dual-view mammogram matching. In : International MICCAI Workshop on Machine Learning in Medical Imaging (MLMI), 2020. p. 312-321.
- (Yan et al., 2019b) **Yutong Yan**, Pierre-Henri Conze, Etienne Decencière, Mathieu Lamard, Gwenolé Quellec, Béatrice Cochener and Gouenou Coatrieux. Cascaded multi-scale convolutional encoder-decoders for breast mass segmentation in high-resolution mammograms. In : International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2019, p. 6738-6741.
- (Yan et al., 2021c) **Yutong Yan**, Pierre-Henri Conze, Mathieu Lamard, Heng Zhang, Gwenolé Quellec, Béatrice Cochener and Gouenou Coatrieux. Deep active learning for dual-view mammogram analysis. International MICCAI Workshop on Machine Learning in Medical Imaging (MLMI), 2021.
- (Yan et al., 2021e) **Yutong Yan**, Pierre-Henri Conze, Gwenolé Quellec, Pascale Massin, Mathieu Lamard, Gouenou Coatrieux and Béatrice Cochener. Longitudinal detection of diabetic retinopathy early severity grade changes using deep learning. International MICCAI Workshop on Ophthalmic Medical Image Analysis (OMIA), 2021.

Bibliography

Bibliography

Abe, H., MacMahon, H., Engelmann, R., Li, Q., Shiraishi, J., Katsuragawa, S., Aoyama, M., Ishida, T., Ashizawa, K., Metz, C. E., et al., (2003), Computer-aided diagnosis in chest radiography: results of large-scale observer tests at the 1996–2001 RSNA scientific assemblies, *Radiographics*, *231*, 255–265.

Abràmoff, M. D., Garvin, M. K., & Sonka, M., (2010), Retinal imaging and image analysis, *IEEE Reviews in Biomedical Engineering*, *3*, 169–208.

Adal, K. M., Van Etten, P. G., Martinez, J. P., Rouwen, K. W., Vermeer, K. A., & van Vliet, L. J., (2017), An automated system for the detection and classification of retinal changes due to red lesions in longitudinal fundus images, *IEEE Transactions on Biomedical Engineering*, *656*, 1382–1390.

Agarwal, R., Diaz, O., Lladó, X., Yap, M. H., & Martı, R., (2019), Automatic mass detection in mammograms using deep convolutional neural networks, *Journal of Medical Imaging*, *63*, 1–9.

Akselrod-Ballin, A., Karlinsky, L., Hazan, A., Bakalo, R., Horesh, A. B., Shoshan, Y., & Barkan, E., (2017), Deep learning for automatic detection of abnormal findings in breast mammography. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*.

Al-antari, M. A., Al-masni, M. A., Choi, M.-T., Han, S.-M., & Kim, T.-S., (2018), A fully integrated computer-aided diagnosis system for digital X-ray mammograms via deep learning detection, segmentation, and classification, *International Journal of Medical Informatics*, *117*, 44–54.

Alaverdyan, Z., Jung, J., Bouet, R., & Lartizien, C., (2020), Regularized siamese neural network for unsupervised outlier detection on brain multiparametric magnetic resonance imaging: application to epilepsy lesion screening, *Medical Image Analysis*, *60*, 101618.

Amit, G., Hashoul, S., Kisilev, P., Ophir, B., Walach, E., & Zlotnick, A., (2015), Automatic dual-view mass detection in full-field digital mammograms, *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 44–52.

Arevalo, J., González, F. A., Ramos-Pollán, R., Oliveira, J. L., & Lopez, M. A. G., (2015), Convolutional neural networks for mammography mass lesion classification, *IEEE Engineering in Medicine and Biology Society*.

Asiri, N., Hussain, M., Al Adel, F., & Alzaidi, N., (2019), Deep learning based computer-aided diagnosis systems for diabetic retinopathy: A survey, *Artificial Intelligence in Medicine*, *99*, 101701.

Badrinarayanan, V., Kendall, A., & Cipolla, R., (2017), Segnet: A deep convolutional encoder-decoder architecture for image segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *3912*, 2481–2495.

Bengio, Y., Goodfellow, I., & Courville, A., (2017), *Deep learning* (Vol. 1), MIT Press.

Bernardes, R., Nunes, S., Pereira, I., Torrent, T., Rosa, A., Coelho, D., & Cunha-Vaz, J., (2009), Computer-assisted microaneurysm turnover in the early stages of diabetic retinopathy, *Ophthalmologica*, *2235*, 284–291.

Bowyer, K., Kopans, D., Kegelmeyer, W., Moore, R., Sallam, M., Chang, K., & Woods, K., (1996), The digital database for screening mammography, *Third International Workshop on Digital Mammography*, *58*, 27.

Budd, S., Robinson, E. C., & Kainz, B., (2019), A survey on active learning and human-in-the-loop deep learning for medical image analysis, *arXiv preprint arXiv:1910.02923*.

