ABSTRACT

Title of Dissertation: DEEP LEARNING APPLICATIONS IN BONE MINERAL DENSITY ESTIMATION, SPINE VERTEBRA DETECTION, AND LIVER TUMOR SEGMENTATION

Fakai Wang
Doctor of Philosophy, 2023

Dissertation Directed by: Professor Min Wu
Department of Electronic and Computer Engineering
University of Maryland

As the aging population and related health concerns emerge in more countries than ever, we face many challenges such as the availability, quality, and cost of medical resources. Thanks to the development of machine learning and computer vision in recent years, Deep Learning (DL) can help solve some medical problems. The diagnosis of various diseases (such as spine disorders, low bone mineral density, and liver cancer) relies on X-rays or Computed Tomography (CT). DL models could automatically analyze these radiography scans and help with the diagnosis. Different organs and diseases have distinct characteristics, requiring customized algorithms and models. In this dissertation, we investigate several Computer Aided-Diagnosis (CAD) tasks and present corresponding DL solutions.

Deep Learning has multiple advantages. Firstly, DL models could uncover underlying

health issues invisible to humans. One example is the opportunistic screening of Osteoporosis through chest X-ray. We develop DL models, utilizing chest film to predict bone mineral density, which helps prevent bone fractures. Humans could not tell anything about bone density in the chest film, but DL models could reliably make the prediction. The second advantage is accuracy and efficiency. Reading radiography is tedious, requiring years of expertise. This is particularly true when a radiologist needs to localize potential liver tumors by looking through tens of CT slices, spending several minutes. Deep learning models could localize and identify the tumors within seconds, greatly reducing human labor. Experiments show DL models can pick up small tumors, which are hardly seen by the naked eye.

Attention should be paid to deep learning limitations. Firstly, DL models lack explainability. Deep learning models store diagnostic knowledge and statistical patterns in their parameters, which are obscure to humans. Secondly, uncertainty exists for rare diseases. If not exposed to rare cases, the models would yield uncertain outcomes. Thirdly, training AI models are subject to high-quality data and the labeling quality varies from clinical practice. Despite the challenges and issues, deep learning models are promising to promote medical diagnosis and benefit society.

DEEP LEARNING APPLICATIONS IN BONE MINERAL DENSITY ESTIMATION, SPINE VERTEBRA DETECTION, AND LIVER TUMOR SEGMENTATION

by

Fakai Wang

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2023

Advisory Committee:
Professor Min Wu, Chair/Advisor
Professor Matthias Zwicker,
Professor Joseph F. JaJa,
Professor Gang Qu,
Dr. Le Lu

Dedication

To my family in Harbin, China, who give me emotional support.

Acknowledgments

I owe my gratitude to all who help me in research, experiment, paper writing, publication, life advice.

First and foremost I'd like to thank my PhD advisor, Professor Min Wu. She has given valuable advice in many aspects, including project details, paper reviewing, academic presentation, career development. The MAST group meetings have taught me a lot of lessons, which are fruitful and memorable. I also want to give thanks to all the MAST members, for we discuss interesting ideas and we grow together.

I would like to thank Dr. Le Lu, who provides the internship opportunities, mental support. He has given me important guidance on career development, research attitudes based on his long-term commitment to the medical imaging area. His research advice and encouragement have always been a powerful source for me, to aim high and work hard.

I would like to thank Dr. Shun Miao, who leads me into the medical imaging research area. He gives me a lot of mental support and technical help, in order for me to get involved and progress. "The initial step is the hardest", I get confidence and inspirations from working with him. I will always appreciate the research attitude and ability he teaches me.

I would like to thank Dr. Ling Zhang, who gives me research directions and guidance. He provides the initial plans for the liver tumor detection project, and gives

detailed and insightful suggestions. He gives great help with the RSNA abstract publication, collaboration with doctors, pushing forward the project, and research strategies.

I appreciate the fantastic PAII people, who devote wholeheartedly to medical imaging research. I own many thanks to Dr. Kang Zheng, who provides great help in the spine vertebra detection project and Bone Mineral Density estimation projects. I want to thank Yizhi Chen, since she organized many lunches and activities which I enjoyed a lot. I also want to thank Dr. Yingda Xia and Dr. Ke Yan, who give many helpful ideas with the liver tumor detection project. All the colleagues in PAII are very helpful, which I will always be grateful.

I would like to thank my collaborators in Chang Gung Memorial Hospital (CGMH), who provide the research agenda and high-quality data. Many thanks to Dr. Chang-Fu Kuo, who gives direction and clinical knowledge in the Rheumatoid Arthritis (RA) project, Bone Mineral Density estimation projects. I own many thanks to Una Chen, for she helps collect data and shows great patience in model verification. I also want to thank Dr. Chi-Tung Cheng and Dr. Chien-Wei Peng, who give expert knowledge in liver disease, teach me to read the CT image, and help with the liver tumor labeling. It is a fortune to work with all the diligent doctors in my PhD career, and I will always remember!

Table of Contents

Bibliography 135

List of Tables

List of Figures

List of Abbreviations

ACC	Accuracy
AI	Artificial Intelligent
AUC	Area Under Curve
BMD	Bone Mineral Density
CAD	Computer-Aided Diagnosis
CNN	Convolutional Neural Network
CT	Computed Tomography
CV	Computer Vision
CXR	Chest X-ray
DICOM	Digital Imaging and Communications in Medicine
DL	Deep Learning
FN	False Negative
FP	False Positive
HCC	Hepatic Cellular Carcinoma
LR, lr	Learning Rate
META	Metastasis
ML	Machine Learning
MRI	Magnetic Resonance Image
NIFTI	Neuroimaging Informatics Technology Initiative
PACS	Picture Archiving and Communication System
RA	Rheumatoid Arthritis
SENS	Sensitivity
SPEC	Specificity
TN	True Negative
TP	True Positive

Chapter 1

Introduction

As an important part of Computer-Aided Diagnosis (CAD), medical image analysis feeds into medical scanning and applies machine learning techniques to recognize clinically valuable clues automatically in order to guide doctors in the screening, diagnosis, treatment, and operation planning of various diseases. While medical scanning has been ubiquitously applied to human bones, organs, and soft tissues, inspection and statistics still heavily rely on manual work. Medical practitioners and computer engineers have long sought to automate imaging analysis for improved efficiency and quality, but the process has been impeded by the lack of adequate computer algorithms and models. With the emergence of advanced computing hardware, large-scale datasets, and deep learning models, computer vision-related areas have been revolutionized in the past 15 years. Deep learning models outperform humans in many visual recognition tasks, and the same trends are also applied to many medical imaging problems.

Preciseness and accuracy are required in the clinical setting, medical imaging tasks depend not only on computer vision technologies but more importantly on quality data and clinical knowledge. Each type of study subject (organs, bones, or soft tissues) can

have various distributions of anomalies. Therefore it calls for specialized medical image analysis models for a particular scanning modality of the specific body part. In this dissertation, we look at three medical image analysis tasks, namely Bone Mineral Density estimation from Chest X-ray images, spine vertebra detection via CT images, and liver tumor detection via CT images. These projects cover multiple task categories, including regression, classification, localization, segmentation, and detection.

Before we delve into a particular task, we look at the general concepts, goals, and methodologies of medical image analysis. In this chapter, we take an overview of the study topics in this dissertation. We will provide the background as well as motivations for each medical imaging problem. The readers can go to individual topics (Chapter 3,4,5) directly if interested. Chapter 2 gives more emphasis on the technological foundations of computer vision and deep learning, as well as the critical aspects of medical imaging analysis. The readers can find this fundamental knowledge in many online resources, and we put them here simply for easy reference.

1.1 Medical Image analysis tasks

Medical image analysis is the process of extracting meaningful patterns from 2D or 3D images. Hospitals conduct medical image scanning for disease screening, diagnosis, and treatment planning. Without resection, doctors can find important clues for many diseases from the medical scans alone, such as bone fractures, organ tumors, and coronary artery anomalies. However, reading medical scans requires many years of training and can be a tedious job when a radiologist need to read hundreds of patient records daily.

Computer algorithms can automate this process, extract target information and store it in a database consistently, with the help of computer vision models.

Computer vision deals with object classification, object detection, segmentation, and geometry sensing in general. Everyday life is usually represented by RGB images or videos, which is a single or a sequence of 2D projections captured by consumer cameras. Recognizing real-world visionary concepts through 2D pixel representations lies at the heart of computer vision. Machine learning models are designed to delineate vision signals from simple lines, corners, outlines, and complex textures. The enthusiasm for machine vision has gone a long way, beginning with traditional methods such as handcrafted feature matching and registration. Statistical methods have helped advance computer vision, but still far from applicable in many scenarios. Recently, learning-based models and novel layers such as convolutional kernels, and activation functions, are the main force that enables computer vision applications for real-world general usage.

Medical image analysis depends on the technological advancements of computer vision. Many medical tasks are computer vision problems by nature, such as detecting tumors in the organ, classifying the microscopic anatomy of biological tissues (histology), and predicting bone mineral densities from bone textures. Therefore similar computational models can be utilized in the medical imaging domain, to help with disease discovery and monitoring. There are many differences and similarities between everyday life vision and medical scans. Above all, medical imaging has much higher requirements for preciseness and accuracy in the processes of scanning, recognizing, and determining, because of medical regulations and ethics. Detailed differences in terms of imaging modality, model architecture, and analyzing metrics will be elaborated in Chapter 2.

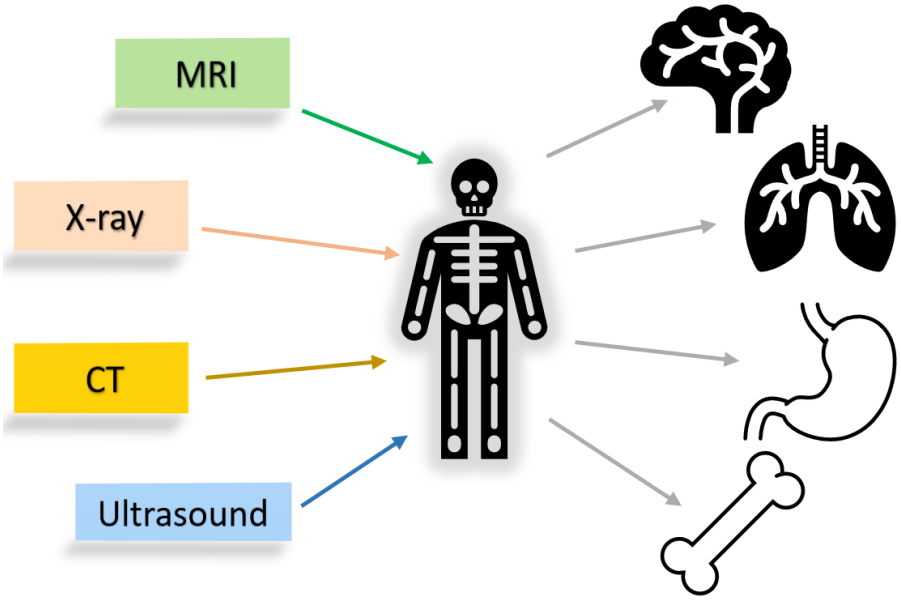

Figure 1.1: Medical imaging utilizes different scanning modalities on a variety of body parts. Many diseases manifest structural changes inside the body, which can be captured by Ultrasonography, X-ray, MRI, and CT. In the diagnosis scenario, the scanning usually focuses on certain organs with specific imaging equipment.

1.1.1 Medical image analysis tasks

Medical imaging has been widely used in the diagnosis and monitoring of many fatal diseases, such as cancer and Alzheimer's Disease (AD). Both cancer and AD would develop lesions or texture changes inside the organs for a period of time, before showing obvious symptoms. Before the signs or symptoms appear, the initial recognition of these diseases is mostly through screening, including blood tests, and radiology scanning. The suspected candidates then need finer evidence such as a pathology study of organ tissues through biopsy before making the final confirmation.

According to the Global Cancer Statistics [1], an estimated 19.3 million new cancer cases and almost 10.0 million cancer deaths occurred in 2020. The leading cancers by incidence rate are female breast cancer (11.7%), lung cancer (11.4%), colorectal cancer(10.0

%), prostate cancer(7.3%), and stomach cancer(5.6%). By the mortality number, the ranks become lung cancer (18%), colorectal (9.4%), liver (8.3%), stomach (7.7%), and female breast cancer(6.9%). It is a consensus that building sustainable infrastructure for cancer prevention in transitioning countries is of critical importance. Medical imaging screening of cancers enables cancer detection at an early stage, making room for better progressive treatment. X-ray and CT images have been well adapted to screen many cancers, such as breast cancer, lung cancer, colorectal cancer, liver cancer, and stomach cancer.

Being the 6th leading cause of death and projected to increase over the years to come in the United States [2], Alzheimer's Disease (AD) is the most common cognition-degeneration disease, characterized by memory loss and brain changes. As neurons die and connections break down, the brain regions may shrink. Depending on the Alzheimer's stages, the degree of brain atrophy would cause varied brain volume loss. A CT or MRI scan of the brain serves as an important clue for doctors to tell whether Alzheimer's Disease exists. However, Alzheimer's Disease may affect the brain years before the symptom manifestation, therefore the diagnosis may be delayed and the early treatment is missed. If machine algorithms could detect early signs of brain changes, then the patient could opt for medicine and preparation earlier, which is critical to slow down the deterioration.

Medical imaging is an important modality for disease progression analysis and health monitoring. For example, during regular health screens, patients take examinations such as chest X-rays, and abdominal ultrasonography. If the technician finds any abnormal changes in the body, further tests would be arranged. Another example is the Rheumatoid Arthritis (RA) classification, where the doctor needs to examine hand X-ray or foot X-ray

to determine the severity of RA. These processes all call for specially trained personnel to examine with great care. Computer vision workflows and machine learning models have a high potential to improve both efficiency and quality, assisting doctors to find and locate potential lesions.

1.1.2 Deep learning advancements in computer vision

Deep learning methodologies have dominated computer vision research nowadays, and outperform humans in various benchmarks. In our society, many duties or jobs can be partly or wholly formulated as task-specific computer vision problems. For example, facial recognition in the cell phone unlocking function is based on computer vision and machine learning techniques, and the complex process can be decomposed into visionary information extraction, feature registration, characteristic comparison, and identity determination. Another example is Optical Character Recognition, where machines can instantly transform text-containing images into sequences of words and sentences, ready for language translation and other instructions. In both face recognition and OCT tasks, deep learning models can work reliably with high accuracy.

Deep learning techniques for computer vision have seen rapid growth over the past 15 years. Though the neural network, backpropagation, and convolutional layer have been proposed for more than 30 years (before 1990), machine vision does not evolve into being widely applicable until the dramatic advancement of information technology introduced by the Internet and new computing hardware. The democratization of consumer cameras and the exponential growth of picture sharing on social media make it possible to create

large-scale image datasets, such as ImageNet (2008), and MS COCO (2014). The fast advancement of General Purpose GPU and computational libraries (CUDA toolkit, cuDNN library, Torch, Caffe, etc.) significantly reduces the technical difficulties for the wide adoption and research of deep learning. Harnessing all the hardware, software, and knowledge in deep learning communities, entry-level engineers can train and deploy powerful machine learning models to solve complex tasks, which is unimaginable before 2010.

The rapid development of deep learning is based on several new components which solve long-standing challenges. Network layers (convolutional, dropout, ReLu, skip connection, etc.) help solve the computational problems in neural networks, such as local pattern learning, gradient vanishing, and overfitting. The loss functions (weight decay, Kullback–Leibler, Dice similarity, etc.) enable improved training processes and training goals. Novel training schemes (unsupervised, semi-supervised, generative adversarial, reinforced) have enlarged the scopes and capabilities of deep learning in solving computer vision problems. Deep learning models could solve many computer vision tasks that have defined scopes of patterns and logical relations. The architecture and training process of deep learning models closely depend on the data characteristics and task definition, and practitioners often need theory and experience to make the model converge and perform well.

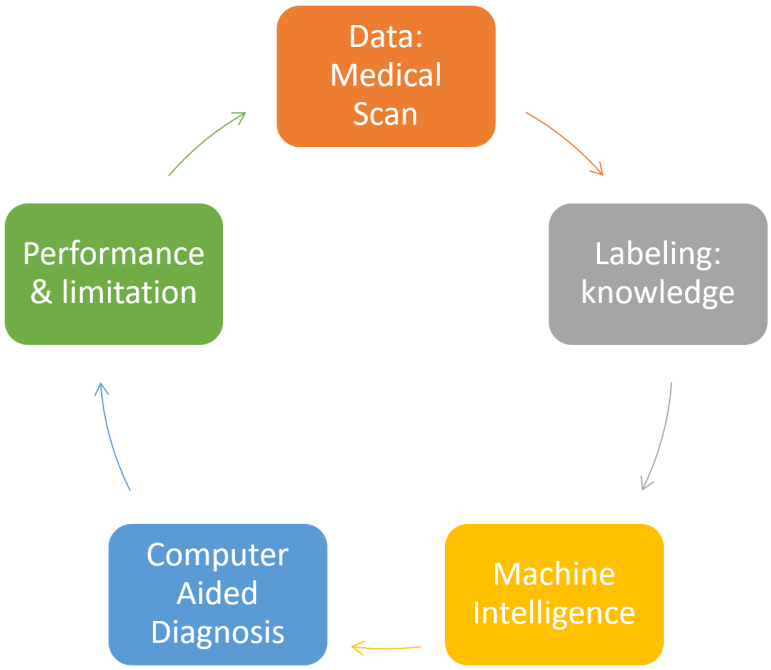

Figure 1.2: Procedures involved in medical imaging task. The data acquisition, labeling, model formulation, application integration, and clinical verification all require knowledge and consideration in the medical domain.

1.1.3 The formulation of medical imaging tasks

Medical image analysis differs from general computer vision in several aspects. In the medical imaging domain, the study is fixed on body parts, with limited types of imaging modalities which have standard scanning definitions. The main purpose of medical imaging is to learn the visual patterns of predefined lesions or diseases, while the purposes of general computer vision range from object classification, detection, and segmentation to action recognition, and behavior analysis. Though the task is much simpler in medical image analysis, the requirements of precision and interpretability are much higher. Cross-hospital verification of the model performance is always demanded before it goes into clinical deployment. In many scenarios, machine learning models serve as facilitating or supporting tools, and it is the doctor who makes the final decision.

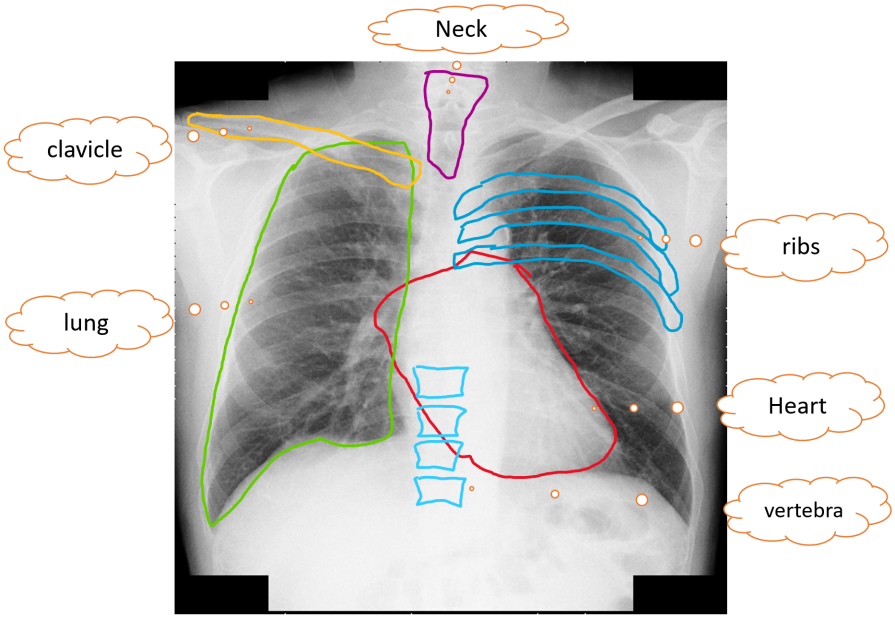

Figure 1.3: Chest X-ray covers multiple organs, and bones, which can serve many purposes in the screening scenario. Radiologists can examine if any part goes wrong. Many anomalies can be spotted, such as rib fractures, cardiomegaly, compression fracture in the spine, and lung nodules.

Data collection and distribution regulations are of special concern for medical imaging tasks. Due to the intrinsic nature of body scanning, data privacy concerns make it hard to collect large-scale or cross-regional datasets, which is different from other vision tasks. Considering that only some portion of hospital patients give permits for medical image usage, the collected data may not fully represent the disease distribution in the whole population. Depending on the medical purposes of each study, the task definition can have large variations from the doctor, the engineer, and the implementation. Based on private data and specifically defined tasks, deep learning models may not be directly comparable between different studies.

The goals and metrics of medical imaging tasks are within the scope of the related radiology practices and may vary in different organs or body parts. In the screening

scenario, an X-ray or CT image usually covers a large region of the body which makes it possible to detect any occurring anomaly in multiple organs or parts. The medical imaging models, therefore, are expected to segment the organs first, then classify each organ as normal or abnormal. An example is the chest X-ray imaging in Figure 1.3. Upon finding irregular changes, the patient can take further examinations under the guidance of physicians. In diagnosis scenarios, machine models are expected to not only detect the illness but also make finer-level predictions of the disease-incurred changes. For example, a lung lesion detection model should be able to determine if lesions exist, tell the lesion types, and summarize the lesion statistics. When the prediction is not certain, probabilistic interpretation should accompany the result. In the longitudinal scenario 1.4, researchers want to train machine learning models based on the medical history and current examinations of the patients, to forecast future disease development. In order to predict disease evolution, the model workflow needs to incorporate a larger learning context, including the joint representation of different modalities, the temporal changing pattern of the disease in the general, environmental, and genetic specification of the patient, and the effects of medical treatments.

1.1.4 The critical aspects for medical image analysis

Data labeling quality plays an important role in lesion segmentation. There are several aspects of labeling quality. As organs would have many associated diseases, medical experts with professional knowledge are needed to determine lesion types. For some datasets, the labeling process should be guided by radiology and pathology reports,

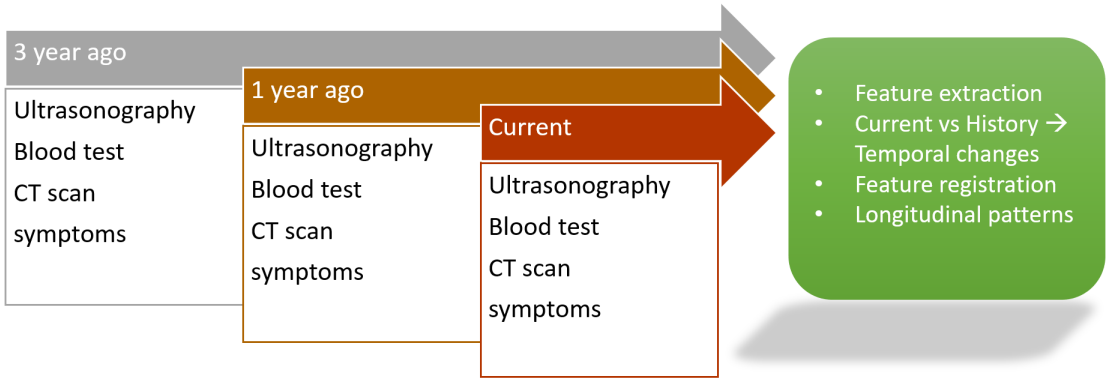

Figure 1.4: The longitudinal study can be designed in various ways. For some chronic diseases, periodic monitoring of body changes is necessary. For example, older patients with Hepatic B Virus are advised to take the examination of the liver at intervals by the doctor, multiple modalities (including ultrasound, blood test, and sometimes the CT) are needed to monitor the liver status. By comparing with previous health records, doctors can prescribe treatments. If some lesion exists, comparing historical measurements (size, density) can help with lesion classification and progression prediction.

which provide insights from radiologists and histologists. Sometimes human eyes could not distinguish lesions by checking the medical images, but representative visual patterns may actually exist. Given correctly labeled lesion masks or bounding boxes, deep learning models can pick up such subtle appearance nuances to store in the embedded features.

The metrics for medical image analysis come from different perspectives. From a pure computer vision point, lesion-level recall, accuracy, False Positive (FP), and False Negative (FN) are usually desired. The segmentation tasks also compare the performance of Dice coefficients at different levels. In disease screening settings, patient-level metrics such as sensitivity, and specificity are more important, since the primary goal is to early detect harmful diseases. Besides, there are many technological metrics to evaluate prediction quality, such as Pearson correlation coefficients, Area Under Curve, and Mean Squared Error.

Different from other vision tasks, medical imaging models go through strict scrutiny

before clinical deployment. Restrained by data availability and engineering flaws, machine learning models usually have many limitations. Without training and validation on samples from multiple medical centers on a large scale, models would not be considered fully representative or robust. Even when we restrict the application scenarios to specific goals for certain populations, third-party verification is necessary. As a facilitating tool for screening and diagnosis, deep learning models would function as effective helping hands for medical practitioners.

1.2 The overview of medical image tasks in this thesis

Having introduced the general concepts for medical image analysis as well as the related computational tools, we will take a tour of the main tasks in this dissertation. These three medical imaging tasks are all based on close collaboration between deep learning engineers and experienced hospital doctors. The trained models in the experiments have been delivered to doctors for larger-scale verification. The implementation principles not only consider engineering novelty but also lie in consistency with clinical requirements. In each task, all the relevant details are published. More details can be found in individual chapters.

1.2.1 Bone Mineral Density estimation from chest X-ray images

Osteoporosis is a metabolic disease widely affecting older people, characterized by extremely low Bone Mineral Density (BMD). In 2010 adults aged 50 years and older in the US have an overall 10.3% prevalence of osteoporosis, and it is estimated that 10.2

Category	BMD	T-score	Description
Young 30	---	1	Reference=1
All People	---	0	Average=0
Normal	---	(-1, 1)	Low Risk
Osteopenia	---	(-1, -2.5)	Med Risk
Osteoporosis	---	< -2.5	High Risk

Figure 1.5: The bone mineral density status defined by the World Health Organization. The DXA machine measures the density at the hip or lumbar vertebra to get the BMD scores, varying by the device manufacturer. The vendor-dependent BMD scores are normalized into the standard T-score range, which can be used to determine patients' bone density status.

million older adults had osteoporosis [3]. The overall low bone mass prevalence was 43.9%, and it is estimated that 43.4 million older adults had low bone mass. Symptoms include back pain, loss of height, and fatigue, but people tend to ignore these symptoms due to a lack of awareness. As a silent disease, osteoporosis can cause serious injuries before the patients finally get aware. The bones become fragile when the mineral density is below a certain threshold, which may easily lead to bone fractures.

In current clinical practice, Dual-energy X-ray Absorptiometry (DEXA/DXA) is used as the gold standard of bone mineral density. The specially trained operator would navigate the scan on the lower spine and hips. The DXA BMD can be used to diagnose osteoporosis or osteopenia. However, due to the low availability of DXA services and low awareness of bone mineral loss, DXA services are not adequately performed around the world. So it would make a big difference if low bone mineral density conditions are detected in regular health screening. Many non-DXA BMD measurement methods have been proposed and analyzed, but currently, there are still no clinically verified applications

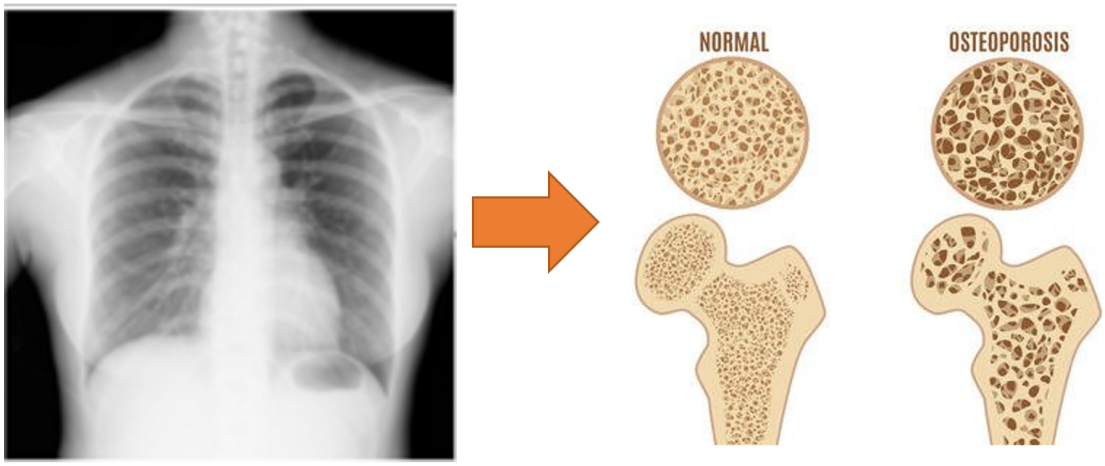

Figure 1.6: The purpose of the chest X-ray BMD project is to investigate if we can predict bone density status from the plain film.

deployed in scale. What's more, many of the non-DXA methodologies suffer from critical drawbacks.

One effective way to improve the health of the whole society is to screen disease during opportunistic scanning. As the economy develops, normal body checkups become increasingly accessible to all age groups. Chest X-ray is widely available, due to its low radiation (0.1 mSv) and low cost. If we can utilize chest X-ray images to predict BMD, many osteoporosis or low bone mineral density cases can be discovered in the early stage before causing serious troubles. Previous works have shown X-ray images in the hip or lumbar spine region could be used to predict BMD with high correlation, and it is highly likely that chest X-ray images contain usable bone mineral density information.

Next, we need to identify the challenges associated with the chest X-ray BMD prediction task. The whole chest contains many anatomical structures, such as shoulder bones, clavicle bones, neck bones, spine vertebra, ribcage, and many organs. Some patients have implantation inside the upper body, and some patients suffer from diseases

that alter the organ textures. Due to the variations from the scanning device and actual operation setting, chest X-ray image quality would vary accordingly. In order to handle the geometry variations, bone localization, and normalization are needed. In order to be more robust to all kinds of diseases, regional and global information should jointly make reliable predictions. To make the task simpler and practical, quality assurance before the model training is also necessary.

Based on the above task challenges and technical requirements, we develop the Attentive Multi-ROI model to predict BMD from chest X-ray images. We first train a graph convolutional neural network (GCN) for bone localization in the chest. The localized bone key points (14 points at regional bone centers) are then used to crop bone patches. Secondly, we adopt the VGG16 architecture to extract features. Thirdly, we explore the correlations among local bones to generate the global feature through transformer modules. During model training, multiple regions of interest (ROI) are utilized to predict BMD and calculate the loss. During inference, only the predicted BMD from the global feature is used. We also experiment with other model variations, to verify the module functionalities.

We collaborate closely with doctors for data collection and result analysis. We conduct extensive experiments to compare performances from different models. The proposed model achieves 90% sensitivity and 90% specificity for osteoporosis classification. The model predicted BMD has a strong correlation with the ground truth (Pearson correlation coefficient 0.894 on lumbar 1). For model verification purposes, we develop a browser-based interface for collaborating doctors to easily upload and process chest X-ray images. Without complex operations on the running server, the doctors can send inference tasks

remotely through web browsers.

1.2.2 Spine vertebra localization and identification via CT

The human spine is an important structure for body motion. The sequential vertebra and surrounding mussels are responsible for a variety of gestures and action support. The spine also protects the spinal cord, which is part of the central nervous system. The spine diseases include injuries, infections, and aging-related bone changes. Orthopedists need to first localize and identify the vertebra in medical image scans for further diagnosis. Fully automatic vertebra localization and identification would benefit vertebra segmentation results as well. The labels in segmentation masks can be re-assigned from accurate vertebra centroid and identity. With precise segmentation and improved detection of vertebrae, doctors can conduct examinations more efficiently.

The movable part of the human spine has 24 vertebrae, including 7 cervical vertebrae (C1 - C7), 12 thoracic vertebrae (T1 - T12), and 5 lumbar vertebrae (L1 - L5). The sacrum and coccyx vertebra are fused together and could not move. The spine vertebra size increases as the index increase from top to bottom, bearing more and more body weight. There are obvious landmark structures in some vertebrae, such as C7, T10, and L5. But the other vertebrae are not easily recognizable because of the similarities with the neighboring vertebrae. Due to the vertebra similarities in terms of bone structure and contexts, machine learning models may assign wrong identities.

There are many challenges for vertebra detection in CT images. Besides the similarities existing in the human spine vertebra, scanning variations and patient conditions also pose

difficulties. Sometimes patients only scan a small part of the spine, to avoid excessive radiation. A small field of view contains fewer landmark structures, which can be used to anchor the vertebra sequence. Some patients have metal implantation or severe spine curvature, which impedes correct localization. The scanning device may also introduce noises or biases.

To address the above challenges, we put forward the anatomy-constrained optimization method to localize and identify vertebrae with high accuracy. We utilize the 3D U-Net to get the voxel-wise probability maps for each vertebra in the first place. Then we aggregate the vertebra centroids together in the 3D space to find the spine line. Afterward, we transform the 3D vertebra centroid probability maps into 1D signals, retaining the spatial distances between vertebrae. In the straightened signal line, vertebra centroids and discs are represented as peaks and valleys. A final sequence of vertebrae can be obtained from the normalized 1D signals. By modeling anatomy constraints explicitly, the optimization process would find out a viable solution that is both physically reasonable and robust to scanning challenges.

We evaluate the proposed method on a public benchmark and achieve state-of-the-art performance. Through visualization and comparison with ground truth, our predictions are correct most of the time. The failures occur when the CT scan is of a small field of view or there exist extreme spine curvatures which would cause ambiguities. Some people may have an abnormal number of vertebrae in the spine, and this would also cause uncertain outputs.

1.2.3 Liver tumor segmentation and detection from CT images

Liver cancer is 6th cancer by incidence but ranks the 3rd by cancer mortality. The liver is the largest organ in the human body and has many critical functions. It regulates the blood chemicals, involving multiple systems in the body. The liver has complex connections with other organs through vessels and ducts. The liver is immediately boarding the portal vein, the gallbladder, the spleen, and the stomach. Digestion or blood problems would unavoidably affect liver functionalities. As a regenerative organ, the liver has the ability to repair and renew, before irreversible changes occur. There are many factors contributing to liver anomaly, such as drinking alcohol, smoking cigarettes, obesity, and hepatic virus infection.

As an important organ in terms of both size and functionalities, liver diseases develop in various and complex ways, and they can go into several different stages of texture changes before the malignant tumors form. Common liver texture changes include fatty liver, fibrosis, and cirrhosis. There are many types of lesions in the liver, and the most common ones include cysts, hemangioma, and focal nodular hyperplasia (FNH). There are mainly three types of malignant tumors in the liver, namely Hepatic Cellular Carcinoma (HCC), Cholangio, and Metastasis (Meta, migrating from other organs). The lesions can appear anywhere inside the liver, next to the hepatic vessels or duct, on the liver boundary. Sometimes the liver may go through morphological changes, such as enlargement or shrinkage. These all add up to the detection difficulties for liver lesions.

The liver tumor detection task also suffers from data availability and labeling problem. Since the liver is very large and complex, it calls for a large amount of high-quality

liver CT images for training and validation purposes. As liver tumor labeling needs professional knowledge and pathology support, close collaboration with liver experts is essential. Due to privacy regulations and data sharing restrictions, there are no public liver CT datasets with pathology-verified multi-lesion labeling. Some datasets such as LiTS or MSD liver only contain one phase CT images without differentiation of liver lesion types. Therefore, we collaborate with a large regional hospital to curate a high-quality liver dataset with multi-lesion multi-organ segmentation masks. We develop the labeling procedures, guidelines, and customized labeling tools. In the end, we obtain 1633 patient cases, with four-phase CT images having segmentation masks of 7 organs and 6 types of liver lesions.

We develop the 2-stage liver tumor detection workflow which yields both high sensitivity and specificity for malignancy detection. We adopt the 3D U-Net for organ segmentation and lesion segmentation in the first place, which gives us the probability maps for individual classes. Doubling the lesion prediction intensity in the probability maps would increase the lesion sensitivity, at a cost of more false positives. To suppress the false positives, we train a dedicated lesion segmentation model which only feeds on patches in the liver region. To make the lesion model robust to all kinds of suspected liver textures, we conduct context-aware lesion augmentation, which randomly combines the lesion patch with non-lesion patches of the same liver. In this way, the lesion model learns the better ability to distinguish true lesions from suspected textures. The segmentation result of the first model and the lesion reclassification result of the second model are combined to form a consensus, which yields the best result.

We experiment and test our proposed working pipeline in both the four-phase setting

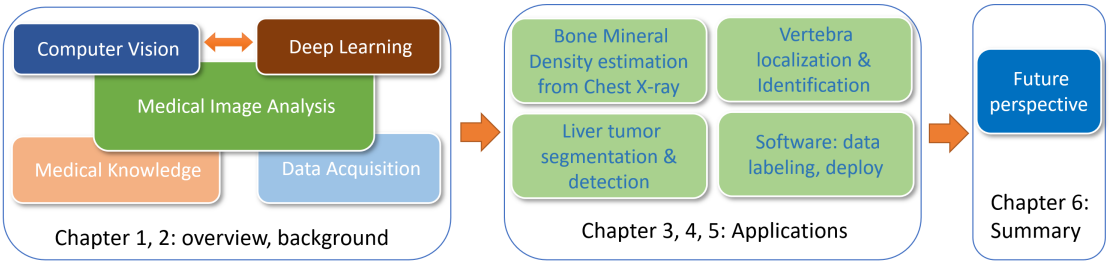

Figure 1.7: The arrangement of all the chapters. The first part gives the reader the necessary background, including medical image analysis and related technologies. The second part contains three projects, applying deep learning models to solve real-world medical tasks. The last part is the summary of the dissertation, discussing the future outlooks of medical image analysis.

and the non-contrast setting. In the diagnosis scenario, doctors would use all four CT phases to identify the liver tumor types, guiding further treatment. In the non-contrast or opportunistic setting, the patient goes for a regular body checkup or some preliminary examination, with the purpose of determining the malignancy's existence. Testing on 331 cases more than half of which have some liver disease, the proposed pipeline achieves promising results in both scenarios. The proposed pipeline can achieve 99% sensitivity and 97% specificity for malignancy detection at the patient level in the non-contrast setting, which holds strong clinical potential.

1.3 Outline of this Dissertation

In Chapter 2, we will take a closer look at the technology foundations for medical image analysis. Medical image analysis tasks originate from clinical reality and can be reformulated as computer vision tasks. Common computer vision principles would generally apply to medical image analysis tasks as well. However, due to the nature of medical scanning and privacy restrictions, there are many differences. Any computer-

aided diagnosis system needs to comply with medical imaging characteristics, for both ethical and technological purposes. Deep learning has revolutionized computer vision research in recent 10 years, and medical imaging tasks are also significantly improved. Given enough labeled data, many radiology tasks have been fully solved, though with some restrictions. Such examples include organ segmentation in CT images, pancreas tumor detection via CT, and so on.

In Chapter 3, we will delve into the chest BMD project. Osteoporosis or low bone mineral density affects more than half of the population over 65 years old. Low bone mineral density may cause serious injuries such as fragile fractures in old people, which can be life-threatening. If we can detect the low bone mineral density at an early stage and treat the disease as prescribed by the physician, bone mineral loss speed can be largely reduced. As the most common radiology scanning, chest X-ray images would make a big difference if BMD can be inferred from chest X-ray at certain accuracy. We will look at the necessary factors such as data collection, deep learning models, experiments, and result analysis to verify the assumption.

In Chapter 4, we look at the vertebra localization and identification problem. As a major bone structure in the human anatomy, spine vertebrae bear the upper body weights and function as an important mechanism for body actions. When the patient takes a CT scan of the lateral spine for the doctor to examine, the first step is to localize and identify vertebrae. However, due to the structural and contextual similarities, there can be ambiguities distinguishing the vertebra identities. What's more, it is desirable if the computer algorithms could automatically localize the centroid of vertebrae. With the centroid and identity of all vertebrae in the CT scan, it would be much easier for automatic

diagnosis of vertebra conditions.

In Chapter 5, we continue medical image analysis with CT scans but focus on the liver tumor detection task. Automatic liver tumor detection has long been sought after in clinical settings, which could help with the early detection of liver cancer and reduce the radiologists' workload substantially. However, there are many challenges due to complex liver conditions and lesion ambiguities. Short of public liver CT datasets, it becomes even harder. We will identify the characteristics of liver tumor detection tasks, investigate the challenges, and put forward our solutions. The whole pipeline of automatic liver tumor detection is discussed, including data curation, workflow design, 2-stage reliable prediction, and result analysis.

In Chapter 6, we discuss medical image analysis development opportunities and challenges. Computer vision and deep learning methodology have revolutionized vision-related tasks, and it is no surprise that medical image analysis would play more important roles in the clinical environment. Machine-backed automation will surpass human inspections of radiography images in many scenarios, and the challenges mainly lie in data-sharing regulations, diagnostic procedures, and clinical integration of computer-aided diagnosis. We should be confident that deep learning methodologies will benefit us all in near future.