Byra, M., Jarosik, P., Szubert, A., Galperin, M., Ojeda-Fournier, H., Olson, L., O'Boyle, M., Comstock, C., & Andre, M., (2020), Breast mass segmentation in ultrasound with selective kernel U-Net convolutional neural network, *Biomedical Signal Processing and Control*, *61*.

Caballo, M., Pangallo, D. R., Mann, R. M., & Sechopoulos, I., (2020), Deep learning-based segmentation of breast masses in dedicated breast CT imaging: Radiomic feature stability between radiologists and artificial intelligence, *Computers in Biology and Medicine*, *118*.

Cai, J.-J., Tang, J., Chen, Q.-G., Hu, Y., Wang, X., & Huang, S.-J., (2019), Multi-view active learning for video recommendation., *International Joint Conference on Artificial Intelligence*, 2053–2059.

Carneiro, G., Nascimento, J., & Bradley, A. P., (2015), Unregistered multiview mammogram analysis with pre-trained deep learning models, *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 652–660.

Caruana, R., (1997), Multitask learning, *Machine learning*, *281*, 41–75.

Chan, H.-P., Hadjiiski, L. M., & Samala, R. K., (2020), Computer-aided diagnosis in the era of deep learning, *Medical Physics*, *475*, 218–227.

Choi, H., & Jin, K. H., (2016), Fast and robust segmentation of the striatum using deep convolutional neural networks, *Journal of Neuroscience Methods*, *274*, 146–153.

Choukroun, Y., Bakalo, R., Ben-Ari, R., Akselrod-Ballin, A., Barkan, E., & Kisilev, P., (2017), Mammogram classification and abnormality detection from nonlocal labels using deep multiple instance neural network, *Eurographics Workshop on Visual Computing for Biology and Medicine*.

Cong, L., Feng, W., Yao, Z., Zhou, X., & Xiao, W., (2020), Deep learning model as a new trend in computer-aided diagnosis of tumor pathology for lung cancer, *Journal of Cancer*, *1112*, 3615.

Conze, P.-H., Brochard, S., Burdin, V., Sheehan, F. T., & Pons, C., (2020), Healthy versus pathological learning transferability in shoulder muscle MRI segmentation using deep convolutional encoder-decoders, *Computerized Medical Imaging and Graphics*.

Conze, P.-H., Kavur, A. E., Gall, E. C.-L., Gezer, N. S., Meur, Y. L., Selver, M. A., & Rousseau, F., (2021), Abdominal multi-organ segmentation with cascaded convolutional and adversarial deep networks, *Artificial Intelligence in Medicine*, *117*.

Cover, T., & Hart, P., (1967), Nearest neighbor pattern classification, *IEEE Transactions on Information Theory*, *131*, 21–27.

Dai, J., Li, Y., He, K., & Sun, J., (2016), R-FCN: Object detection via region-based fully convolutional networks, *Advances in Neural Information Processing Systems*, 379–387.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L., (2009), ImageNet: A large-scale hierarchical image database, *IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Dhungel, N., Carneiro, G., & Bradley, A. P., (2017a), A deep learning approach for the analysis of masses in mammograms with minimal user intervention, *Medical Image Analysis*, *37*, 114–128.

Dhungel, N., Carneiro, G., & Bradley, A. P., (2017b), Fully automated classification of mammograms using deep residual neural networks, *IEEE International Symposium on Biomedical Imaging*, 310–314.

Doi, K., (2007), Computer-aided diagnosis in medical imaging: historical review, current status and future potential, *Computerized Medical Imaging and Graphics*, *314-5*, 198–211.

D'Orsi, C. J., (1996), The American college of radiology mammography lexicon: an initial attempt to standardize terminology., *American Journal of Roentgenology*, *1664*, 779–780.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S., (2017), Dermatologist-level classification of skin cancer with deep neural networks, *Nature*, *5427639*, 115–118.

Everingham, M., Eslami, S. M. A., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A., (2015), The Pascal visual object classes challenge: A retrospective, *International Journal of Computer Vision*, *1111*, 98–136.

Geras, K. J., Wolfson, S., Shen, Y., Wu, N., Kim, S., Kim, E., Heacock, L., Parikh, U., Moy, L., & Cho, K., (2017), High-resolution breast cancer screening with multi-view deep convolutional neural networks, *arXiv preprint arXiv:1703.07047*.

Gharaibeh, N., Al-Hazaimeh, O. M., Al-Naami, B., & Nahar, K. M., (2018), An effective image processing method for detection of diabetic retinopathy diseases from retinal fundus images, *International Journal of Signal and Imaging Systems Engineering*, *114*, 206–216.

Girshick, R., (2015), Fast R-CNN, *Proceedings of the IEEE International Conference on Computer Vision*, 1440–1448.

Girshick, R., Donahue, J., Darrell, T., & Malik, J., (2014), Rich feature hierarchies for accurate object detection and semantic segmentation, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 580–587.