Chapter 2

Deep Learning and Medical Image Analysis

2.1 Introduction

In this chapter, we take a look at the background of medical imaging analysis. Firstly we will start with the role of medical imaging in clinical settings. Secondly, we define the medical imaging tasks from a computer vision perspective. We will see the similarities and differences between general computer vision tasks and medical imaging tasks. Thirdly, We overview computer vision techniques and milestones. Lastly, we look into deep learning developments, with a focus on models, and datasets related to medical imaging tasks in this thesis. Now let us start with common types of medical scanning.

2.1.1 Medical imaging modalities

Medical ultrasonography Medical ultrasonography creates images of internal body structures to measure characteristics such as distance, velocities, and lesion presence. It is real-time, portable, low-cost, and radiation-free [4]. Even though it avoids the use of ionizing radiation, it has many shortcomings. Ultrasonography requires a specially

trained operator to actively inspect and cooperate with the patient, and it also suffers from noises or obscurity from bone or air. It has a limited field of view and is hard to produce stable and consistent images compared to X-rays. There are many configurations for ultrasound, such as operating mode and imaging techniques. Ultrasound examination is widely used as the screening tool for many diseases, such as liver tumors, stomach lesions, and soft tissues around bones. Once suspect lesions or anomaly is found, doctors may refer the patients to more accurate imaging modalities such as CT or MRI.

X-ray image X-ray is the most common diagnostic imaging modality. The X-ray machine sends out electromagnetic waves which pass through the body, and the film receptor on the other side captures the radiation signals [5]. Dense parts such as bone, and some lesions would absorb the X-ray more thus rendering the corresponding regions darker in the X-ray film. Conditions such as bone fractures and tumor formation can be detected if the anomaly is noticeable. Compressing a 3D body structure into a 2D imaging scan, the overlapping or cluttering pixels may leave out important details. Thus stereo scannings such as CT or MRI are usually required once the doctor finds suspicious scanning results. X-ray scanning on different body parts is used in various scenarios, with varied radiation absorption ratios. Radiation may affect vulnerable populations such as children and pregnant.

Computed Tomography Computed tomography (CT) produces a voxel-level image scanning of organs or parts by processing multiple X-ray measurements taken from different angles with tomographic reconstruction algorithms [6]. CT service become widely available

in the last 30 years, and about half of CT scanning in the US are contrast-enhanced which shows the lesions or organs with more clarity. 3D imaging scanning serves as an important organ or tissue examination tool for doctors in screening, diagnosis, measuring, and monitoring many diseases. However, a chest or abdominal CT can have as much as 100 times body radiation absorption compared to a chest X-ray. Low-dose CT or organ-focused CT may be preferred especially for the sensitive population.

Magnetic Resonance Imaging MRI produces 3D scanning of the body, without radiation. Instead, it utilizes magnetic fields and a special computer to take high-resolution pictures of the body part, which could show bones as well as soft tissues [7]. Having higher resolution and better visual qualities for soft tissues, MRI has advantages for examining subtle changes in the brain or soft tissues. But the complex operation and long scanning time may cause discomfort to patients. Therefore MRI is not as widely used as X-ray/CT in regular body screening. Instead, its advantages get explored in disease diagnosis, staging, and follow-up. Paired with corresponding configurations and scanning processes, MRI can be adapted to the best visual quality for many body parts.

2.1.2 Medical imaging analysis as Computer vision tasks

Common medical imaging equipment and trained staff are widely available in many countries, and radiology reading and interpretation are crucial for disease screening, diagnosis, treatment planning, and monitoring. X-ray, CT, and MRI are increasingly used to screen organs, bones, and soft tissues since they are usually accurate enough and affordable. The scanning process requires professional operators, who know the

steps and cautions. Afterward, radiologists read and explain the radiology images, to localize and determine lesions or organ changes. The radiology interpretation process, which is knowledge-intensive and time-consuming, can be costly. What's more, human interpretation may not be complete or consistent, due to subtle lesion variations and negligence. Therefore automatic machine interpretation with human-level accuracy is long sought after in practice. When computer-aided diagnosis equipped with competent machine learning models is deployed in hospitals for screening or diagnosis purposes, it boosts workflow efficiency and lowers medical costs, promoting affordable and high-quality health care.

Automatic intelligent medical image analysis can outperform human beings and bring in extra benefits in many scenarios. For example, some bone fractures only cause nuance changes in the chest X-ray, which is hard to find by the eye. Deep learning models can be trained to discover such underlying fractures and inform the patients. Another example is early tumor detection. Cancer is a major health threat in many countries, and a critical aspect of good treatment is early detection. Tumors are usually small during the early stage, and hard to notice during body screening with X-ray or CT images. Deep learning models could learn to detect small lesions or lesions on the organ margin. The radiologist could harness the automatic findings from the model-based segmentation pipeline, and re-examine the suspected lesions, to substantially improve the sensitivity and reduce the cost.

For some disease diagnoses, novel applications can be created with the medical image analysis models. For example, Rheumatoid Arthritis (RA) can be staged by the Joint Space Narrowing (JSN) degree in the hand X-ray images, requiring special orthopedics

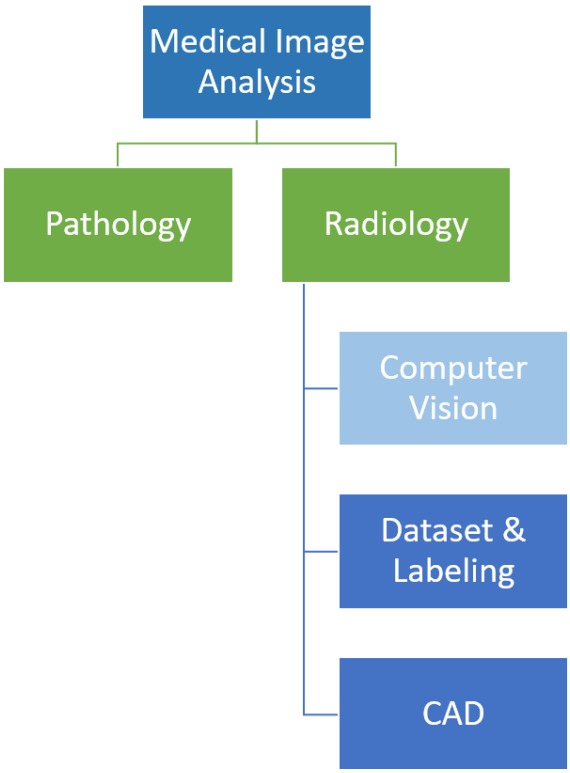

Figure 2.1: Medical Imaging Analysis depends on radiology scanning and may rely on the pathology report. Machine learning application in medical imaging analysis involves multiple steps such as data collection, labeling, and model development.

knowledge which limits the service availability. Machine learning models can learn to classify the JSN degree and make consistent predictions. With the help of a deep learning pipeline, machines can early detect or classify RA from regular hand X-ray images. Another example is predicting BMD from X-ray images. Normally, BMD examination is performed with the DXA machine, which is not widely available. Computer vision models developed with deep learning techniques can be utilized to predict BMD with adequate accuracy for osteoporosis screening.

There are many considerations when formulating medical imaging problems as computer vision tasks. Medical image analysis has many similarities and differences with general computer vision problems. Both of the task scopes include classification,

segmentation, and detection. Many well-studied computer vision methodologies apply to medical image tasks, such as semi-supervised learning, transfer learning, and image augmentation. However, collected as human body scans, medical image data only has small local variations in regional structures, texture, and lesions because of the similarity in human anatomy. The training aims to make machine learning models sensitive to these subtle details, which is essential to medical findings. So the pattern distinctions among labels are regional and in small magnitude. However, pattern clues in RGB images would lie in large and complex relationships in much larger space ranges. Based on the fundamental differences in patterns and purposes, pixel or voxel segmentation capturing local nuances is more crucial for medical image tasks. In this sense, the segmentation masks not only contain organ statistics but also provide flexible and detailed lesion detection results, out-competing the detection or regression results in many diagnosis scenarios.

2.2 The background and key factors for computer vision

Artificial Intelligence (AI) generally includes all man-made machinery, algorithms, and methodology that can handle complex tasks requiring memory and reasoning. Humans have long sought after automatic and intelligent machines that could accomplish complicated jobs to facilitate production and improve quality. As a key aspect of AI applications, Computer Vision (CV) deals with recognizing image patterns, such as classification, detection, and segmentation of defined objects, activity, and more targets. Computer vision develops alongside AI advancement in recent decades, from initial conceptualization in the 1960s to established traditional machine learning methodology in the 1990s, and to

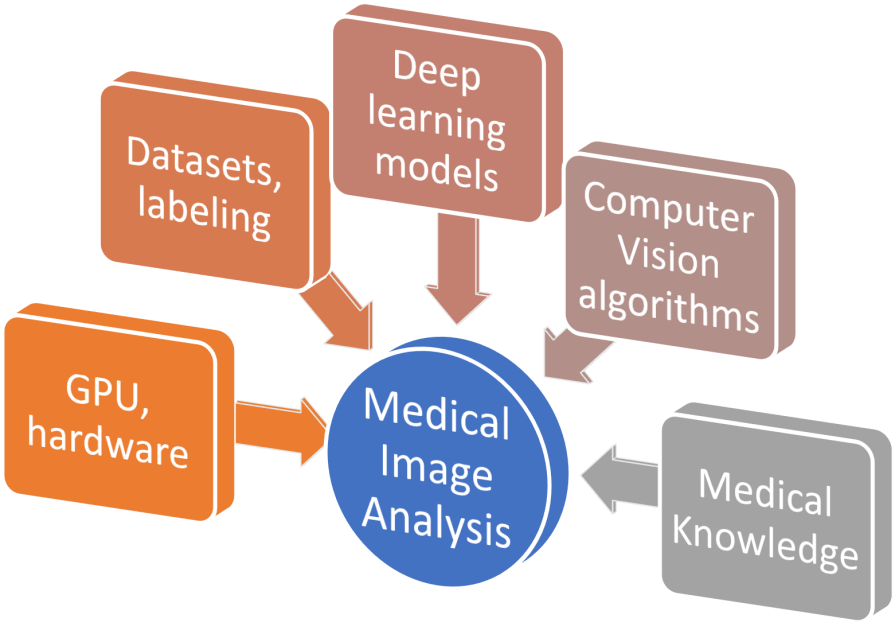

Figure 2.2: Medical image analysis factors. Researches and applications rely on the medical understanding of diseases, computing hardware, deep learning platform, image processing algorithms, and deep learning models.

Deep Learning (DL) era in the 2010s.

Same to the critical factors for any AI applications, computer vision advancements build on high-performance computing hardware, high-quality datasets, and improved machine learning algorithms. Before General-purpose computing on graphics processing units (GPGPU) was introduced at the beginning of the 21st century, machine learning algorithms run on the CPU only, which usually has a limited number of processors. The hardware limitation constrains the implementation algorithms for computer vision tasks, and algorithms need to consider both the time complexity and the space complexity. Due to the limited spread of sampling devices such as cameras, and medical radiology equipment, data collections are usually of small scale and simple label in many tasks before the 2010s.

With the advancement of computing hardware (faster CPU, GPGPU, CUDA), more

complex machine learning algorithms are experimented and implemented. In the social media era, images can be captured and shared more easily than ever. With the coming of Alexnet and following deep convolutional neural networks, pattern recognition enters the deep learning era. Researchers and engineers build and share the whole stack of technology components for computer vision, such as hardware computing facilities, fundamental libraries of mathematics and computing acceleration, software libraries for machine learning algorithms, deep learning platforms utilizing GPU, and pre-trained models. What's more, many institutions are willing to share data collection with labels, which makes it possible for anyone to build AI models.

2.2.1 Representative computer vision techniques

Researchers have developed traditional methods for computer vision applications for decades, usually with hand-crafted features, application-specific patterns, low-resolution inputs, and small-amount datasets. Traditional pattern recognition methods include hand-crafted feature matching, neural networks, support vector machines, principal component analysis, and decision trees. Depending on the task purposes, these methods may have different usages in image classification, registration, and clustering. But they all suffer from many limitations, including small image dimension, simple target pattern, too much hand engineering, little robustness, and limited generalization. The limitations come from many aspects, such as limited computing power, lack of modern computation theory and tools, and the difficulty of data acquisition. When technology development breaks up these ties, traditional solutions soon give way to deep learning models in recent 20 years.

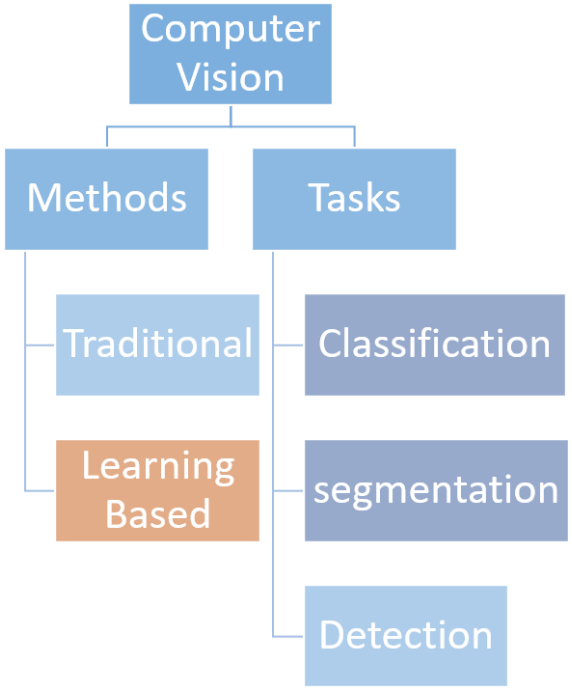

Figure 2.3: Computer vision tasks and methods. The traditional methods contain all the non-learning methodologies, such as hand-crafted feature matching, linear regression, SVM, and decision trees. The learning-based methods upgrade the model by adjusting parameters iteratively in an automatic way, to search for the best-fitting solution.

Inspired by the biological computation process, deep learning methodology implements parameter auto-fitting solutions with layered architectures. Relying on specialized layers (convolutional layers, ReLu layers, dropout layer, skip connections, etc.) and the optimized parameter updating schemes, deep learning models are much more adaptive and resilient than traditional ones.

The deep neural network is inspired by the functioning process of biological neurons in the vision system [8]. In the visual perception system of humans or other mammals, the retina transforms the light into neuronal signals, which can be further transmitted into the visual cortex. It still remains a mystery how exactly the brain processes visual signals, but it is reasonable to assume that there are many biological components dedicated to

specific functionalities. Humans learn to perceive still or moving objects for a long time during infant periods, and this knowledge accumulates in the visual system. Humans have more advanced visual recognition abilities than other animals because humans learn more about logical concepts beyond object appearances. It is also sensible to assume that our brain functions in 'sequential concept-specific layers', and the initial layers are in charge of points, color, distances, lines, and corners. The intermediate layers process the shapes, speed, and contours. The advanced layers recognize more complex objects, and actions, and associate them with brain memories or reactions.

Artificial Neural Networks have been proposed to solve practical problems but encountered many limitations in the last century. The backpropagation algorithm would suffer from gradient vanishing when the network becomes deeper. For computer vision tasks, linear mapping layers could not capture texture patterns due to a lack of computational locality. Before solving the visual pattern recognition and gradient vanishing challenges, deep-layer neural-like artificial systems are not possible. Researchers in the 1980s and 1990s begin to adopt the convolutional kernels in neural networks [9] [10] [11], which captures the visual elements in cascaded manners. With the advent of novel non-linear layers (ReLu) and more regularization techniques (dropout), very deep convolutional neural network architectures (AlexNet, VGG, etc.) become possible and popular.

2.2.2 Representative models, techniques, datasets for deep learning

In the past decade, deep learning models become dominant in general computer vision applications. Deep learning models such as AlexNet [**?**], VGGNet [12], and

ResNet [13] have achieved higher and higher accuracy on the ImageNet [14] challenge. Thanks to the deep learning platforms (Caffe, TensorFlow, PyTorch, etc.) and deep learning model zoos, an entry-level researcher could train or infer general vision tasks in a very short time. The flourishing developments of deep learning draw attention from all areas, increasingly producing large and high-quality datasets. Deep learning building blocks are summarized in Figure 2.4.

Regularization layers expand the modeling capacity and can help avoid overfitting or gradient saturation problems. Non-linear activation functions such as ReLu and its variants increase the modeling capacity by enabling deeper layers without severe gradient saturation challenges in the Sigmoid or Tanh functions. The dropout layer forces the model to become resilient against perturbations by randomly suppressing some neuron outputs during the training stage, which make more units to be effectively functioning. During the loss calculation, the weight decay reduces the complexity of a model and prevents overfitting by adding a weight regularization term.

Novel network architectures help solve fundamental computer vision tasks such as image classification, segmentation, object detection, and image generation. Take the segmentation networks as an example, the U-shape architecture and its variants employ the encoder-decoder workflow and shortcut connections to achieve pixel-level or voxel-level classifications. After passing through layers of mapping and transformation, the encoder layers learn a good feature embedding. The decoding layers extract the embedded information and combine it with corresponding shortcut high-resolution intermediate information of the encoding layers, to generate desired outputs. Encoder-decoder models have been widely used in image denoising, object or semantic segmentation, and variational auto-

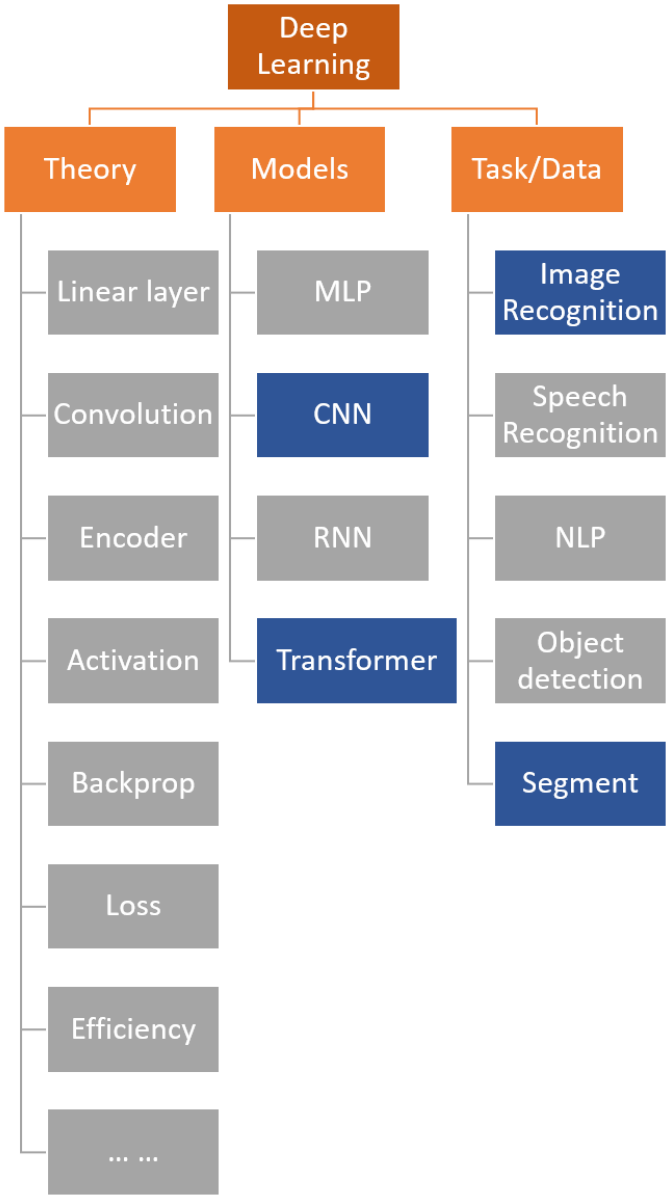

Figure 2.4: Deep learning models are ubiquitously adopted in computer vision, natural language processing, and speech recognition tasks. Although tasks differ by scopes, goals, and datasets, they share similar working principles, underlying components, and training techniques.

encoding.

New visual datasets and benchmarks promote technology upgrading and deep learning adoption. The ImageNet challenge contains more than 10 million images, which serves as a standard benchmark for all vision classification models. Model weights pre-trained

on ImageNet contain the representation knowledge, ready for reuse in other vision tasks through transfer learning. For example, we can keep most of the intermediate layers unchanged, and retrain the initial convolutional layer and the last classification layer, when transferring the representation knowledge learned in the ImageNet dataset, to other vision tasks such as the CIFAR-10 and Fashion-MNIST. Transfer learning not only reduces training time but also substantially improves the performance of tasks with limited data.

2.3 Applying deep learning in medical image analysis

2.3.1 Deep learning for medical image analysis

Medical image analysis plays a crucial role in many disease diagnoses. Ultrasound, X-ray, CT, and MRI take the scanning image of some body parts, to visualize abnormal changes inside. Traditionally, it requires medical specialists or radiologists' time and patience to inspect and determine, introducing the high labor cost. However, human judgments unavoidably bring in inconsistency and error. Machine-facilitated medical image reading or computer-aided diagnosis may significantly reduce human labor and improve the diagnosis quality, in terms of speed, cost, accuracy, and reliability.

Deep learning has been applied in the Computed-Aided Diagnosis of many diseases. For example, Chen-I Hsieh *et al.* [15] presents an automated tool to identify fractures, predict BMD, and evaluate fracture risk using plain radiography. Kang *et al.* [16] proposes a semi-supervised self-training algorithm to train a BMD regression model. Hoo-Chang *et al.* [17] investigate the CNN architectures and datasets for medical imaging tasks. Deep learning models have shown efficacy in lesion detection and classification for various

diseases [18]. One example is lung lesion detection, which draws a large amount of social attention during the COVID-19 outbreak. COVID-19 has many symptoms, and one of them shows in the lung. A deep learning model could take in the 3D thoracic CT image, and detect the lung lesions. The diagnosis could be affected by many factors, such as patient medical history, scanning setting, the lesion ambiguities. Trained on a large amount of human-labeled CT scans, deep learning models could help doctors identify lung parenchyma changes automatically and efficiently, which helps relieve the shortage of medical personnel.

2.3.2 Medical Datasets

In recent years, the medical imaging community has curated many public datasets and recognition challenges, which significantly speed up research activities. For example, Xiaosong *et al.* [19] presents the "ChestX-ray8" database, which comprises 108,948 frontal-view X-ray images of 32,717 unique patients with the text-mined eight disease image labels from the associated radiological reports using natural language processing. Antonelli *et al.* [20] organizes the MSD challenge, comprised of different targets, modalities, and challenging characteristics. The MSD challenge provides datasets for different organs, such as the brain, heart, hippocampus, liver, lung, pancreas, prostate, colon, hepatic vessels, and spleen. With the availability of high-quality datasets, researchers propose a variety of models and compare them against each other to boost technology improvement.

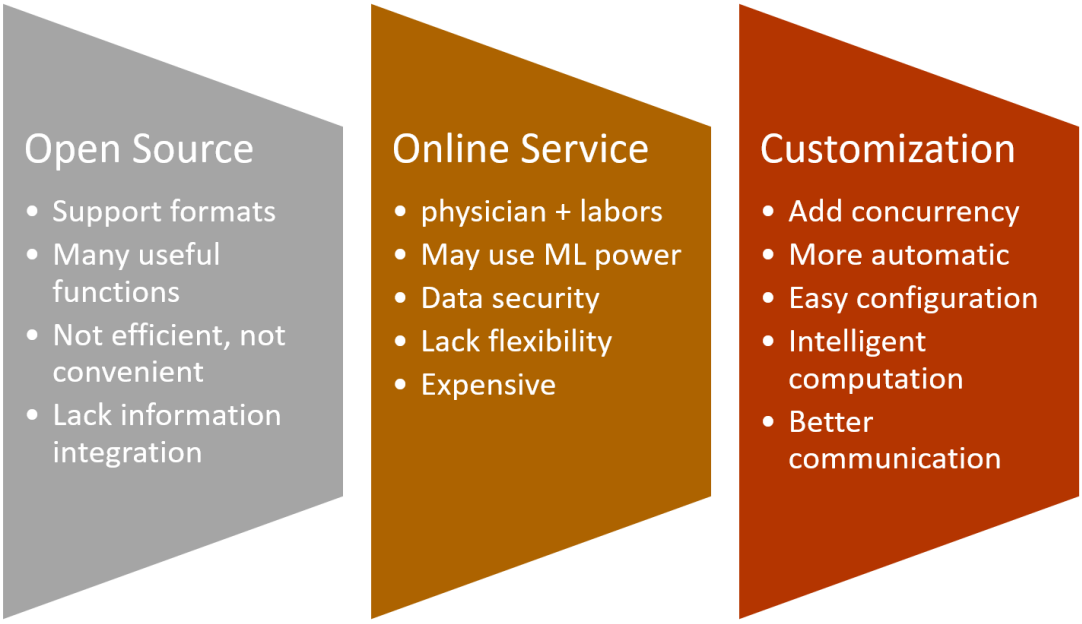

Figure 2.5: There are multiple choices for the labeling tool. We can utilize open-source desktop software, which is free and reliable. We can also pay for online labeling services to label medical images, which has advantages and drawbacks. When we are working on a large number of images, the most effective solution needs customization of the desired functions and features.

2.3.3 Data labeling

One of the central problems in medical image modeling is data labeling. It requires hundreds or thousands of patient records for a task, and it calls for both medical knowledge and engineering specifications to conduct the labeling. A data labeling or curation committee should set up the labeling targets, protocols, procedures, software tools, guidelines, the quality review rules. For a particular labeling task in an organ, the committee should define lesion types, characteristics, visual appearances, and exceptions. As there can be many rare or exceptional cases in any organ, explicit and complete labeling rules should be defined correspondingly.

As data labeling is a technical and laborious job, the software tool is crucial for both

efficiency and cost. There are different kinds of labeling platforms. Open source labeling tools (ITK-snap, 3D Slicer) usually support mainstream medical image formats (DICOM, NIfTI), and have integrated many useful functions (draw, erase, zoom, selection, multiple/3D views, statistics, display contrast). However, existing open-source tools lack many modern features, such as file management, auto loading, auto contrast, label-color configuration, concurrent display for multi-phase CTs, and intelligent mask editing.

Besides traditional desktop software, online labeling services for medical images are also popular. Online services (egg: iMerit - Medical Imaging) provide support from both professional physicians and trained labeling laborers. The experienced teams would develop labeling guidance and distribute the labeling work to individuals. Machine auto-labeling combined with human correction can effectively boost efficiency. However, labeling services also suffer from several drawbacks, such as data security concerns, high expense, and labeling errors.

To ensure high-quality labeling results, medical imaging tasks may need careful design in all the relevant steps. Due to the differences between targets and medical images, the labeling steps and actions may vary. It is best to customize the labeling tool for better workflow integration. If communications for review and discussion of labeling are needed, there should be convenient reviewing facilities. When a patient case has multiple medical images or masks, the software should load and display them simultaneously. The tool should also be able to load medical reports (radiology, pathology) automatically, for improved efficiency.

2.3.4 Limitations

Although it has been proved that deep learning models have outperformed human in many vision recognition benchmarks, machine models still has many challenges or limitations in medical image analysis. First of all, due to data sharing difficulties and the heterogeneous medical standards across the globe, it is hard to develop a universal medical model for particular diseases. Unlike general computer vision, researchers in medical image analysis are not able to share patient data or trained models freely. Models of the same purpose may be studied and trained independently in different hospitals. Secondly, the ground truth labeling for medical images needs to consider many aspects. People in different regions or communities may have varied disease distributions, and hospitals may provide various kinds of treatment based on patient history, financial status, doctor's professional ability, and so on. The labeling quality may have substantial variations even between doctors, which adds up to the difficulty of obtaining a large and uniform dataset. Thirdly, computer-aided protocols and regulations fall far behind technology development. Human doctors are trained for years in medical schools and hospitals, and they must get professional certificates before working in the diagnosis or treatment of diseases. As an emerging role, machine learning models have not been well defined in terms of application scopes, steps, limitations, and liabilities. Without adequate discussion and legislation, machine learning models would not be fully trusted or deployed on a large scale.

2.4 Human-computer interaction and the software tools

There are many steps and operations which require human-computer interactions in medical imaging tasks. During the data inspection, the human operator needs to see the medical image details and related summaries. For the abdominal scanning utilizing the contrast-enhanced CT images, the operators need to check through all CT scanning phases and compare for the differences to find potential lesions. For verification and failure analysis in the deep learning modeling process, the engineers need to visualize the image masks along with the CT image, and sometimes we need to simultaneously visualize several masks. During the mask labeling process, lesion information from the radiologist, biopsy analyzers, and disease experts is needed for accurate guidance. And the human-computer interaction software needs to present the information in reports, and lesion annotations. When clinical users try to gain realistic geometry outlooks of the scanned organs, the visualization software needs to present flexible 3D views of the corresponding masks of the organs, lesions, and bones.

2.4.1 Integrated labeling tool for CT images

In Figure 2.6, the labeling process involves many roles and experiment steps. Therefore the labeling tool needs to meet the corresponding requirements. To tackle the challenges, the customized software uses various components with integrated internal logic. It should incorporate many functionalities such as case management, simultaneous multiple displaying, clear mask showing, assistant mask computation, and convenient mask editing. To maximize the working efficiency, the labeling tool needs to carefully connect user-interacting operations

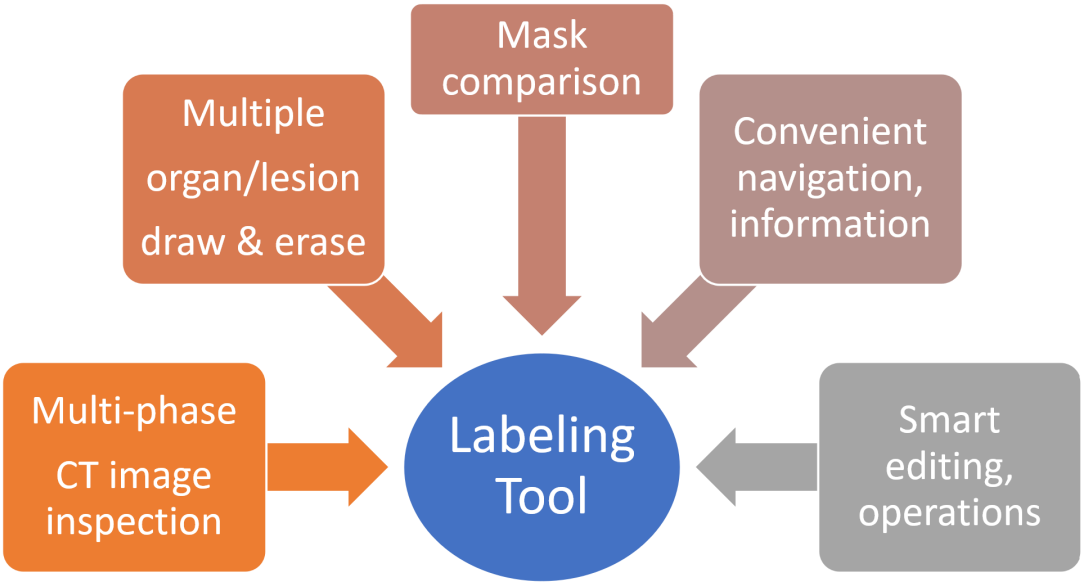

Figure 2.6: There are multiple requirements for the labeling process. A handy tool would effectively reduce human label, increase working efficiency and improve result quality.

with visualization components.

The software design takes the above requirements into consideration and the CT labeler in Fig 2.7 could satisfy desired specifications. The user could set the image path and choose the flexible mask path, enabling storage separation between the CT images (very large) and the mask files (relatively small, frequently transferred among workstations). The system could automatically load the clicked case and compute the connected regions in the masks. In the *Detailed Information* section, medical reports (radiology, pathology) could show up automatically. Other case information or labeling comments could also show up. When the user clicks on a *region* in the *Region list*, the displays automatically jump to the starting slice of the region with proper zooming scale. And region statistics (volume, density) also shows up in each display window correspondingly. In the *Visual settings* section, the user could configure CT images for four display windows, which are synchronized and maneuvered simultaneously. The user

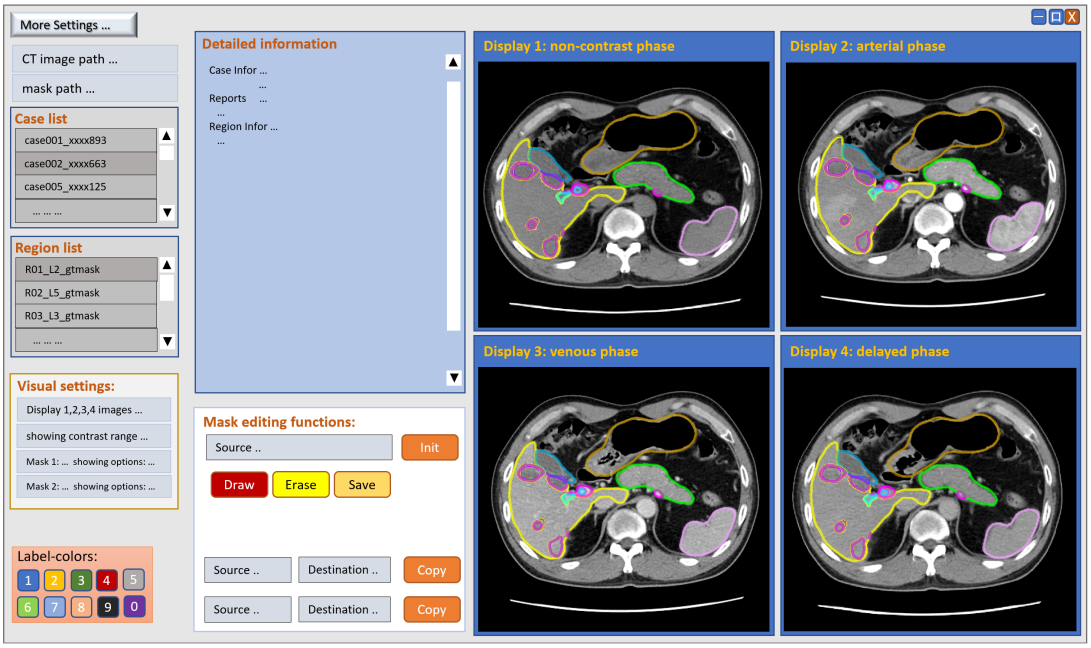

Figure 2.7: The CT labeler interface. The user can configure all the necessary settings conveniently, and the software makes the best effort to show information in an intelligent way.

could also set multiple masks, each with unique displaying styles (contour thickness, contour transparency, mask transparency). The labeling colors and names for different labels can be configured by a configuration file. The software features efficient mask editing abilities, such as mask initialization, contour-based drawing/erasing, and region-specific cross-mask copying.

The implementation of the labeling tool is based on Python and QT. PyQt [21] provides well-documented usages of GUI widgets and signal/slot mechanisms. Multi-threading is used for the CT image loading, and connected region computation, for a smooth user experience during case switching. The display windows can be configured to show up in separate windows, for higher display resolution. The medical reports are in *CSV* format, and the user can configure the showing columns conveniently. All the user configurations can be saved and loaded automatically.

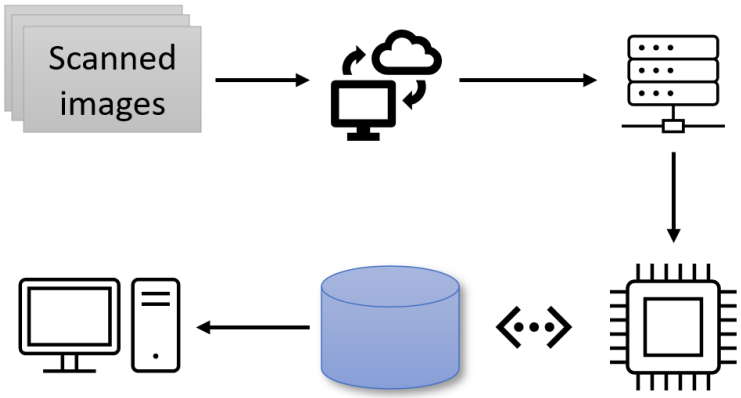

Figure 2.8: During clinical deployment, radiologists or doctors want to know the model prediction as soon as possible. The whole process includes network connection, data request, prediction computation, and result storage. The system should function smoothly with the least delay.

2.4.2 Web service for prediction result retrieval

In many clinical scenarios, it is desired that the model could return prediction results immediately after finishing the computation. During the anomaly detection scenario, the patient always wants to know the results instantly. Therefore, the result presentation should follow the computation process closely. Firstly, the medical images are sent to the backend server through network interfaces for data storage and inference-request registration. Secondly, the data pre-processing and the prediction model are invoked to serve the existing cases in the job pool. Thirdly, the prediction results are recorded and returned to the calling client through the network. Then the client side could conveniently show the machine predictions, in dedicated applications or web browsers. These steps are illustrated in Fig 2.8

One possible workflow for the web serving solution is depicted in Figure 2.9. The project manager receives HTTP requests and manages the project repository. The

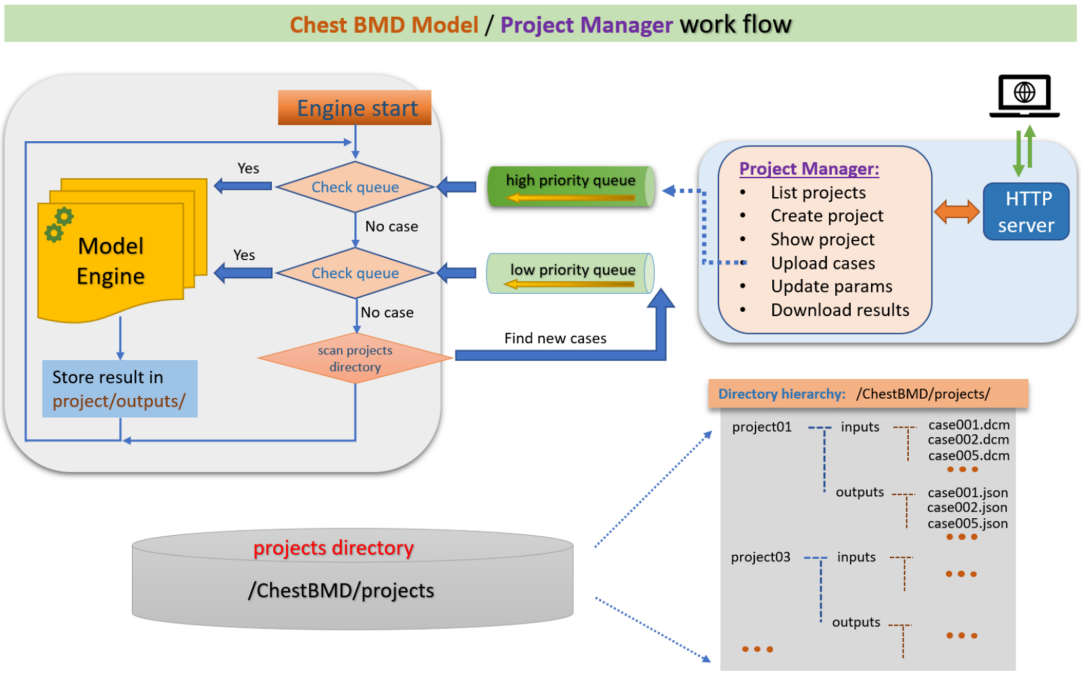

Figure 2.9: The proposed workflow of the backend server. The *Project Manager* works in a Docker image for more flexibility. The system has been shown to work properly for the *Chest BMD* project in the collaborating hospital.

operations are based on the internal storage hierarchy, listed at the bottom of Figure 2.9. Each *project* is isolated from others, which is useful for multiple users in practice. It provides a list of function interfaces, including *List projects*, *Create projects*, *Show project*, *Upload cases*, *Update parameters*, *Download results*, etc. The new-coming cases would be added to processing pools, where the model engine would constantly check. The prediction would be placed in the corresponding case folder, and the project manager could update the status and communicate with the calling client.

The HTTP service is based on Flask server [22] and Jinja templates [23]. The project manager and model engineer run in the same Docker image. The HTTP server can handle more than 1000 network requests per second, though the model engine inference would lag behind. The user could make requests through web browsers or dedicated

clients easily, and the results can be displayed as desired. In practice, multiple doctors could use the system concurrently without interference. It maximizes the utilization of computation resources while maintaining all the data and results in a structured way.

Chapter 3

Opportunistic Screening of Osteoporosis Using Plain Film Chest X-ray

3.1 Background

Osteoporosis is a common chronic metabolic bone disease often under-diagnosed and under-treated due to the limited access to bone mineral density (BMD) examinations, *e.g.*, via Dual-energy X-ray Absorptiometry (DXA). This paper proposes a method to predict BMD from Chest X-ray (CXR), one of the most commonly accessible and low-cost medical imaging examinations. Our method first automatically detects Regions of Interest (ROIs) of local CXR bone structures. Then a multi-ROI deep model with a transformer encoder is developed to exploit both local and global information in the chest X-ray image for accurate BMD estimation. Our method is evaluated on 13719 CXR patient cases with ground truth BMD measured by the gold standard DXA. The model predicted BMD has a strong correlation with the ground truth (Pearson correlation coefficient *0.894* on lumbar 1). When applied in osteoporosis screening, it achieves a high classification performance (average AUC of *0.968*). As the first effort of using CXR scans

to predict BMD, the proposed algorithm holds strong potential for early osteoporosis screening and public health promotion.