Girshick, R., Donahue, J., Darrell, T., & Malik, J., (2015), Region-based convolutional networks for accurate object detection and segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *381*, 142–158.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y., (2014), Generative adversarial nets. *Advances in Neural Information Processing Systems* (pp. 2672–2680).

Gu, X., Shi, Z., & Ma, J., (2018), Multi-view learning for mammogram analysis: Auto-diagnosis models for breast cancer, *IEEE International Conference on Smart Internet of Things*, 149–153.

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al., (2016), Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs, *Jama*, *31622*, 2402–2410.

Hadsell, R., Chopra, S., & LeCun, Y., (2006), Dimensionality reduction by learning an invariant mapping, *IEEE Conference on Computer Vision and Pattern Recognition*, 1735–1742.

Haggstrom, M. et al., (2014), Medical gallery of mikael haggstrom 2014, *Wiki Journal of Medicine*, *12*, 1.

Hamidinekoo, A., Denton, E., Rampun, A., Honnor, K., & Zwiggelaar, R., (2018), Deep learning in mammography and breast histology, an overview and future trends, *Medical Image Analysis*, *47*, 45–67.

Han, X., Leung, T., Jia, Y., Sukthankar, R., & Berg, A. C., (2015), MatchNet: Unifying feature and metric learning for patch-based matching, *IEEE Conference on Computer Vision and Pattern Recognition*, 3279–3286.

He, K., Gkioxari, G., Dollár, P., & Girshick, R., (2017), Mask R-CNN, *Proceedings of the IEEE International Conference on Computer Vision*, 2961–2969.

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

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

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B., (1998), Support vector machines, *IEEE Intelligent Systems and their Applications*, *134*, 18–28.

Hizukuri, A., Nakayama, R., & Ashiba, H., (2017), Segmentation method of breast Masses on ultrasonographic images using level set method based on statistical model, *Journal of Biomedical Science and Engineering*, *104*.

Hmida, M., Hamrouni, K., Solaiman, B., & Boussetta, S., (2017), An Efficient Method for Breast Mass Segmentation and Classification in Mammographic Images, *International Journal of Advanced Computer Science and Applications*, *811*, 256–262.

Hu, J., Shen, L., & Sun, G., (2018), Squeeze-and-excitation networks, *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7132–7141.

Ioffe, S., & Szegedy, C., (2015), Batch normalization: Accelerating deep network training by reducing internal covariate shift, *International Conference on Machine Learning*, 448–456.

Jaderberg, M., Simonyan, K., Zisserman, A., & Kavukcuoglu, K., (2015), Spatial transformer networks, *arXiv preprint arXiv:1506.02025*.

Jørgensen, K. J., & Bewley, S., (2015), Breast cancer screening viewpoint of the IARC Working Group, *The New England Journal of Medicine*.

Jouirou, A., Baâzaoui, A., & Barhoumi, W., (2019), Multi-view information fusion in mammograms: A comprehensive overview, *Information Fusion*, *52*.

Jung, H., Kim, B., Lee, I., Yoo, M., Lee, J., Ham, S., Woo, O., & Kang, J., (2018), Detection of masses in mammograms using a one-stage object detector based on a deep convolutional neural network, *PloS One*, *139*.

Kanbayti, I. H., Rae, W. I., McEntee, M. F., Al-Foheidi, M., Ashour, S., Turson, S. A., & Ekpo, E. U., (2020), Is mammographic density a marker of breast cancer phenotypes?, *Cancer Causes & Control*.

Koch, G., Zemel, R., & Salakhutdinov, R., (2015), Siamese neural networks for one-shot image recognition, *ICML Deep Learning Workshop*.

Kooi, T., Litjens, G., Van Ginneken, B., Gubern-Mérida, A., Sánchez, C. I., Mann, R., den Heeten, A., & Karssemeijer, N., (2017), Large scale deep learning for computer aided detection of mammographic lesions, *Medical Image Analysis*, *35*, 303–312.

Kozegar, E., Soryani, M., Minaei, B., Domingues, I., et al., (2013), Assessment of a novel mass detection algorithm in mammograms, *Journal of Cancer Research and Therapeutics*, *94*, 592.

Krizhevsky, A., Sutskever, I., & Hinton, G. E., (2012), Imagenet classification with deep convolutional neural networks, *Advances in Neural Information Processing Systems*, 1097–1105.

LeCun, Y., Bengio, Y., & Hinton, G., (2015), Deep learning, *Nature*, *5217553*, 436–444.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D., (1989), Backpropagation applied to handwritten zip code recognition, *Neural Computation*, *14*, 541–551.

Lee, R., Gimenez, F., Hoogi, A., Kawai Miyake, K., Gorovoy, M., & Rubin, D., (2017), A curated mammography data set for use in computer-aided detection and diagnosis research, *Scientific Data*, *4*, 170177.