3.2 chapter Introduction

Osteoporosis is the most common chronic metabolic bone disease, characterized by low bone mineral density (BMD) and decreased bone strength. With an aging population and longer life span, osteoporosis is becoming a global epidemic, affecting more than 200 million people worldwide [24]. Osteoporosis increases the risk of fragility fractures, which are associated with disability, fatality, reduced life quality, and financial burden to the family and society. While with an early diagnosis and treatment, osteoporosis can be prevented or managed, osteoporosis is often underdiagnosed and under-treated among the population at risk [25]. More than half of insufficiency fractures occur in individuals who have never been screened for osteoporosis [26]. The under-diagnosis and under-treatment of osteoporosis are mainly due to 1) low osteoporosis awareness and 2) limited accessibility of Dual-energy X-ray Absorptiometry (DXA) examination.

Opportunistic screening of osteoporosis is an emerging research field in recent years [27–30]. It aims at reusing medical images originally taken for other indications to screen for osteoporosis, which offers an opportunity to increase the screening rate at no additional cost. As the most commonly prescribed medical image scanning, plain films' excellent spatial resolution permits the delineation of fine bony micro-structure that may correlate well with the BMD. We hypothesize that specific regions of interest (ROI) in the standard chest X-rays (CXR) may help the osteoporosis screening.

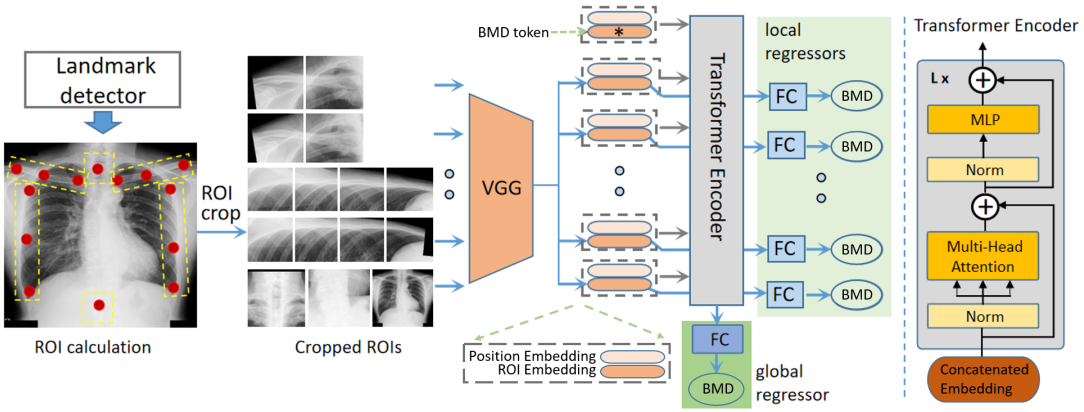

Figure 3.1: The proposed working pipeline. The landmark detector locates key bone points (red dots) on CXR images, then we crop and normalize 14 ROIs. These modalities go through a shared feature extractor (VGG16). The global regressor (dark green) works on the global feature generated by the Transformer Encoder [32] (grey). During training, results of both local regressors (light green) and the global regressor are used for loss calculation and backpropagation.

This work introduces a method to estimate the BMD from CXR to screen osteoporosis. Our method first locates anatomical bone landmarks and extracts multiple ROIs as imaging biomarkers for osteoporosis. Then We propose a novel network architecture that jointly processes the ROIs with learnable feature weight adjustment to estimate the BMDs. We experiment on 13719 CXRs with paired DXA BMDs (ground truth). This paper extends from a preliminary work [31]. In summary, our contributions are three-fold: 1) to our best knowledge, we are the first to develop models using CXR to estimate BMD 2) we propose the anatomy-aware Attentive Multi-ROI model to combine global and local information for accurate BMD estimation. 3) Our method achieves clinically useful osteoporosis screening performance.

3.3 Related work

3.3.1 Bone Mineral Density estimation and early screening

BMD examination via DXA machines is essential for osteoporosis determination and fracture risk assessment. In practice, bone densities of young adults are used as the reference where the standard deviations (SD) are used as the measuring unit (T-score) [33]. *1* unit of the T-score represents *1* unit of SD of the density, and *0* T-score represents the mean density of all the references. T-score at the spine, hip, or mid-radius lower than *-2.5* is considered osteoporosis. T-score between -2.5 and -1 is considered osteopenia, and T-score above -1 is considered normal. DXA is considered as the gold standard of care for osteoporosis screening, particularly for the aging population. However, it is underutilized for many reasons such as costs, availability and the requirement for experienced technicians. Although DXA is clinically well-validated as the modality of choice to measure BMD in many scenarios for diverse populations, DXA services are not widely available for general screening. The mere DXA-based diagnosis could not ensure the in-time bone quality evaluation for asymptomatic patients, since more fractures occur without reaching the severity of the osteoporosis [34] [26]. On the other hand, X-ray imaging (especially chest X-ray) is the most common radiological imaging examination broadly covering many asymptomatic patients, making it ideal as an opportunistic and preventive osteoporosis screening protocol.

Many works have discussed osteoporosis screening from non-DXA examinations. Firstly, the Quantitative Computed Tomography (QCT) in the abdomen or chest can

be re-used without additional radiation exposure or cost [35] [36] [37] [38] [39]. The *Hounsfield Units* of the QCT scans correlate well with DXA BMD scores for low BMD diagnosis [36] [40]. QCT has the advantage of the three-dimensional assessment of the structural and geometric properties of the examined bone [41], but QCT is not widely available. Secondly, the Quantitative Ultra Sound (QUS) based techniques have advantages of safety, low cost, operating flexibility [41] [42] [43]. X-ray-based examinations (DXA, QCT) are not suitable for the sensitive such as young children or the pregnant because of ionizing radiation. DXA or CT machines occupy extensive space and require specially trained operators, impeding general screening. QUS avoids these shortcomings, but QUS methods do not have standards on skeletal measuring sites, performance criteria, or normative reference data in the clinical setting [42].

Lastly, the plain film or X-ray, as the most common radiography examination, can also be utilized for osteoporosis screening. The existing techniques (DXA, QCT, QUS) all work on specific bone regions concerning the score from the areal or volumetric mass. The X-ray images however contain not only the bone textures but also other contexts. Since osteoporosis is a metabolic bone disease with complex manifestation, involving a larger context could capture the density relation which benefits BMD estimation. Hip X-ray-based BMD estimation has been verified in osteoporosis screening [44]. We investigate the chest X-ray-based BMD estimation with a focus on the input modality, model architectures, and prediction applicability.

3.3.2 Convolutional neural network and self-attention mechanism

Convolutional Neural Networks (CNN) have succeeded in medical image analysis [45] [46] [47] [48], partly because the hierarchical visual patterns echo the inductive biases learned by CNN layers. However, the inductive biases including translation equivalence and locality are less important for BMD pattern learning. The texture contrast among neighboring pixels and regions has more BMD cues. But the bare CNN backbones operate locally, failing to compare regional contexts [49] [50] [32]. Some papers exploit textual relationships by enhancing spatial feature encodings [49] or through channel-wise feature recalibration [51]. CCNet [52] proposes the criss-cross attention module to harvest the contextual information on the criss-cross path. LR-Net [53] presents the local relation layer (Local Relation Network) that adaptively determines aggregation weights based on the compositional relationships. GCNet [54] unifies the simplified non-local network [49] and SENet [51] into a general framework for global context modeling.

Inspired by the *Transformer* success in language tasks [55] [56] [57] [58], emerging works employ the *Transformer* modules to replace or facilitate the convolutional layers for visual tasks [32] [59] [60] [61]. The *Transformer Encoder* learns the global relationship through repetitive layers of *Multi-Head Self-attention* and *Multi-Layer Perception* operations. Attention Augmentation [50] augments convolutional operators with a self-attention mechanism by concatenating convolutional feature maps with a set of self-attention feature maps. The iGPT [62] train the transformer model [57] on pixel sequences to generate coherent image completions in unsupervised settings.

The CXR BMD task requires the model to capture both local textures and regional

relations automatically. A mindfully tailored combination of CNN and Transformer could harness their strengths. Convolutional layers can capture inductive biases such as translation equivariance and locality, while the transformer encoder enables global feature interaction. The purpose of the transformer encoder is to use the self-attention module to exchange information across different bone structures. Therefore, we employ both the convolutional feature extractor and self-attention fusion module in our proposed *Attentive Multi-ROI* model.

3.4 Methodology

3.4.1 Task Overview

In the opportunistic screening setting, the input is a chest X-Ray image. Our goal is to predict the BMD of lumbar vertebrae (L1, L2, L3, L4), alarming the patient of possible low BMD or osteoporosis conditions. Our hypothesis is that the BMD information lies in the CXR patterns, both in individual bone textures (local information) and in the overall combination of chest bone contexts (global information). Directly feeding the whole chest image into a CNN model for BMD prediction (**Baseline**) is intuitive and viable, but it lacks localized and contrastive bone information among the chest regions. The proposed pipeline in Figure 5.2 consists of chest landmark detection, bone ROI cropping, local texture extraction, global feature fusion, and BMD regression. We get the key landmarks for the chest from a Graph Convolution Network (GCN) model [63]. Then bone regions are cropped accordingly and fed into the proposed Attentive Multi-ROI model. The CNN layers learn local textures and patterns, extracting individual features. The Transformer

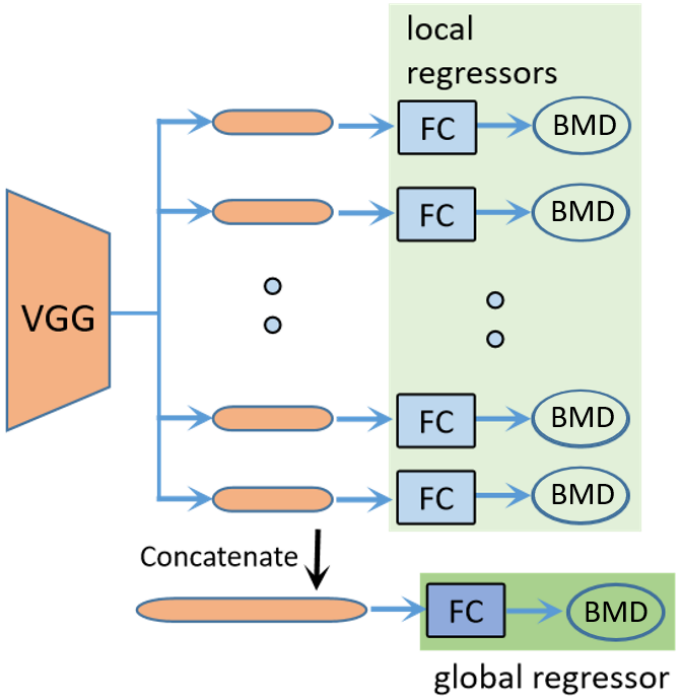

Figure 3.2: The plain fusion process in the Multi-ROI model. Individual feature vectors are concatenated as one before going through the global regressor.

Encoder refines individual representations to enable inter-regional feature interaction. At the end of the pipeline, local regressors and global regressors make BMD predictions on corresponding features. The model variants (in Fig 3.2 Fig3.3) are studied in Experiments section.

3.4.2 Automatic ROI Extraction

As the first step in the proposed pipeline, bone selection and region cropping prepare the model inputs. There are multiple bones in the chest area, bearing varied importance for BMD prediction. Although all bones provide density information due to their metabolic nature, the model should focus on the most effective regions. It is also unclear if the combination of distinct bone patterns is essential for this task. To learn representations of

the local textures and to explore the correlation among different regions, medical experts advise us to extract ROIs for clavicle bone, cervical vertebra, lumbar vertebra, and ribcage edges. We avoid the central part of the chest X-ray where cardiac or pulmonary diseases may significantly influence the appearance. In the end, our model works on the ROI croppings of left/right clavicle bones, cervical spine, left/right rib-cage area, and T12 vertebra.

We utilize the Graph Convolution Network (GCN) based Deep Adaptive Graph (DAG) [63] to automatically detect critical landmarks in the chest. We identify 16 landmarks in Figure 5.2, which include 1) 3 points on the left/right clavicles, 2) 4 points along the left/right rib cages, 3) 1 point on the C7 vertebrae, 4) 1 point on the T12 vertebra. We manually labeled 1000 cases (16 landmarks on each CXR scan) as the training samples for the DAG model. The resulting landmark detector can reliably extract all the key points. Given the key points for each bone, we crop the corresponding bone regions. However, different bones have distinct shapes and sizes, so we further sub-split the wider and higher ones. As seen in the cropped ROIs 5.2, there are 2, 2, 4, 4, 1, and 1 croppings for the left clavicle, right clavicle, left ribcage, right ribcage, cervical, and lumbar respectively. These arrangements are based on the bone size and width/height ratio. Besides these 14 local ROIs, we also include the whole CXR image as one modality. These 15 ROIs are resized and normalized before going through CNN layers.

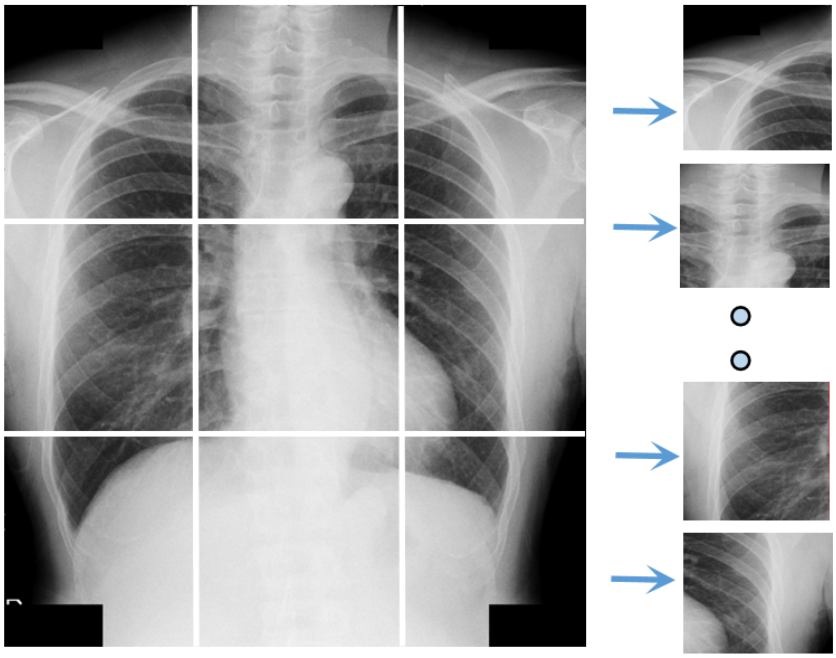

Figure 3.3: Patch generation of Multi-Patch model. The CXR image is split into *3*x*3* non-overlapping regions.

3.4.3 Hybrid architecture of convolution and self-attention

The feature extractor backbone is VGG16, and we extract local patterns separately for each bone region (ROI). Since there is little variation in the chest outlines among different people, the overall appearance alone is not enough to determine BMD. Instead, finer-level texture and pixel densities around the bones tell the distinctions. So individual ROI features are extracted independently. VGG16 is better than more complex backbones because its relatively shallow layers fit the simplicity of bone texture characteristics. We use *average_pool* to reduce the spatial dimensions on the VGG16 feature outputs, generating local representations $\mathbf{f}_i \in \mathbb{R}^D,\ i \in [1, N]$, N=15 is the number of ROIs, D=512 is the feature dimension. However, the local textures and pattern combinations could have distinct manifestations in different people for the same BMD value [33] [34]. To make

reliable density predictions, it should be addressed from the global level to account for the intractable variations resulting from diseases, scanning settings, and noises.

We employ the *Transformer Encoder* [55], which has been commonly applied in NLP and vision tasks [64] [58] [32] [65], to learn the global feature. The feature fusion process adjusts the individual features through layers of Multi-head Self-Attention (MSA) and Multiple Linear Perception (MLP) units, where the weighted relations are learned automatically [55]. Similar to *BERT*'s [CLS] token [56], we prepend a learnable [BMD] token embedding $\mathbf{E}_{bmd} \in \mathbb{R}^D$ as the target holder to increase robustness. In Equation 3.1, $\mathbf{f}_i \in \mathbb{R}^D, i \in [1, N]$ (dark orange ovals in Figure 5.2) is the CNN feature for the ith ROI, $\mathbf{E}_{bmd} \in \mathbb{R}^D$ (the dark orange oval with * inside) is the [BMD] token feature, $\mathbf{E}_{pos,i} \in \mathbb{R}^D, i \in [1, N+1]$ (light orange ovals) represents the learnable positional embeddings for *N* ROI and *1* [CLS] token, $\mathbf{z}_0 \in \mathbb{R}^{D\times(N+1)}$ (deep orange oval) is the concatenated embedding fed into the transformer encoder (grey box). The randomly initialized and learnable position embeddings $\mathbf{E}_{pos}$ are essential to keep spatial identity during the self-attention computation since there is no explicit sequential or grammatical order among ROI patches. In Equation 3.2 and 3.3, the alternating operations of MSA and MLP refine the feature representations. In our proposed model the encoder consists of $L = 6$ layers similar to [32]. Each layer consists of *Layer Norm*, *MSA*, *Layer Norm*, *MLP*. In Equation 3.4, the *mean* of (N+1) adjusted feature embeddings $\mathbf{z}_L^i \in \mathbb{R}^D$ are used as the global feature.

$$\mathbf{z}_0 = [\mathbf{E}_{bmd}; \mathbf{f}_1; \mathbf{f}_2; ...; \mathbf{f}_N] + \mathbf{E}_{pos} \tag{3.1}$$

$$\mathbf{z}'_l = MSA(LN(\mathbf{z}_{l-1})) + \mathbf{z}_{l-1}, \quad l = 1...L \tag{3.2}$$

$$\mathbf{z}_l = MLP(LN(\mathbf{z}'_l)) + \mathbf{z}'_l, \quad l = 1...L \tag{3.3}$$

$$\mathbf{f}_{global} = \frac{1}{N+1} \sum_{i \in [1, N+1]} \mathbf{z}^i_L \tag{3.4}$$

The core module of the transformer encoder [55] [32] is the *Multi-head Self-Attention*, illustrated in the *QKV* form in Equation 3.5-3.8. In Equation 3.5, $\mathbf{z} \in \mathbb{R}^{(N+1)\times D}$ represent the feature embeddings, the *query (**q**)*, *key (**k**)*, *value (**v**)* are the projection of $\mathbf{z}$ on matrix mapping $\mathbf{U}_{qkv}$. $k = 6$ is the *head* number, $D_h = D/k$ is the feature size in each head. In Equation 3.6, A is the weight matrix and A_{ij} represents the pairwise similarity between the ith and the jth features. In Equation 3.6 and 3.7, the Self-Attention (SA) module adjusts feature embedding according to weighted affinity. Regions of higher correlation contribute more, and others share less feature exchange. In this way, the individual feature becomes more robust to patient variations or noises. In the transformer encoder layer, the MLP consists of two linear layers with one non-linear *GELU* layer. The Transformer Encoder module uses constant latent vector size D=512 through all layers, the same size as the VGG16 feature. The Layer Norm (LN) and residual connections are applied before

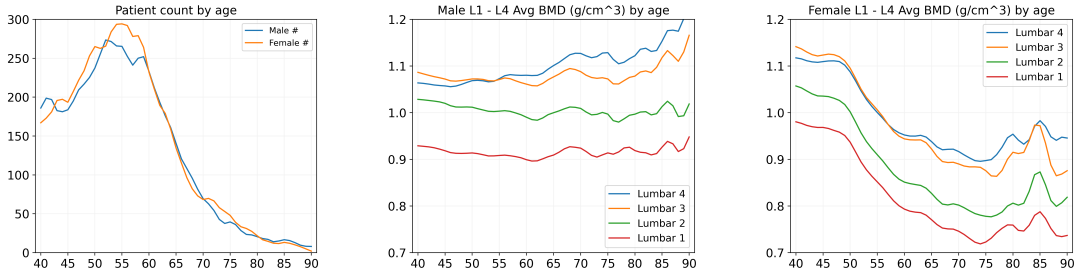

Figure 3.4: DXA BMD averages across ages for both genders. Different lumbar vertebrae (L1, L2, L3, L4) are drawn separately.

and after every block respectively, to keep the training stable.

$$[\mathbf{q},\mathbf{k},\mathbf{v}] = \mathbf{z}\mathbf{U}_{qkv} \quad \mathbf{U}_{qkv} \in \mathbb{R}^{D\times 3D_h} \tag{3.5}$$

$$A = softmax(\mathbf{q}\mathbf{k}^T/\sqrt{D_h}) \quad A \in \mathbb{R}^{N\times N} \tag{3.6}$$

$$SA(\mathbf{z}) = A\mathbf{v} \tag{3.7}$$

$$MSA(\mathbf{z}) = [SA_1(\mathbf{z}); SA_2(\mathbf{z}); ...; SA_k(\mathbf{z})]\mathbf{U}_{msa} \tag{3.8}$$

$$\mathbf{U}_{msa} \in \mathbb{R}^{k\cdot D_h\times D}$$

3.4.4 BMD Estimation via Joint Analysis of the ROIs

We apply *N* local regressors and *1* global regressor on the local ROI features $\mathbf{f}_i \in \mathbb{R}^D, i \in [1, N]$ and global feature $\mathbf{f}_{global} \in \mathbb{R}^D$ respectively for BMD prediction, represented as *FC* in Figure 5.2. All the regressors consist of two linear layers and one *ReLu* non-linear layer. We employ L2 loss on the predictions. The local regressions are active during training iterations to regularize feature representations but are ignored during evaluation, while the global regression is used all the time. During validation or inference, only the global BMD output is used. By jointly utilizing the global and local features in the chest and jointly promoting feature exchange through the transformer encoder, the network is capable of extracting BMD patterns on different scales for robust regression.