Lehman, C. D., Wellman, R. D., Buist, D. S., Kerlikowske, K., Tosteson, A. N., & Miglioretti, D. L., (2015), Diagnostic accuracy of digital screening mammography with and without computer-aided detection, *Journal of the American Medical Association: Internal Medicine*, *17511*, 1828–1837.

Lévy, D., & Jain, A., (2016), Breast mass classification from mammograms using deep convolutional neural networks, *arXiv preprint arXiv:1612.00542*.

Li, H., & Yin, Z., (2020), Attention, suggestion and annotation: A deep active learning framework for biomedical image segmentation, *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 3–13.

Li, H., Chen, D., Nailon, W. H., Davies, M. E., & Laurenson, D., (2018), Improved breast mass segmentation in mammograms with conditional residual u-net. *Image Analysis for Moving Organ, Breast, and Thoracic Images* (pp. 81–89).

Lin, M., Chen, Q., & Yan, S., (2013), Network in network, *arXiv preprint arXiv:1312.4400*.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S., (2017a), Feature pyramid networks for object detection, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2117–2125.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P., (2017b), Focal loss for dense object detection, *IEEE International Conference on Computer Vision*, 2980–2988.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L., (2014), Microsoft COCO: Common objects in context, *European Conference on Computer Vision*, 740–755.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., & Sánchez, C. I., (2017), A survey on deep learning in medical image analysis, *Medical Image Analysis*, *42*, 60–88.

Liu, H., Yue, K., Cheng, S., Pan, C., Sun, J., & Li, W., (2020), Hybrid Model Structure for Diabetic Retinopathy Classification, *Journal of Healthcare Engineering*, *2020*.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C., (2016), SSD: Single shot multibox detector, *European Conference on Computer Vision*, 21–37.

Liu, Y., Ren, L., Cao, X., & Tong, Y., (2020), Breast tumors recognition based on edge feature extraction using support vector machine, *Biomedical Signal Processing and Control*, *58*.

Long, J., Shelhamer, E., & Darrell, T., (2015), Fully convolutional networks for semantic segmentation, *IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440.

Lowe, D. G., (1999), Object recognition from local scale-invariant features, *Proceedings of the seventh IEEE International Conference on Computer Vision*, *2*, 1150–1157.

Ma, J., Liang, S., Li, X., Li, H., Menze, B. H., Zhang, R., & Zheng, W.-S., (2019), Cross-view relation networks for mammogram mass detection, *arXiv preprint arXiv:1907.00528*.

Massin, P., Chabouis, A., Erginay, A., Viens-Bitker, C., Lecleire-Collet, A., Meas, T., Guillausseau, P.-J., Choupot, G., André, B., & Denormandie, P., (2008), OPHDIAT©: A telemedical network screening system for diabetic retinopathy in the Île-de-France, *Diabetes & Metabolism*, *343*, 227–234.

Matheus, B. R. N., & Schiabel, H., (2011), Online mammographic images database for development and comparison of CAD schemes, *Journal of Digital Imaging*, *243*, 500–506.

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., et al., (2020), International evaluation of an AI system for breast cancer screening, *Nature*, *577*, 89–94.

Moreira, I. C., Amaral, I. F., Domingues, I., Cardoso, A. J. M., Cardoso, M. J., & Cardoso, J. S., (2012), INbreast: toward a full-field digital mammographic database., *Academic Radiology*.

Muslea, I., Minton, S., & Knoblock, C. A., (2006), Active learning with multiple views, *Journal of Artificial Intelligence Research*, *27*, 203–233.

Myers, E. R., Moorman, P., Gierisch, J. M., Havrilesky, L. J., Grimm, L. J., Ghate, S., Davidson, B., Mongtomery, R. C., Crowley, M. J., McCrory, D. C., et al., (2015), Benefits and harms of breast cancer screening: a systematic review, *Journal of the American Medical Association*, *31415*, 1615–1634.

Narasimha-Iyer, H., Can, A., Roysam, B., Tanenbaum, H. L., & Majerovics, A., (2007), Integrated analysis of vascular and nonvascular changes from color retinal fundus image sequences, *IEEE Transactions on Biomedical Engineering*, *548*, 1436–1445.

Ogurtsova, K., da Rocha Fernandes, J., Huang, Y., Linnenkamp, U., Guariguata, L., Cho, N. H., Cavan, D., Shaw, J., & Makaroff, L., (2017), IDF Diabetes Atlas: Global estimates for the prevalence of diabetes for 2015 and 2040, *Diabetes Research and Clinical Practice*, *128*, 40–50.

Oliver, A., Tortajada, M., Lladó, X., Freixenet, J., Ganau, S., Tortajada, L., Vilagran, M., Sentıs, M., & Martı, R., (2015), Breast density analysis using an automatic density segmentation algorithm, *Journal of Digital Imaging*, *285*, 604–612.