3.4.5 Implementation Details

We work on a workstation with Intel Xeon W-2295 CPU @ 3.00GHz, 132 GB RAM, and 4 NVIDIA Quadro RTX 8000 GPUs. Our models are implemented with PyTorch. The input images/ROIs sizes are set as *(256, 256)* by default for the best results. The training augmentations include scaling, rotation, translation, and random flip. The SGD optimizer has a learning rate of 0.0001, and a weight decay of 4e-4. All models are trained for 100 epochs. The four components in our model, VGG16 feature extractor, transformer encoder, local regressors, and global regressors, occupy 14.8M, 7.9M, 2.1M, and 0.13M parameters respectively, which sum to 25M parameters.

Table 3.1: Performance comparison. Our proposed (the Attentive Multi-ROI model, **Proposed**) performs the best of all. R-value, MAE, SD, and AUC represent Pearson correlation coefficient, Mean Absolute Error, Standard Deviation, and Area Under Curve respectively.

Model	L1			L2			L3			L4			Average	
	R-Val	MAE	AUC	R-Val	MAE	AUC	R-Val	MAE	AUC	R-Val	MAE	AUC	R-Val	MAE±SD
Base	0.859	0.069	0.952	0.87	0.078	0.96	0.86	0.08	0.966	0.823	0.091	0.963	0.853	0.08±0.066
MultiPatch	0.873	0.054	0.958	0.877	0.061	0.967	0.873	0.063	0.971	0.834	0.075	0.965	0.864	0.063±0.053
AttMultiPatch	0.885	0.052	0.964	0.888	0.059	**0.967**	0.882	0.061	0.969	0.846	0.072	0.964	0.875	0.061±0.051
MultiROI	0.883	0.052	0.959	0.887	0.059	0.966	0.879	0.062	0.972	0.837	0.074	0.966	0.871	0.062±0.052
Proposed	**0.894**	**0.049**	**0.964**	**0.899**	**0.055**	0.966	**0.887**	**0.059**	**0.972**	**0.855**	**0.069**	**0.968**	**0.884**	**0.058±0.05**

3.5 Experiments

3.5.1 Data collection

The data comes from Chang Gung Research Database [66], Chang Gung Memorial Hospital, Taiwan. We follow the Helsinki declaration with ethical permission number IRB-202100564B0 (The correlation between a chest X-ray and bone density). In the database, we searched patients with both DXA and CXR taken within 2 weeks from patients undergoing annual health checkups. The images are chest plain films in DICOM format where patient information has been removed to protect privacy. The DXA machine is GE Lunar, X-ray detector is Canon CXDI 710C. The CXR view is PA, the voltage is 115/120 kV, and the pixel spacing is 0.16*0.16 or 0.125*0.125 (mm*mm). We exclude unsuitable cases such as implantation and bone fracture by running quality assessment preprocessing steps in [44].

3.5.2 Experiment Setup

Dataset. We collected 13719 frontal views CXR scans, with paired DXA BMD scores (on four lumbar vertebrae L1 - L4) as ground truth. All experiments use the same data split, with 11024 and 2695 patient cases for training/validation and testing, respectively.

There is no patient overlapping between data splits. The model train-val/test for different lumbar vertebrae is conducted in four separate experiments. For a particular lumbar BMD model, it is trained using 4-fold cross-validation, with a train/validation ratio of 3 to 1. The ensembles of predictions from all 4-fold models on the testing set are reported as final results.

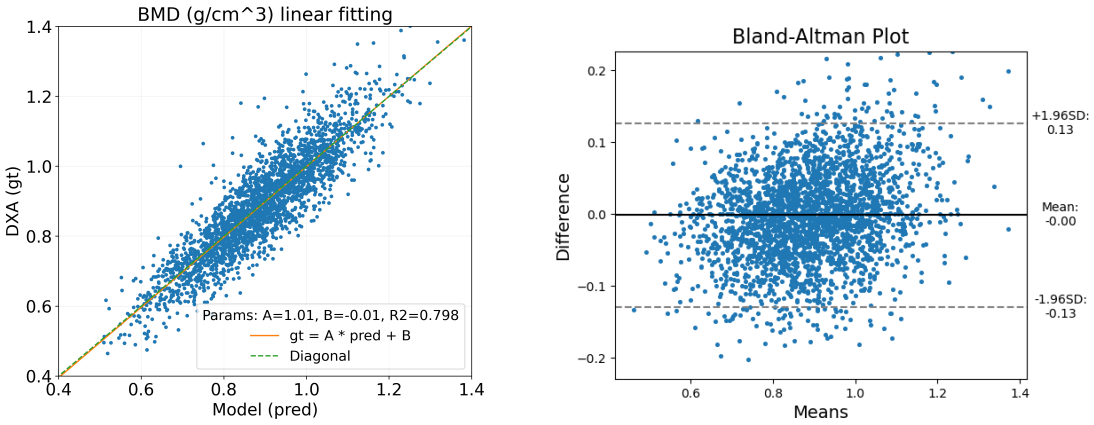

Figure 3.5: Illustrate the proposed model results on Lumbar 1 BMD, each prediction is compared with its paired DXA BMD value. In the linear fitting line (left), **A** and **B** represent **slop** and **y-intercept**, **R2** is the coefficient of determination. In the Bland-Altman plot (right), the horizontal axis is the mean, and the vertical axis is the difference between each pair.

3.5.3 Data distribution

3.5.3.1 Patient statistics by gender and age

Our data is from annual body check-ups in a regional hospital and the study population is limited to East Asian Han Chinese. The BMD-related characteristics may not represent the Taiwan general population. In our task setting, we only select patients aged from 40 to 90 since this age range is clinically relevant [33]. In Figure 3.4 (1), the data count peaks

between 50 and 60.

3.5.3.2 The mean BMD values and caveats

In Figure 3.4 (2)(3), each curve describes the average BMD values at each age vertebra-wise. Data above 75 years old has abnormal BMD mean and variance out of sample scarcity. Lumbar 1 has the smallest mean value at all ages. The bumps or uprisings in the BMD curves result from the data source (annual checkups) characteristics and the BMD distribution between neighboring age groups may not be consistent. Take female Lumbar 4 BMD values as an example, the 95% Confidence Intervals (**CI**) of patients aged 74, 75, 76, 77, 78, 79, 80 are 0.902±0.394, 0.882±0.449, 0.912±0.494, 0.913±0.288, 0.88±0.272, 1.003±0.55, 0.944±0.406 respectively.

3.5.3.3 Confidence Intervals by BMD status

The model utilizes only BMD values during training/inference, ignoring gender or age information. The 95% CIs in BMD range for Lumbar 1 *normal*, *osteopenia*, *osteoporosis* respectively are 1.0±0.17, 0.81±0.08, 0.65±0.1, for Lumbar 2 are 1.07±0.21, 0.85±0.08, 0.68±0.1, for Lumbar 3 are 1.14±0.23, 0.9±0.08, 0.74±0.11, for Lumbar 4 are 1.13±0.25, 0.88±0.08, 0.72±0.12. Notably, Lumbar 1 osteopenia (38%) and osteoporosis (11%) amount similarly to normal (51%) cases, while Lumbar 4 normal amount (70%) is much larger than osteopenia (23%) and osteoporosis (7%).

Table 3.2: The Attentive Multi-ROI model classification characteristics using different prediction thresholds. Ground truth osteoporosis uses T-score of -2.5 as the judging threshold. Prediction classification use either unified (-1.75, -2, -2.25, -2.5) T-score thresholds for all vertebra or *Flex* thresholds. *Flex* Thresholds are -2.2,-2.1,-2.0,-1.9 for L1,L2,L3,L4 respectively.

T-score	L1		L2		L3		L4		Average	
Thresholds	Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec
-1.75	93.9%	85.1%	91.7%	89.1%	93.4%	91.1%	85.0%	93.1%	91.0%	89.6%
-2	86.2%	91.0%	85.0%	93.0%	84.8%	94.8%	75.4%	95.4%	82.8%	93.6%
-2.25	77.6%	95.3%	76.3%	95.8%	72.0%	96.9%	68.9%	97.4%	73.7%	96.4%
-2.5	61.9%	97.5%	63.5%	97.9%	60.7%	98.3%	54.5%	98.9%	60.2%	98.1%
Flex	79.8%	94.7%	82.3%	94.4%	84.8%	94.8%	78.4%	94.4%	81.3%	94.6%

3.5.4 Performance Metrics

Our vertebra-level metrics include the *Mean Absolute Error* (MAE), the *Pearson Correlation Coefficient* (R-value), *Area Under Curve* (AUC), sensitivity, specificity, coefficient of determination (R squared or R^2) of the linear fitting curve, the standard deviation of the prediction errors. The patient-level metrics are sensitivity and specificity. MAE measures the averaged absolute differences between the predicted and ground truth. The R-value measures the linear correlation between the predicted and the ground truth, only considering the sequential correlation regardless of the absolute values. For osteoporosis classification, BMD values are transformed into T-score values by checking the transforming table in the DXA machine [33] [34]. In the T-score range, the AUC measures accumulated true positive (osteoporosis) rate under different judging thresholds for osteoporosis classification. The sensitivity and specificity are also for classification purposes. The linear fitting curve illustrates the general correspondence between prediction and ground truth. The coefficient of the determinant quantifies the fitting goodness. We also draw the Bland-Altman plot which shows the standard deviation limits and prediction error distribution.

3.5.5 Attentive Multi-ROI model performance (vertebra level)

Each lumbar BMD model is trained four times using a 4-fold cross-validation setting. The prediction ensemble of these models on the testing set is recorded as the final result in Table 3.1. The proposed model and its variants have substantially lower MAE than the Base model. The proposed model outperforms others by at least 1% in terms of R-value for all BMD tasks, which clearly demonstrates its superiority. While the R-value and MAE are error measurements, AUC scores evaluate osteoporosis classification ability in Figure 3.6. L4 has a larger AUC score because its osteoporosis ratio (7%) is much smaller than normal (70%) or osteopenia (23%), while L1 is the opposite. We show the sensitivity and specificity in Table 3.2. Applying the **Flex** thresholds, the model achieves high averaged sensitivity (81.3%) and specificity (94.6%).

To show the performance intuitively, we draw the linear fitting line and the Bland-Altman plot for L1 predictions in Figure 3.5. The *intersect* (-0.01, close to 0) and *slop* (1.01, close to 1) of the linear fitting line demonstrates the general correctness, and the *R-squared* (0.798) measures the closeness between predictions and the ground truth. In the Bland-Altman plot, value errors are drawn against value means for each prediction and DXA BMD pair. The relatively small standard deviation (0.065) and the concentrated scattering further prove the performance consistency. Outliers beyond the $\pm$1.96SD limits occupy a small portion of all. Linear fitting and Bland-Altman plots of other lumbar vertebrae lead to similar conclusions. Error analysis is in the Ablation section.

3.5.6 The patient-level osteoporosis classification

For opportunistic screening, the goal is to inform osteoporosis risks using a routine tool such as CXR. Although we predict the BMD of four lumbar vertebrae with four separate models, we aim to generate unified alarming signals. Therefore we calculate the patient-level osteoporosis classification performance. Each patient is either **normal** (all four lumbar T-scores are larger than -2.5) or **osteoporosis** (any lumbar vertebra has a T-score smaller than -2.5). The patient distribution is in Figure 3.7, where there are 2267 normal cases, and 390 osteoporosis cases.

Applying different prediction T-score thresholds, we calculate the sensitivity and specificity to evaluate the proposed model in Table 3.3. Referring to the proper vertebra-level thresholds in Table 3.2, we only compare four thresholds for the patient-level classification. These settings cover both high sensitivity (0.933) and high specificity (0.966) for osteoporosis. As a balanced configuration, the $Flex$ thresholds (T-score -2.2,-2.1,-2.0,-1.9 for Lumbar 1,2,3,4 respectively) achieves approximately 90% for both osteoporosis sensitivity and specificity. It also achieves 99.7% patient-level osteopenia specificity, implying picking out nearly no patients with healthy BMD. These results strongly support the practical applicability of our proposed model. Applying the $Flex$ thresholds in Figure 3.7, the model can pick out osteoporosis patients very well, with 75.2% **P1**, 92.4% **P2**, 97.1% **P3**, 100% **P4**.

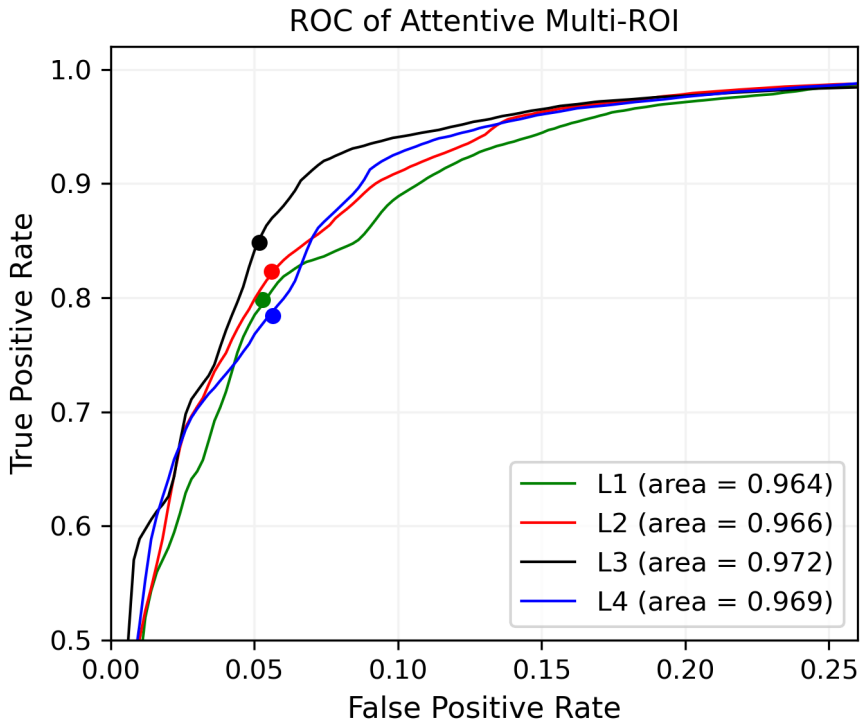

Figure 3.6: The Receiver Operating Characteristic (ROC) Area Under Curve (AUC) of the proposed *Attentive Multi-ROI* model for osteoporosis classification in L1, L2, L3, L4 experiments (only upper left AUC). The colored dots operate on **Flex** T-score thresholds.

3.5.7 The model variants

3.5.7.1 The Baseline model

Since the VGG and Resnet have been shown to work well on hip X-ray BMD estimation [44], they may also succeed in the CXR-based BMD estimation. In the *Baseline* model, we adopt VGG16 as the feature extractor, apply the Global Average Pooling (GAP) to reduce spatial dimension, use two linear layers with *ReLu* non-linearity in the regressor to predict BMD, and use Mean Squared Error (MSE) loss to train the model. Among different *Baseline* input candidate ROIs, the whole chest performs the best. Modalities such as lumbar ROI or cervical ROI get 2% to 5% lower R-value than the

whole chest ROI. In Table 3.1, the *Baseline* model serves as a reference for performance comparison.

3.5.7.2 The Multi-ROI model (MultiROI)

In order to investigate the effect of the *Transformer Encoder* in the proposed Attentive Multi-ROI model, we replace *attentive feature fusion* with *direct feature concatenation* in Figure 3.2 in the *Multi-ROI* model. The concatenated global feature has the length of *512*15*, and the global regressor now has a larger input dimension. The proposed (Attentive Multi-ROI) model has consistent advantages over the *MultiROI* (plain-fusion Multi-ROI) in Table 3.1, which demonstrates the positive effect of *Transformer Encoder*.

3.5.7.3 The Attentive Multi-Patch model (AttMultiPatch)

To investigate the benefits of precise ROI extraction in the proposed *Attentive Multi-ROI*, we replace the *landmark-based ROI extraction* with the *image patch splitting* in the *Attentive Multi-Patch* model (AttMultiPatch). We split the high-resolution chest X-ray image into evenly distributed patches in Figure 3.3. Though lacking the precise landmark-based cropping, the *AttMultiPatch* model is able to learn both individual patch details and inter-patch relations. However, the representative meaning of each patch is less certain due to the scanning variations in patient postures and body sizes. Comparing the *AttMultiPatch* and **Proposed** in Table 3.1, the landmark detection and precise ROIs benefit all four lumbar BMD tasks.

3.5.7.4 The Multi-Patch model (MultiPatch)

To show the effect of *Transformer Encoder* on *Attentive Multi-Patch* model, we train and test the *Multi-Patch* model which instead uses the *plain concatenation*. Their comparisons in Table 3.1 again demonstrate better feature fusion ability in the attention module.

3.5.8 Performance comparisons

To see the advantage of the *Multiple-Modality* inputs working flow in our four models (*MultiPatch, AttMultiPatch, MultiROI, AttMultiROI,*), we compare them with the *Base* in Table 3.1. The *Multiple-Modality* models extract not only global patterns but also detailed local textures, thus achieving better results. With simple patch splitting and attentive fusion, the *AttMultiPatch* model outperforms the *Base* model significantly in terms of R-value and MAE.

The four *Multiple-Modality* models differ by the *ROI extraction* component and the *global fusion* component. To see the component-wise boosting effects from the *landmark-based ROI extraction* and the *Transformer Encoder*, four models are compared to each other with the base being the *MultiPatch* model. Applying the *Transformer Encoder* only (AttMultiPatch) or applying the *landmark-based ROI extraction* only (MultiROI) contributes a similar amount of benefit (about 1% R-value boosting) to the *MultiPatch* model in Table 3.1. Applying these two simultaneously in our proposed *AttMultiROI* model leads to the best performance, with 2% R-value boosting. On the one hand, the precise ROI croppings in the *MultiROI* and *AttMultiROI* models enable more efficient

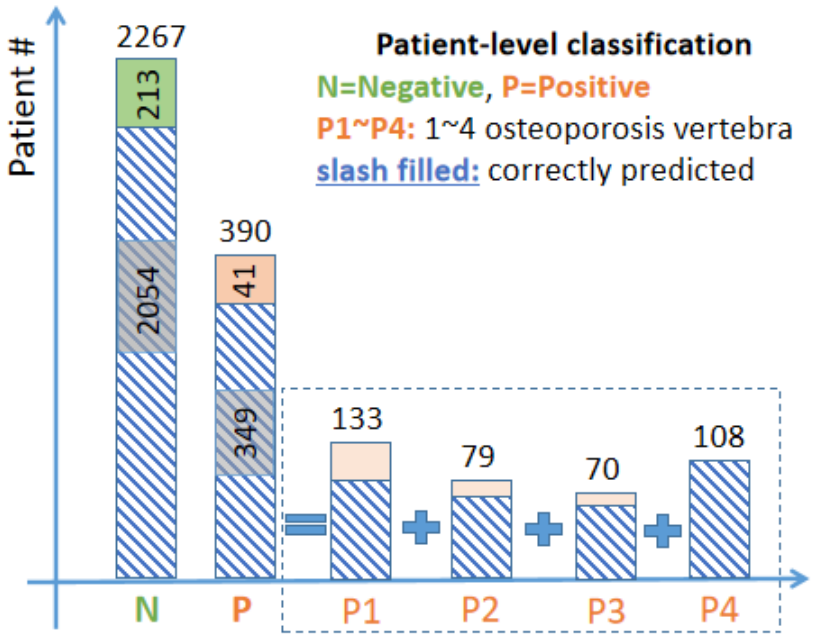

Figure 3.7: Patients with four lumbar records (2657 total) in the testing set are assigned as normal (Negative) or osteoporosis (Positive). The positive cohort can be further decomposed into four bins, according to the number of osteoporosis vertebrae. The shadowing parts are true negatives and true positives from the Attentive Multi-ROI model, applying $Flex$ thresholds.

local texture utilization. On the other hand, the plain concatenation in the *MultiPatch* and *MultiROI* models treat all the individual ROI features as equal which renders the model less robust to occlusion or noises, especially in case of implants or tissue consolidations. The *Transformer Encoder* adjusts the individual features in a learnable and flexible manner, addressing the correlations among chest bones, which leads to improved feature robustness.