Pang, J., Chen, K., Shi, J., Feng, H., Ouyang, W., & Lin, D., (2019), Libra r-cnn: Towards balanced learning for object detection, *IEEE Conference on Computer Vision and Pattern Recognition*, 821–830.

Patel, S. N., Shi, A., Wibbelsman, T. D., & Klufas, M. A., (2020), Ultra-widefield retinal imaging: an update on recent advances, *Therapeutic advances in ophthalmology*, *12*, 2515841419899495.

Perek, S., Hazan, A., Barkan, E., & Akselrod-Ballin, A., (2018), Siamese network for dual-view mammography mass matching. *Image Analysis for Moving Organ, Breast, and Thoracic Images* (pp. 55–63), Springer.

Perek, S., Ness, L., Amit, M., Barkan, E., & Amit, G., (2019), Learning from longitudinal mammography studies, *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 712–720.

Quellec, G., Charrière, K., Boudi, Y., Cochener, B., & Lamard, M., (2017), Deep image mining for diabetic retinopathy screening, *Medical Image Analysis*, *39*, 178–193.

Quellec, G., Lamard, M., Lay, B., Guilcher, A. L., Erginay, A., Cochener, B., & Massin, P., (2019), Instant automatic diagnosis of diabetic retinopathy, *arXiv preprint arXiv:1906.11875*.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A., (2016), You Only Look Once: Unified, real-time object detection, *IEEE Conference on Computer Vision and Pattern Recognition*, 779–788.

Redmon, J., & Farhadi, A., (2018), YOLOv3: an incremental improvement, *arXiv preprint arXiv:1804.02767*.

Ren, S., He, K., Girshick, R., & Sun, J., (2015), Faster R-CNN: towards real-time object detection with region proposal networks, *Advances in Neural Information Processing Systems*, 91–99.

Ribli, D., Horváth, A., Unger, Z., Pollner, P., & Csabai, I., (2018), Detecting and classifying lesions in mammograms with deep learning, *Scientific Reports*, *81*, 4165.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M., (2011), pROC: an open-source package for R and S+ to analyze and compare ROC curves, *BMC Bioinformatics*, *121*, 1–8.

Ronneberger, O., Fischer, P., & Brox, T., (2015a), U-Net: Convolutional Networks for Biomedical Image Segmentation, *Medical Image Computing and Computer-Assisted Intervention*, 234–241.

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

Roth, H. R., Shen, C., Oda, H., Sugino, T., Oda, M., Hayashi, Y., Misawa, K., & Mori, K., (2018), A Multi-scale pyramid of 3D fully convolutional networks for abdominal multi-organ segmentation, *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 417–425.

Ruder, S., (2017), An overview of multi-task learning in deep neural networks, *arXiv preprint arXiv:1706.05098*.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., (2015), ImageNet large scale visual recognition challenge, *International Journal of Computer Vision*, *1153*, 211–252.

Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Unwin, N., Colagiuri, S., Guariguata, L., Motala, A. A., Ogurtsova, K., et al., (2019), Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, *Diabetes Research and Clinical Practice*, *157*, 107843.

Saha, S. K., Xiao, D., Bhuiyan, A., Wong, T. Y., & Kanagasingam, Y., (2019), Color fundus image registration techniques and applications for automated analysis of diabetic retinopathy progression: A review, *Biomedical Signal Processing and Control*, *47*, 288–302.

Salehi, S. S. M., Erdogmus, D., & Gholipour, A., (2017), Auto-context convolutional neural network (auto-net) for brain extraction in magnetic resonance imaging, *IEEE Transactions on Medical Imaging*, *3611*, 2319–2330.

Santeramo, R., Withey, S., & Montana, G., (2018), Longitudinal detection of radiological abnormalities with time-modulated LSTM. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (pp. 326–333).

Santos, M. K., Ferreira Júnior, J. R., Wada, D. T., Tenório, A. P. M., Barbosa, M. H. N., & Marques, P. M. d. A., (2019), Artificial intelligence, machine learning, computer-aided diagnosis, and radiomics: advances in imaging towards to precision medicine, *Radiologia Brasileira*, *526*, 387–396.

Sapate, S., Talbar, S., Mahajan, A., Sable, N., Desai, S., & Thakur, M., (2020), Breast cancer diagnosis using abnormalities on ipsilateral views of digital mammograms, *Biocybernetics and Biomedical Engineering*, *401*, 290–305.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D., (2017), Grad-cam: Visual explanations from deep networks via gradient-based localization, *Proceedings of the IEEE international conference on computer vision*, 618–626.

Sener, O., & Savarese, S., (2017), Active learning for convolutional neural networks: A core-set approach, *arXiv preprint arXiv:1708.00489*.

Shankar, K., Sait, A. R. W., Gupta, D., Lakshmanaprabu, S., Khanna, A., & Pandey, H. M., (2020), Automated detection and classification of fundus diabetic retinopathy images using synergic deep learning model, *Pattern Recognition Letters*, *133*, 210–216.