Table 3.3: Patient-level sensitivity and specificity. Unified thresholds (-1.75, -2, -2.25) ignore the lumbar BMD differences, while $Flex$ (thresholds -2.2,-2.1,-2.0,-1.9 for Lumbar 1,2,3,4) is aware.

T-score	Osteoporosis		Osteopenia	
Thresholds	Sensitivity	Specificity	Sensitivity	Specificity
-1.75	0.933	0.873	0.453	0.995
-2	0.844	0.929	0.341	0.998
-2.25	0.736	0.966	0.255	0.999
Flex	0.895	0.906	0.392	0.997

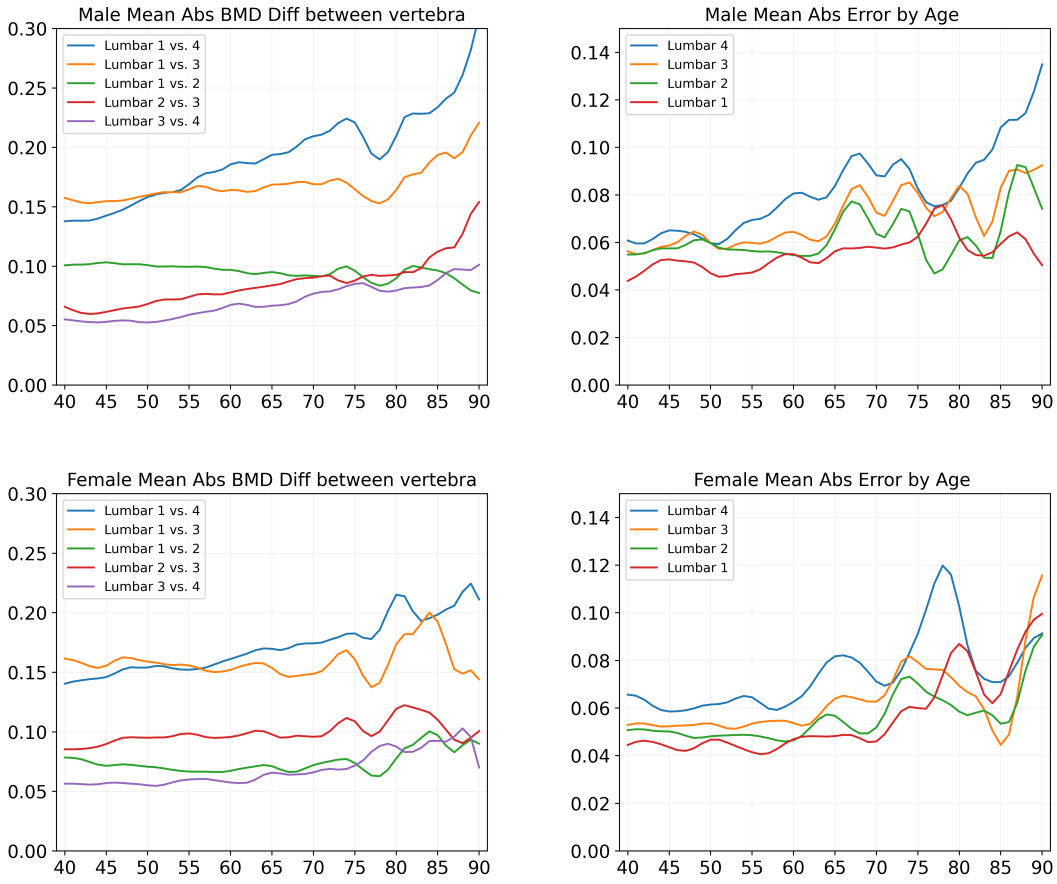

Figure 3.8: Cross-vertebra BMD differences (1)(3). Measure the mean of the absolute difference between vertebra pairs, using DXA BMD (g/cm^3). The mean of absolute prediction error (2)(4). The proposed model (Attentive Multi-ROI) predictions have a satisfactory error range.

3.6 Ablation study

3.6.1 Convolutional neural network backbone selection

To find the best convolutional neural network backbones for feature extraction, we compare VGG13, VGG16, VGG19, Resnet18, Resnet34, Resnet50 [12] [13] with the *Baseline* working flow. The input modalities have a fixed resolution of *(256, 256)*. In our experiments, VGG16 and VGG19 outperform other backbones in different lumbar

BMD prediction tasks, in terms of R-value, MAE, and AUC. The VGG family surpasses Resnet variants in all tasks, suggesting that relatively simple backbones work better on CXR texture recognition. The proper combination of convolutional kernels and backbone depth in VGG16 exploits the CXR textures better than in VGG19, and we use VGG16 as the default feature extractor in all experiments.

3.6.2 Image splitting dimension for the Multi-Patch model

In the *Multi-Patch* model, we split the original CXR image into even patches in Figure 3.3. We train and test the *Multi-Patch* model with different split dimensions N (N rows by N columns, N=2,3,4). Smaller split dimensions (*e.g.*, $N = 2$) produce patches covering a larger field of view, which may impede finer-level texture exploration. Larger split dimensions (*e.g.*, $N = 4$) produce patches covering a smaller area, leading to theme shifting of individual patches due to the varied shape, sizes, and positions in scanning. As a result, N=3 fits the best (used as default).

3.6.3 Determine the proper T-score thresholds

Different lumbar vertebrae have distinct BMD distributions in Figure 3.4. We have known their varied Confidence Intervals for osteoporosis, osteopenia, and normal cases in previous sections. Lumbar 4 on average has higher BMD values than Lumbar 1 and there are many cases where lumbar 4 is normal/osteopenia but lumbar 1 is osteoporosis. This discrepancy leads to prediction sensitivity/specificity differences among Lumbar BMD tasks under the same osteoporosis T-score threshold. What's more, in the opportunistic

setting, the models not only predict individual vertebra BMD but also generate unified alarming information. Therefore the T-score thresholds for osteoporosis predictions on different vertebrae could be adjusted to get consistent sensitivity and specificity.

In Table 3.2, we investigate osteoporosis classification under different T-score thresholds. To get a balanced and applicable result in practice, the sensitivity may be desired above 80%, and specificity to be above 90%. Under the fixed T-score threshold (-1.75, -2, -2.25, or -2.5), the sensitivity and specificity have different ratios across vertebrae. When it is reasonably good on Lumbar 1, the metrics are tilted towards a low sensitivity on Lumbar 4. A balanced performance means similar metrics on all vertebrae, which requires flexible threshold adjustments. The **Flex** achieves this goal with approximately 80% sensitivity and 94% specificity on all vertebrae, producing an equalized alarming degree across vertebrae, better for unified patient-level judgments.

3.6.4 Factors leading to large prediction errors

There are many factors influencing predictions, such as the hardware, scanning settings, and sample characteristics (gender, age, lumbar vertebra). For results on Lumbar 1 in Figure 3.5 (1), points far away from the *Diagonal* are large error predictions, corresponding to points above +1.96SD or below -1.96SD in (2). Beyond the [-1.96SD, +1.96SD] range in the L1 Bland-Altman plot, there are 144 cases. Among them are 68 females and 76 males, which are not gender-specific. They scatter across all ages, with data count distribution similar to Figure 3.4 (1). We also check their scanning hardware settings such as voltage, image spacing, and image size, without finding any

particular insights. Then we inspect their landmark localization and bone patch extraction procedures, but there are no mistakes in the intermediate results either. Visually checking these cases, there are no differences from the more accurately predicted ones. So the large errors are due to unknown factors and remained for future research.

3.6.5 Model performance gaps

To quantitatively investigate the prediction differences among five models, we compare performance metrics in Table 3.1. The R-value and AUC appear similar, but MAE could tell the difference. The proposed model on average reduces the MAE by 27.5% compared to the Base model. Improving the Standard Deviation (SD) of absolute errors from 0.066 to 0.05 also brings in a substantially smaller error variance. The proposed model outperforms its variants steadily in all metrics.

To further test the statistical differences among models, we conduct the t-test of whether the absolute prediction errors have the same mean. For the Lumbar 1 BMD task, we first calculate the absolute prediction errors of all models. Then we calculate the p-values between model pairs. Compared to the proposed model, the Base, the MultiPatch, the AttMultiPatch, and the MultiROI have p-values of 2.9e-49, 3.1e-5, 0.022, and 0.028 respectively. The absolute error p-values in other lumbar BMD tasks have similar relations. The baseline model has a nearly zero p-value with all other models, which implies significant prediction differences.

3.6.6 The model performance boundary

Due to the physiological differences between body parts, the corresponding BMDs would vary. Therefore the cross-bone BMD variances could provide a good hint about the model performance boundary. We calculate the mean absolute DXA BMD differences between vertebrae, plotted by gender and age in Figure 3.8 (1)(3). Samples beyond 75 years old are not accurate out of scarcity. The neighboring bones have relatively small and stable differences, such as Lumbar 1 versus 2, and Lumbar 2 versus 3. As the distance increases, the cross-bone BMD difference increases, such as Lumbar 1 versus 3, and Lumbar 1 versus 4. The average BMD differences between two neighboring bones stay in the $[0.05, 0.10]$ range most of the time.

We plot the Mean Absolute Error (MAE) between prediction and DXA BMD by gender and age in Figure 3.8 (2)(4). Generally, predictions on Lumbar 1 have a smaller absolute error, while Lumbar 4 predictions have a larger error, which is in accordance with their BMD magnitude distribution. For most cases, the prediction errors fall in $[0.04, 0.08]$ range, which is even smaller than the counterparts of neighboring bones. Referencing the mean absolute difference in (1)(3), it serves as the upper bound for cross-bone BMD estimation which implies the model performance boundaries. By comparing (1)(2) or (3)(4) in Figure 3.8, our proposed model marches near this upper limit.

Besides MAE, the DXA BMD R-values also shine a light. The DXA BMD R-value between L1 and L2 is 0.918, between L1 and L3, is 0.878, and between L1 and L4 is 0.807. The model prediction R-values are 0.894, 0.899, 0.887 on L1, L2, and L3 respectively in Table 3.1. Though neighboring bones (L1 and L2) have a higher

BMD correlation than CXR-based prediction, the model predictions have surpassed the unconnected bones (L1 and L3) in providing BMD reference. Given the closeness between vertebrae in terms of both geometric distance and physiological function, our CXR-based model has achieved remarkable performances.

3.7 Discussion

3.7.1 The ground truth DXA BMD limitations

The DXA scan is a 2D projection of the 3D object, unavoidably including noise from posterior parts of the vertebrae to interfere with the vertebral body BMD [67] [68]. Prior works extracted 3D geometric and structural measurements from area-DXA scans to mitigate this shortcoming [67] [68] [69] [70]. Metrics such as the trabecular bone score (TBS, based on lumbar DXA scanning) are developed to provide bone microarchitecture and skeletal information [71] [72] [73]. These DXA augmentations shine insights for accuracy correction and quality assessment. Though opportunistic applications do not require strict accuracy, we should be aware of the limitations of using DXA BMD as the ground truth.

3.7.2 Data source limitations

The data is NOT a randomly collected sample set. Instead, it is a convenience sampling from the 'annual health checkup' population (Taiwan) who need to pay their fees. They are in general healthy enough not to visit clinics. Therefore, the sampling may not reflect the general trends of declining BMD. People younger than 40 or older than 90

are not included due to sample scarcity. In this study, the patient statistics with respect to gender and age may not accurately reflect clinical reality. We exclude cases with implants or bone fractures, which are less frequent in opportunistic settings. Some characteristics and limitations have been discussed in the data distribution section. As our data is from one hospital, the model must be tested in more centers with different hardware scanners before wide-range clinical application.

3.7.3 Result interpretation limitations

Although deep learning models have been successfully applied in many vision and language tasks, application in medical tasks requires more caution. The BMD-related patterns and textures are not visually identifiable by a human. In our experiments, the correlation between chest X-ray images and lumbar DXA BMD is established through training models to predict the paired information. The functioning principles of this process have not been fully examined. Although our model achieves impressive osteoporosis sensitivity and specificity, a finer-level analysis of BMD correlation and prediction is expected in future studies.

3.7.4 Applicability

The performance of using chest X-ray to predict BMD is unlikely to match direct DXA examination on the hip and lumbar. Part of this is explained in the performance boundary section, where L1 and L2 have a higher R-value than model predictions. For the formal judgment of osteoporosis or osteopenia, only DXA BMD on the lumbar or

hip should be considered [33] [34]. CXR-based BMD prediction works as a low-cost and opportunistic way to make alarms instead of determining osteoporosis status, and the patient should take hip/lumbar DXA scans in cases of positive predictions. Since the availability and utilization of CXR are much greater than DXA, opportunistic screening using CXR may increase the 'eligible' screening population by multi-folds. For people older than 70 or with spine diseases, hip BMD is more accurate. Certain populations such as post-menopause females should take comprehensive examinations guided by medical experts [34]. This study points out the possibility of a new multi-step osteoporosis screening strategy, incurring no additional costs while gaining apparent benefits of patient risk stratification.

3.8 Chapter Summary

In this paper, we design deep learning models to estimate lumbar vertebra BMD from chest X-ray images. We propose the anatomy-aware *Attentive Multi-ROI* model that can extract local bone textures and generate robust feature representation. The landmark-based ROI extraction promotes the local feature reliability against scanning variations. The transformer-based encoder improves the system's robustness in case of noises and occlusions. The proposed model achieves good performance on vertebra-level BMD prediction as well as patient-level osteoporosis classification. We conduct detailed comparisons of data distribution and model performance. Through extensive experiments and comprehensive analysis, the model holds great clinical potential for opportunistic screening.

Chapter 4

Vertebra Localization and Identification through Computed Tomography

4.1 Background

Accurate vertebra localization and identification are required in many clinical applications of spine disorder diagnosis and surgery planning. However, significant challenges are posed in this task by highly varying pathologies (such as vertebral compression fracture, scoliosis, and vertebral fixation) and imaging conditions (such as limited field of view and metal streak artifacts). This paper proposes a robust and accurate method that effectively exploits the anatomical knowledge of the spine to facilitate vertebra localization and identification. A key point localization model is trained to produce activation maps of vertebra centers. They are then re-sampled along the spine centerline to produce spine-rectified activation maps, which are further aggregated into 1-D activation signals. Following this, an anatomically-constrained optimization module is introduced to jointly search for the optimal vertebra centers under a soft constraint that regulates the distance between vertebrae and a hard constraint on the consecutive vertebra indices. When

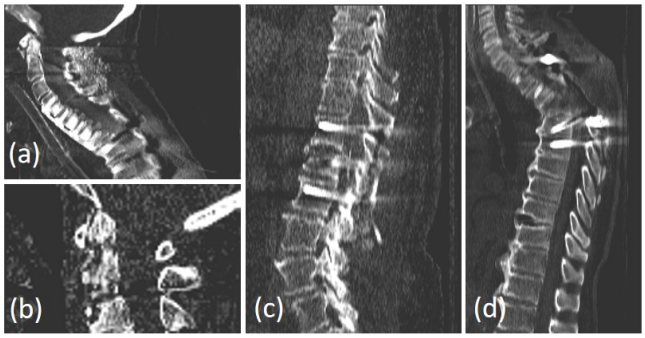

Figure 4.1: Example spine CT images from the SpineWeb benchmark dataset demonstrating the challenges. (a) a small field of view, (b) low image quality, (c,d) metal implants, and severe compression fracture.

being evaluated on a major public benchmark of 302 highly pathological CT images, the proposed method reports a state-of-the-art identification (id.) rate of 97.4% and outperforms the best competing method of 94.7% id. rate by reducing the relative id. error rate by half.

4.2 Introduction

Localization and identification of spine vertebrae in 3-D medical images are key enabling components for computer-aided diagnosis of spine disorders [30]. As a prerequisite step of downstream applications, high accuracies of vertebra localization and identification are frequently demanded. In recent years, many studies have been reported to address this problem, with substantial progress on public benchmarks (e.g., the SpineWeb [74]. However, due to the similar appearances of the spine vertebrae, it remains a daunting task to identify vertebrae with a very high accuracy that meets the requirements of clinical applications.

The challenges in distinguishing vertebrae with similar shapes/appearances are well

recognized by the research community [75–77]. Multiple methods have been proposed to address them by exploiting the anatomical prior knowledge: 1) the spatial order of vertebrae, and 2) the distance between neighboring vertebrae. Spine anatomical knowledge is incorporated into neural networks implicitly using Bi-RNN [75], or explicitly using an information aggregation layer considering the spatial distribution prior to the vertebrae [76]. The anatomical prior has also been used to post-process the neural network output [77]. While steady performance improvements are observed in these works, anatomical knowledge is still not fully utilized. In particular, anatomy-inspired network architectures like Bi-RNN [75] rely on the network to learn the anatomical prior without guaranteed respect to the prior. Building the anatomical knowledge into a network layer [76] or the optimization target [77] makes a compromise that turns the hard constraint (which should be strictly enforced, e.g., the spatial order) into soft constraints that can be violated. As a result, previous methods may produce physically implausible predictions (e.g., vertebrae in reversed order, multiple occurrences of the same vertebra).

Furthermore, while previous methods employ the information exchange mechanisms (e.g., Bi-RNN [75] and message passing [76]) to incorporate the global context, the vertebra label is still classified individually at the output stage for each vertebra without imposing the anatomical constraints. Therefore, these methods completely depend on the information exchange mechanisms to capture and regulate the spatial relationships between vertebrae. Existing fusion mechanisms include 1) recurrent neural network [75], which *encourages* the message passing between vertebrae in a softly learned way instead of *enforcing* it in an anatomy coherent manner; 2) aggregation of the neighboring vertebrae's activation maps [76] following the vertebra distance prior, which is only reliable for short-

range relationships, leaving the global anatomical knowledge insufficiently exploited. A specific optimization formulation is used in [77] to jointly label the vertebrae by formulating a global objective function. However, the Markov modeling of vertebra labels employed in [77] is still limited to capture the short-range relationships and the error accumulates with the Markov steps.

In this work, we propose a vertebra localization and identification method that jointly labels all vertebrae with anatomical constraints to effectively utilize anatomical knowledge. In particular, a key point localization U-Net [78] is trained to predict activation maps for the 26 vertebra centers. Along the automatically calculated spine centerline, the activation maps are warped to rectify the spine and aggregated to form novel 1-D vertebra activation signals. Vertebra localization and identification tasks are then formulated as an optimization problem on the 1-D signals. The spatial order of the vertebrae is guaranteed using a hard constraint to limit the optimization search space. The prior knowledge of the distance between vertebrae is integrated via a soft constraint, i.e., a regularization term in the objective function. The labels for all vertebrae are searched jointly in the constrained search space, which allows the global message to pass among the vertebrae and ensures the anatomical plausibility of the results. We evaluate our method on a main public benchmark from SpineWeb with a training set of 242 CTs and a testing test of 60 CTs. Our method reports the new state-of-the-art identification rate of 97.4%, significantly outperforming the previously best-competing method [77] that achieves a rate of 94.7%.

In summary, our contributions are four-fold. **1)** We propose a simple yet effective approach to aggregate 3-D vertebra activation maps into 1-D signals so that the complexity of the task is significantly reduced. **2)** We exploit the spatial order of the vertebrae as

a hard constraint of the optimization search space, which anatomically ensures plausible outputs. **3)** We introduce the vertebra distance prior as a soft constraint in the optimization of the objective function, flexibly leveraging the relation between vertebrae. **4)** Our method achieves the new state-of-the-art performance by improving the identification accuracy from 94.7% to 97.4% and equivalently cutting the error rate by half.

4.3 Related Work

Vertebra localization and identification task shares fundamental similarities with general landmarks detection tasks, where various formulations and methods have been proposed, including heatmap-base methods [79], coordinate-based [80, 81] and graph-based methods [82,83]. Specialized methods focusing on vertebra localization and identification have also been extensively studied to optimize performance by exploiting the prior knowledge of the spine anatomy.