Shen, H., Tian, K., Dong, P., Zhang, J., Yan, K., Che, S., Yao, J., Luo, P., & Han, X., (2020), Deep Active Learning for Breast Cancer Segmentation on Immunohistochemistry Images, *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 509–518.

Shen, L., Margolies, L. R., Rothstein, J. H., Fluder, E., McBride, R., & Sieh, W., (2019), Deep learning to improve breast cancer detection on screening mammography, *Scientific Reports*, *91*, 1–12.

Shen, R., Yan, K., Tian, K., Jiang, C., & Zhou, K., (2019), Breast mass detection from the digitized X-ray mammograms based on the combination of deep active learning and self-paced learning, *Future Generation Computer Systems*, *101*, 668–679.

Sikder, N., Masud, M., Bairagi, A. K., Arif, A. S. M., Nahid, A.-A., & Alhumyani, H. A., (2021), Severity Classification of Diabetic Retinopathy Using an Ensemble Learning Algorithm through Analyzing Retinal Images, *Symmetry*, *134*, 670.

Simonyan, K., & Zisserman, A., (2014), Very deep convolutional networks for large-scale image recognition, *arXiv preprint arXiv:1409.1556*.

Singel, R. L., Miller, K. D., & Jemal, A., (2018), Cancer statistics, 2018., *CA: A Cancer Journal for Clinicians*, *681*, 7–30.

Singh, V. K., Rashwan, H. A., Romani, S., Akram, F., Pandey, N., Sarker, M. M. K., Saleh, A., Arenas, M., Arquez, M., Puig, D., et al., (2020), Breast tumor segmentation and shape classification in mammograms using generative adversarial and convolutional neural network, *Expert Systems with Applications*, *139*, 112855.

Skarping, I., Förnvik, D., Sartor, H., Heide-Jørgensen, U., Zackrisson, S., & Borgquist, S., (2019), Mammographic density is a potential predictive marker of pathological response after neoadjuvant chemotherapy in breast cancer, *BMC cancer*, *191*, 1–11.

Song, E., Xu, S., Xu, X., Zeng, J., Lan, Y., Zhang, S., & Hung, C.-C., (2010), Hybrid segmentation of mass in mammograms using template matching and dynamic programming, *Academic radiology*, *1711*, 1414–1424.

Stanford Health Care, (2017), Digital Mammography, https://stanfordhealthcare.org/medical-tests/m/mammogram/digital-mammography.html

Suckling J, P., (1994), The mammographic image analysis society digital mammogram database, *Digital Mammo*, 375–386.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F., (2021), Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, *CA: A Cancer Journal for Clinicians*.

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A., (2017), Inception-V4, inception-resnet and the impact of residual connections on learning, *Proceedings of the AAAI Conference on Artificial Intelligence*, *311*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A., (2015), Going deeper with convolutions, *IEEE Conference on Computer Vision and Pattern Recognition*.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z., (2016), Rethinking the inception architecture for computer vision, *IEEE Conference on Computer Vision and Pattern Recognition*, 2818–2826.

Tan, M., & Le, Q. V., (2019), EfficientNet: rethinking model scaling for convolutional neural networks, *arXiv preprint arXiv:1905.11946*.

Tardy, M., & Mateus, D., (2021), Looking for abnormalities in mammograms with self-and weakly supervised reconstruction, *IEEE Transactions on Medical Imaging*.

Thomas, R., Halim, S., Gurudas, S., Sivaprasad, S., & Owens, D., (2019), IDF Diabetes Atlas: A review of studies utilising retinal photography on the global prevalence of diabetes related retinopathy between 2015 and 2018, *Diabetes Research and Clinical Practice*, *157*, 107840.

Tu, Z., & Bai, X., (2010), Auto-context and its application to high-level vision tasks and 3D brain image segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *3210*, 1744–1757.

Uijlings, J. R., Van De Sande, K. E., Gevers, T., & Smeulders, A. W., (2013), Selective search for object recognition, *International Journal of Computer Vision*, *1042*, 154–171.

Vijayarajan, S., & Jaganathan, P., (2014), Breast cancer segmentation and detection using multi-view mammogram, *Academic Journal of Cancer Research*, *72*.

Virmani, J., Agarwal, R. et al., (2019), Effect of despeckle filtering on classification of breast tumors using ultrasound images, *Biocybernetics and Biomedical Engineering*, *392*, 536–560.

Wang, H., Feng, J., Zhang, Z., Su, H., Cui, L., He, H., & Liu, L., (2018), Breast mass classification via deeply integrating the contextual information from multi-view data, *Pattern Recognition*, *80*, 42–52.

Wang, W., & Zhou, Z.-H., (2008), On multi-view active learning and the combination with semi-supervised learning, *Proceedings of the International Conference on Machine Learning*, 1152–1159.

Warren, R. M., Duffy, S., & Bashir, S., (1996), The value of the second view in screening mammography, *The British Journal of Radiology*, *69818*, 105–108.

Wilkinson, C., Ferris III, F. L., Klein, R. E., Lee, P. P., Agardh, C. D., Davis, M., Dills, D., Kampik, A., Pararajasegaram, R., Verdaguer, J. T., et al., (2003), Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales, *Ophthalmology*, *1109*, 1677–1682.

World Health Organization, (2019), World report on vision.

Yan, Y., Conze, P.-H., Decencière, E., Lamard, M., Quellec, G., Cochener, B., & Coatrieux, G., (2019a), Cascaded multi-scale convolutional encoder-decoders for breast mass segmentation in high-resolution mammograms, *IEEE International Engineering in Medicine and Biology*.

Yan, Y., Conze, P.-H., Decencière, E., Lamard, M., Quellec, G., Cochener, B., & Coatrieux, G., (2019b), Cascaded multi-scale convolutional encoder-decoders for breast mass segmentation in high-resolution mammograms, *International Conference of the IEEE Engineering in Medicine and Biology Society*, 6738–6741.

Yan, Y., Conze, P.-H., Quellec, G., Lamard, M., Cochener, B., & Coatrieux, G., (2021a), Two-stage multi-scale mass segmentation from full mammograms, *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 1628–1631.

Yan, Y., Conze, P.-H., Lamard, M., Quellec, G., Cochener, B., & Coatrieux, G., (2020), Multi-tasking Siamese networks for breast mass detection using dual-view mammogram matching, *International Workshop on Machine Learning in Medical Imaging*, 312–321.

Yan, Y., Conze, P.-H., Lamard, M., Quellec, G., Cochener, B., & Coatrieux, G., (2021b), Towards improved breast mass detection using dual-view mammogram matching, *Medical Image Analysis*, *71*, 102083.

Yan, Y., Conze, P.-H., Lamard, M., Zhang, H., Quellec, G., Cochener, B., & Coatrieux, G., (2021c), Deep active learning for dual-view mammogram analysis, *International Workshop on Machine Learning in Medical Imaging*, 180–189.

Yan, Y., Conze, P.-H., Quellec, G., Lamard, M., Cochener, B., & Coatrieux, G., (2021d), Two-stage multi-scale breast mass segmentation for full mammogram analysis without user intervention, *Biocybernetics and Biomedical Engineering*.

Yan, Y., Conze, P.-H., Quellec, G., Massin, P., Lamard, M., Coatrieux, G., & Cochener, B., (2021e), Longitudinal detection of diabetic retinopathy early severity grade changes using deep learning, *International Workshop on Ophthalmic Medical Image Analysis*, 11–20.

Yap, M. H., Goyal, M., Osman, F., Marti, R., Denton, E., Juette, A., & Zwiggelaar, R., (2020), Breast Ultrasound Region of Interest Detection and Lesion Localisation, *Artificial Intelligence in Medicine*.

Zagoruyko, S., & Komodakis, N., (2015), Learning to compare image patches via convolutional neural networks, *IEEE Conference on Computer Vision and Pattern Recognition*, 4353–4361.

Zhang, H., Fromont, E., Lefevre, S., & Avignon, B., (2020), Localize to classify and classify to localize: mutual guidance in object detection, *Proceedings of the Asian Conference on Computer Vision (ACCV)*.

Zhang, H., Fromont, E., Lefevre, S., & Avignon, B., (2021), Guided attentive feature fusion for multispectral pedestrian detection, *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 72–80.

Zhang, X., Zhang, Y., Han, E. Y., Jacobs, N., Han, Q., Wang, X., & Liu, J., (2018), Classification of whole mammogram and tomosynthesis images using deep convolutional neural networks, *IEEE Transactions on Nanobioscience*, *173*, 237–242.

Zhao, Y., Chen, D., Xie, H., Zhang, S., & Gu, L., (2019), Mammographic image classification system via active learning, *Journal of Medical and Biological Engineering*, *394*, 569–582.

Zhou, H., Zaninovich, Y., & Gregory, C., (2017), Mammogram classification using convolutional neural networks, *International Conference on Technology Trends*.

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

Zhu, W., Xiang, X., Tran, T. D., Hager, G. D., & Xie, X., (2018), Adversarial deep structured nets for mass segmentation from mammograms, *IEEE International Symposium on Biomedical Imaging*, 847–850.

Zhu, W., Lou, Q., Vang, Y. S., & Xie, X., (2017), Deep multi-instance networks with sparse label assignment for whole mammogram classification, *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 603–611.

DOCTORAT BIOLOGIE BRETAGNE SANTE LOIRE

UBO université de bretagne occidentale

LATIM

Titre : Analyse d'images médicales par apprentissage profond pour le diagnostic assisté par ordinateur dans un contexte de dépistage

Mot clés : diagnostic assisté par ordinateur, dépistage, apprentissage profond, fusion d'informations, cancer du sein, rétinopathie diabétique

Résumé : L'analyse d'images médicales assistée par ordinateur est cruciale pour l'aide au diagnostic, au pronostic et au suivi thérapeutique. En particulier, le récent développement de techniques issues de l'intelligence artificielle appliquées au diagnostic et au dépistage représente une perspective prometteuse. Pour faire face aux limites des systèmes traditionnels de diagnostic assisté par ordinateur (CAD), nous avons proposé dans cette thèse un ensemble de méthodes d'apprentissage profond efficaces et automatisées, visant à améliorer la prise en charge personnalisée des patients. Dans les contextes de dépistage du cancer du sein et de la rétinopathie diabétique, nous avons principalement étudié trois défis associés à l'analyse d'images médicales assistée par ordinateur : (1) l'identification et la segmentation de lésions à partir d'images acquises à haute résolution, (2) la fusion d'informations multi-vues pour un diagnostic amélioré, et (3) la prédiction longitudinale de changements de grade de sévérité. Notre contribution au premier défi a été de développer deux méthodes dédiées à la segmentation de masses à partir de mammographies natives, à haute résolution. Dans un premier temps, nous avons proposé un pipeline de segmentation entraîné de bout en bout consistant à exploiter le contexte spatial multi-échelle grâce à une cascade d'encodeur-décodeurs convolutifs exploitant le paradigme de l'auto-contexte. Ensuite, nous avons développé une approche alternative à deux étapes, combinant la localisation de masses basée sur l'image entière et exploitant une stratégie de fusion des prédictions effectuées à multiples résolutions et la segmentation de masses sur les régions d'intérêts extraites au moyen d'un réseau profond avec connexions imbriquées et denses. Le deuxième défi a été relevé en tirant profit des informations issues des vues cranio-caudale (CC) et médiolatérale-oblique (MLO) des examens mammographiques. Deux méthodes ont ainsi été proposées. Tout d'abord, une nouvelle approche basée sur l'apprentissage multi-tâches a été introduite fournissant des détections de masses précises ainsi que des correspondances entre masses issues des deux vues. Ensuite, nous avons développé une approche d'apprentissage actif exploitant la cohérence inter-vues pour diminuer la charge d'annotations des cliniciens. Ces méthodes ont démontré l'efficacité de l'intégration d'informations issues de multiples vues pour la détection ou la segmentation. Pour le dernier défi, nous avons analysé des paires d'images de fond d'œil consécutives pour la détection de changements de grade de sévérité de la rétinopathie diabétique. Ces contributions permettent d'analyser automatiquement différentes images médicales dans diverses situations et promettent de fournir un support pertinent pour le développement de systèmes de CAD nouvelle génération.

DOCTORAT BIOLOGIE BRETAGNE SANTE LOIRE

UBO université de bretagne occidentale

LATIM

Title: Medical image analysis with deep learning for computer-aided diagnosis in screening

Keywords: computer-aided diagnosis, pathology screening, deep learning, information fusion, breast cancer, diabetic retinopathy

Abstract: Computer-aided medical image analysis is essential to support clinicians in diagnosis, prognosis and therapy-related decisions through fast, repeatable and objective measurements made by computational resources. In particular, the latest development of artificial intelligence applied to diagnosis and screening represents a promising perspective. In this thesis, we addressed the current limitations of traditional computer-aided diagnosis (CAD) systems by providing efficient and fully-automated deep learning methods towards better interaction-free and more personalized medical care. In the contexts of breast cancer and diabetic retinopathy screening, we investigated three main challenges associated with computer-assisted medical image analysis: (1) identification and segmentation of lesions from high-resolution images, (2) multi-view information fusion for improved diagnosis, and (3) longitudinal prediction of severity grade changes. Our initial contribution to the first challenge was to propose an end-to-end mass segmentation pipeline that exploits long-range multi-scale spatial context through a cascade of convolutional encoder-decoders embedding the auto-context paradigm. Then, as a second contribution, we proposed a two-stage framework combining a deep coarse-scale mass localization involving a multi-scale fusion strategy and a fine-scale mass segmentation. The second challenge was addressed by fusing information arising from two standard mammography views, namely craniocaudal (CC) and mediolateral-oblique (MLO). Two methods were proposed towards this goal. First, a novel approach based on multi-task learning was introduced, combining mass classification with dual-view mass matching between CC/MLO mammograms. Then, we applied a label-efficient deep active learning approach that exploits dual-view consistency to mitigate the labeling workload of clinicians. These methods demonstrate the effectiveness of integrating multi-view information for detection or segmentation purposes. For the last challenge, we incorporated the prior screening of fundus images to address the referable diabetic retinopathy severity change detection. All these contributions can automatically analyze different medical images in various situations and are promising to provide relevant support for the development of the next generation of CAD systems.