Early works rely on hand-crafted low-level image features and/or a priori knowledge. Glocker *et al.* [84] propose to use regression forests and probabilistic graphical models to handle arbitrary field-of-view CT scans. They [85] further transform the sparse centroid annotations into dense probabilistic labels for classifier training. Zhan *et al.* [86] use a hierarchical strategy to learn detectors dedicated to distinctive vertebrae and non-distinctive vertebrae. While these methods produce promising results, due to the limited modeling power of hand-crafted features, they lack robustness and produce erroneous results on challenging pathological images. In addition, they fail to exploit the global contextual information to facilitate vertebra identification.

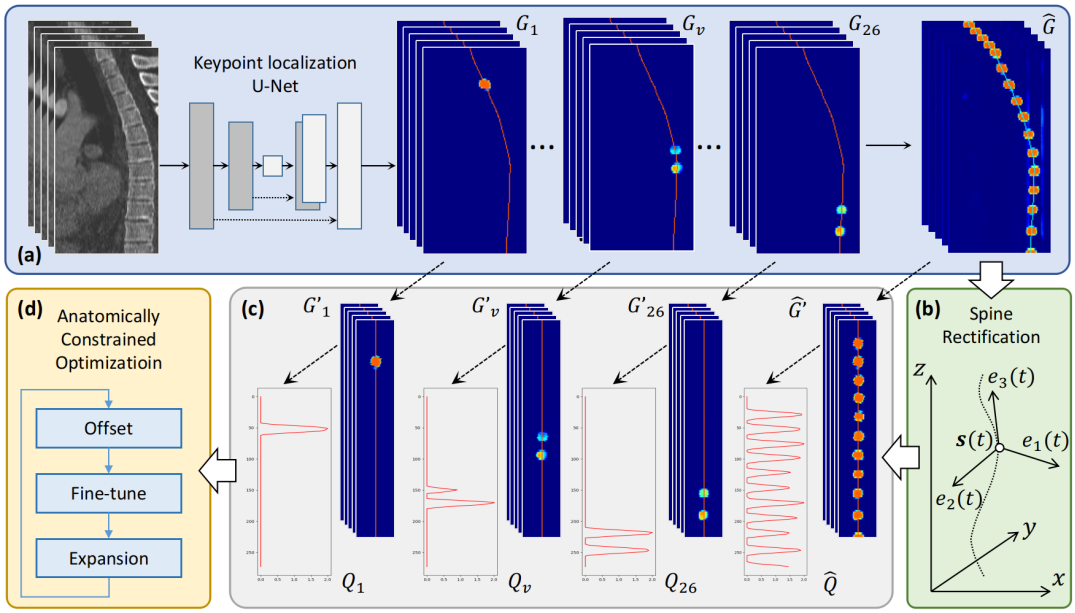

Figure 4.2: Overview of the proposed system. (a) 26 vertebra activation maps $\{G_v\}$ produced by the key point localization nnU-Net, and the all vertebrae activation map $\hat{G}$ produced by aggregating $\{G_v\}$. The centerline of the spine is marked in the activation maps. (b) Spine rectification operator derived from the spine centerline $\mathbf{s}(t)$ in $\hat{G}$. (b) Spine rectified activation maps G'_v and $\hat{G}$ produced by applying spine rectification on $\{G_v\}$ and $\hat{G}$. (c) The 1-D vertebra activation signals $\{Q_v\}$ and $\hat{Q}$ produced by spatially aggregating $\{G_v\}$ and $\hat{G}$. (d) The anatomically constrained optimization module was applied to the 1-D activation signals.

Deep neural networks are employed to detect spine vertebrae and achieve substantially improved performance. A few publications [87,88] employ a convolutional neural network (CNN) to directly detect the vertebra centers. Fully convolutional network (FCN) [89] has also been adopted for the vertebra center detection task [75, 76, 90]. These methods achieve the vertebra localization and identification tasks jointly in one stage. Others employ multiple stages to locate and identify the vertebrae, which can be categorized into top-down [91, 92] or bottom-up strategies. A top-down scheme locates the whole spine first and detects individual vertebrae next. A bottom-up strategy first detects the landmarks of all vertebrae and then classifies them into the respective vertebrae [93, 94].

Many techniques have studied the use of prior knowledge of spine anatomy to facilitate vertebra localization and identification [75–77, 86, 87, 90, 95]. Domain expert knowledge is used to categorize vertebrae into anchor and bundle sets and treat them differently [86]. Markov modeling is adopted to label vertebrae by preserving the consecutive order [77]. Various attempts have been made to automatically learn the knowledge in a data-driven manner [75, 76, 90, 95]. Bi-directional recurrent neural network (RNN) is adopted to enable the model to capture the spatial relations of predictions in different regions [75, 90]. A message-passing mechanism is used to exploit the prior distribution of the distance between vertebrae to regulate the prediction [76]. Adversarial learning has also been employed to encode and impose the anatomical prior [95]. The multi-stage methods [91–94] embed the knowledge of the spine anatomy in their top-down and bottom-up representations.

4.4 Methods

Given a CT image/scan of size $W \times H \times L$, denoted as $I \in \mathbb{R}^{W \times H \times L}$, the goal of vertebra localization and identification is to detect the centers of the spine vertebrae that are present in I and identify their labels. There are in total 26 vertebra labels, including 7 cervical, 12 thoracic, 5 lumbar, and 2 sacrum vertebrae. The model takes the image I as input and outputs the centers of the detected vertebrae $\mathbf{P} = \{x_v, y_v, z_v\}, v \in V$, where $V \subseteq \{1, 2, \dots 26\}$ denotes the indices of any detected vertebrae. For all images in training, the vertebra center annotations $\mathbf{P}$ are provided. Our proposed system consists of three steps: 1) training a U-Net key point detection model to estimate 26 vertebra

activation maps; 2) spine rectification to produce a 1-D activation signal; 3) anatomically constrained optimization to detect vertebra centers from the 1-D signal.

4.4.1 Generation of Vertebra Activation Map

In the first step, we train a key point localization model using U-Net as the backbone network to produce activation maps of 26 vertebra centers. This model is trained using the widely adopted multi-channel activation map regression approach. The multi-channel ground-truth activation maps are generated using Gaussian distribution centered on the spatial coordinates of the vertebra centers. The model is trained using L2 loss on the predicted and ground-truth activation maps. The produced activation maps are denoted as $G_v \in \mathbb{R}^{W \times H \times L}, \quad v \in \{1, 2, \ldots, 26\}$. Although each activation map channel is trained to activate around the center of the corresponding vertebra, due to the repetitive visual patterns of the vertebrae, it is not uncommon for the heap map to falsely activate on the wrong vertebrae, or activate on multiple vertebrae, as shown in Fig. 5.2(a).

Standard key point localization methods process the model-predicted activation map channels individually (e.g., taking the pixel with the maximum activation or taking the centroid) to obtain the key point detection results. A similar approach has also been adopted to produce vertebra localization and identification results [87]. Instead of directly processing the activation map channels to obtain vertebra centers, we propose anatomy-driven processing to achieve robust and accurate vertebra localization and identification, as described in the next two sections.

4.4.2 From 3-D to 1-D Spine Rectification

After obtaining the 3-D vertebra activation maps, we extract the centerline of the spine and aggregate them along the centerline to produce a 1-D vertebra activation signal. The 26 activation maps for individual vertebrae are combined into one activation map:

$$\hat{G} = \sum_{v=1}^{26} G_v, \tag{4.1}$$

which represents the probability of *any* vertebra center without differentiating their indices. While the individual activation map often falsely activates in the wrong vertebrae due to the repetitive image pattern, the activations are typically only around vertebra centers. Therefore, by combining them into one, the centers of all vertebrae are activated, as shown in Fig. 5.2(a).

The centerline of the spine is then computed from the combined activation map $\hat{G}$. It is extracted by tracing the mass centers of the axial slices of $\hat{G}$, calculated as the average coordinates of pixels with activation above 0.5. The extracted centerline is denoted as $\mathbf{s}(t) = (x(t), y(t), z(t))$, where t is the arc-length parameterization. Given the spine centerline, the activation maps G_v are warped so that the centerline becomes straight after warping. Specifically, we calculate a moving local coordinate system along the centerline, denoted as $\langle \mathbf{e}_1(t), \mathbf{e}_2(t), \mathbf{e}_3(t) \rangle$. The three axes are chosen as:

- $\mathbf{e}_3(t)$: the tangent vector of the curve $\mathbf{s}(t)$.
- $\mathbf{e}_2(t)$: the unit vector in the normal plane of $\mathbf{s}(t)$ with the minimum angle to the y-axis of the image (i.e., the patient's front direction).

- $\mathbf{e}_1(t)$: the cross product of $\mathbf{e}_2(t)$ and $\mathbf{e}_3(t)$.

Intuitively, the axes $\mathbf{e}_1(t)$ and $\mathbf{e}_2(t)$ span the normal plane of the spine centerline, where $\mathbf{e}_1(t)$ points at the patient's anterior direction and $\mathbf{e}_2(t)$ directs at the patient's right. Given the centerline and the local coordinate systems, we produce spine rectified activation maps G'_v and $\hat{G}'$ by warping G_v and $\hat{G}$, calculated as:

$$G'_v(x, y, z) = G_v(\mathbf{s}(z) + \mathbf{e}_1(x) + \mathbf{e}_2(y)), \tag{4.2}$$

where $G_v(\cdot)$ denotes the linear interpolation of G_v at the given coordinate. This warping operator can be seen as re-sampling G_v in the normal planes of the spine centerline. In the rectified maps, the spine centerline is straight along the z axis, as shown in Fig. 5.2(c). The anterior and right directions of each vertebra are aligned with the x and y axes.

The rectified activation maps G'_v and $\hat{G}'$ are further processed to produce 1-D signals of vertebra activation, denoted as Q_v and $\hat{Q}$, respectively. Specifically, values in G'_v are summed along the x and y axes, written as

$$Q_v(z) = \sum_{x,y} G'_v(x, y, z). \tag{4.3}$$

The produced 1-D signal indicates the likelihood of vertebra centers at given locations z on the spine centerline. The advantages of the 1-D signal are two-fold: 1) by aggregating the activations in the normal plane, the signal of vertebra centers is strengthened, resulting in a more distinct activation profile, 2) by reducing the spine localization search space to 1-D, the searching complexity is significantly reduced, making it possible and affordable

to adopt more complex optimization approaches. Despite the strengthened activation, false activations in the original activation maps are carried over to the 1-D signal, resulting in false activations in the 1-D signal, as shown in Fig. 5.2(c).

4.4.3 Anatomically-constrained Optimization

Problem Formulation. Given the 1-D response signals $\{Q_v\}$ and $\hat{Q}$, we localize and identify the vertebra centers by solving an optimization problem. Denoting N as the number of detected vertebrae and v_l as the lowest index among them, since the detected vertebrae must be consecutive, their indices can be represented by $[v_l, v_l + N - 1]$. The locations of the detected vertebrae are denoted as $\mathbf{k} = \{k_i\}_{i\in[0,N-1]}$, where i is the vertebra's index relative to v_l. Therefore, k_i indicates the location of the vertebra with absolute index $v_l + i$. Note that since N can be represented by $\mathbf{k}$, we drop N from the parameters for the sake of notation simplicity. The parameters $(v_l, \mathbf{k})$ are optimized to minimize the following energy function:

$$\mathcal{L}(v_l,\mathbf{k}) = -\sum_{i=0}^{N-1} \lambda_{v_l+i} Q_{v_l+i}(k_i) + \sum_{i=2}^{N-2} R(k_i - k_{i-1}, k_{i+1} - k_i). \tag{4.4}$$

$Q_{v_l+i}(k_i)$ is the activation value of the vertebra with the absolute index $v = v_l + i$. $R(\cdot,\cdot)$ is a regularization term that encourages the distances between neighboring vertebrae to be similar, written as:

$$R(a, b) = \exp(\max(\frac{a}{b}, \frac{b}{a})). \tag{4.5}$$

λ_v denotes the weights of the 26 vertebrae. Inspired by the use of anchor vertebrae in [86], throughout our experiments, we treat the two vertebrae at the ends of the spine (C1: Cervical-1, C2: Cervical-2, S1: the first Sacrum, S2: the last Sacrum) as anchors and set their weights λ_v as 2. For all other vertebrae, the weights are set to 1. Intuitively, these vertebrae (C1, C2, S1, S2) at the ends of the spine have more distinct appearances and therefore are given more weight than others.

In the above optimization formulation, we jointly search the vertebra centers to maximize the total vertebra activation score while keeping the distances between vertebra centers regulated. The search space of $(v_l, \mathbf{k})$ implicitly imposes a hard constraint that the detected vertebrae must be consecutive with the indices from v_l to $v_l + N - 1$.

Optimization Scheme. The optimization problem is solved by an initialization step followed by iterative updates. The parameters $(v_l, \mathbf{k})$ are searched in the space: $v_l \in [1, 26]$, $k_i \in [0, L]$. We initialize $v_l = 1$ and the vertebra centers $\mathbf{k}$ as the coordinates of local maxima of $\hat{Q}$ sequentially (*i.e.*, , $k_{i+1} > k_i$). After the initialization, we iteratively apply three operations to search the parameters, namely 1) *offset*, 2) *fine-tune*, and 3) *expansion*.

In the *offset* operation, v_l is optimized via exhaustive search:

$$v_l \leftarrow \arg\min_{v_l} \mathcal{L}(v_l, \mathbf{k}). \tag{4.6}$$

In the *fine-tune* operation, $\{k_v\}$ is optimized via Hill Climbing optimization [96]:

$$\mathbf{k} \leftarrow \arg\min_{\mathbf{k}} \mathcal{L}(v_l, \mathbf{k}). \tag{4.7}$$

The fine-tuning operation adjusts the vertebra centers to minimize the total energy concerning both the individual activation Q_v and the distance regularization.

In the *expansion* operation, a new vertebra center is inserted to $\mathbf{k}$ between $(u, u+1)$. Specifically, the expanded $\mathbf{k}$ is denoted as $E(\mathbf{k}, u)$:

$$\mathbf{k} \leftarrow E(\mathbf{k}, u) = \begin{cases} k_i & \text{if } i \leq u \\ (k_i + k_{i+1})/2 & \text{if } i = u+1 \\ k_{i+1} & \text{if } i > u \end{cases} \tag{4.8}$$

The insertion location u is searched by minimizing the energy function below:

$$u = \arg \min_{u \in [0, N-2]} \mathcal{L}(v_l, E(\mathbf{k}, u)). \tag{4.9}$$

The expansion operation addresses missed vertebrae that are not captured by the local maxima of $\hat{Q}$.

These three operations are iteratively applied until the energy term starts to increase (i.e., indicating convergence). The parameters $(v_l, \mathbf{k})$ associated with the lowest $\mathcal{L}$ during the process are taken as the optimization output. The pseudo-code of the proposed optimization scheme is shown in Algorithm 1. After localizing the vertebra centers from the 1-D signals, their coordinates are mapped back to the 3-D CT image following the reverse spatial mapping of the spine rectification to produce the final 3-D localization results.

Algorithm 1: Optimization

Input: $Q_{v=1,\ldots,26}(z)$ and $\hat{Q}(z)$
$v_l \leftarrow 1$;
$\mathbf{k} \leftarrow$ the coordinates of local maxima of $\hat{Q}(z)$;
$\mathcal{L}_{min} \leftarrow \infty$;
while *true* **do**
  $v_l \leftarrow \arg\min_{v_l} \mathcal{L}(v_l, \mathbf{k})$ // offset ;
  if $\mathcal{L}(v_l, \mathbf{k}) < \mathcal{L}_{min}$ **then**
    $\mathcal{L}_{min} \leftarrow \mathcal{L}(v_l, \mathbf{k})$;
  else
    return $(v_l, \mathbf{k})$ associated with the lowest $\mathcal{L}$;
  end
  $\mathbf{k} \leftarrow \arg\min_{\mathbf{k}} \mathcal{L}(v_l, \mathbf{k})$ // fine-tune ;
  $u \leftarrow \arg\min_{u \in [0, N-2]} \mathcal{L}(v_l, E(\mathbf{k}, u))$;
  $\mathbf{k} \leftarrow E(\mathbf{k}, u)$ // expansion ;
end
Result: $(v_l, N, \mathbf{k})$

4.5 Experiments

4.5.1 Experiment Setup

Dataset. We have conducted extensive experiments on the public dataset provided by SpineWeb [74]. The dataset consists of 302 CT scans with vertebra center annotations.

Table 4.1: Comparison of our method with state-of-the-art methods on the SpineWeb test set of 60 CT images. The mean and standard deviation of the localization error (mm) and the identification rate (%) for different spine regions and their averages are reported.

Method	Cervical			Thoracic			Lumbar			All		
	Mean Error	Std of Error	Id Rate	Mean Error	Std of Error	Id Rate	Mean Error	Std of Error	Id Rate	Mean Error	Std of Error	Id Rate
Glocker *et al.* [85]	6.81	10.0	88.8	17.6	22.3	61.8	13.1	12.5	79.9	13.2	17.8	74.0
McCouat *et al.* [92]	3.93	5.27	90.6	6.61	7.40	79.8	5.39	8.70	92.0	5.60	7.10	85.8
Jakubicek *et al.* [91]	4.21	-	-	5.34	-	-	6.64	-	-	5.08	3.95	90.9
Chen *et al.* [87]	5.12	8.22	91.8	11.4	16.5	76.4	8.42	8.62	88.1	8.82	13.0	84.2
Sekuboy *et al.* [95]	5.90	5.50	89.9	6.80	5.90	86.2	5.80	6.60	91.4	6.20	4.10	88.5
Yang *et al.* [76]	5.60	4.00	92.0	9.20	7.90	81.0	11.0	10.8	83.0	8.60	7.80	85.0
Liao *et al.* [75]	4.48	4.56	95.1	7.78	10.2	84.0	5.61	7.68	92.2	6.47	8.56	88.3
Qin *et al.* [90]	**2.20**	5.60	90.8	3.40	6.50	86.7	2.90	4.30	89.7	2.90	5.80	89.0
Chen *et al.* [77]	2.50	3.66	89.5	2.63	3.25	95.3	**2.19**	1.82	**100**	2.56	3.15	94.7
Ours	2.40	**1.18**	**96.8**	**2.35**	**1.28**	**97.8**	3.19	**1.69**	97.2	**2.55**	**1.40**	**97.4**

This dataset is commonly considered challenging and representative of this task, due to various pathologies and imaging conditions that include severe scoliosis, vertebral fractures, metal implants, and small field-of-view (FOV). In our experiments, we adopt the same dataset split as previous methods [75–77, 84, 87], where 242 CT scans from 125 patients are used for training and the remaining 60 CT scans are held out for testing.

Metrics. We adopt the two commonly used evaluation metrics: *identification rate* and *localization error*. The identification rate measures the percentage of vertebrae that are successfully identified. A vertebra is considered as correctly identified if the detected vertebra center and the ground truth are mutually the closest and their distance is within 20 mm. Localization error measures the mean and standard deviation of localization errors (in mm) of correctly identified vertebrae. The evaluation metrics are calculated for the vertebrae overall, as well as separately for different spine regions (*i.e.*, , cervical, thoracic, and lumbar vertebrae).

4.5.2 Implementation Details

We trained our model on a workstation with Intel Xeon CPU E5-2650 v4 CPU @ 2.2 GHz, 132 GB RAM, and 4 NVIDIA TITAN V GPUs. Our method is implemented in PyTorch. The key point localization model is implemented using nnU-Net [97] [98]. CT images are re-sampled to $0.3 \times 0.3 \times 1.25$ mm spacing. During training, we crop 3-D patches of size $128 \times 160 \times 64$ voxels from each CT scan as input. For inference, we apply the trained model on non-overlapping patches of the same size to obtain the localization activation maps for the full image. The SGD optimizer with a learning rate

of 0.01, a weight decay of 3e-5, and a mini-batch size of 2 is used to train the model for $1,000$ epochs.

4.5.3 Quantitative Comparison with Previous State-of-the-art Methods

We compare our method with 9 baseline methods, including a classic method with hand-crafted feature [84], multi-stage methods [91,92], techniques with data-driven anatomical prior [75, 76, 90, 95] and methods with anatomy inspired architectures [77, 87]. The results are summarized in Table 4.1. Overall, our method significantly outperforms all comparative methods, reporting an id. rate of 97.4% and a mean error of 2.55mm. The closest competitor, Chen *et al.* [77], reports an id. rate of 94.7% and a mean error of 2.56 mm. We reduce the id. error rate significantly from 5.3% to 2.6%, by absolute 2.7% (or relative 50.9%). When evaluated on three spine regions separately, the id. rates of our method are still better than all comparison methods, except for the lumbar region when compared to Chen *et al.* [77]. On cervical and thoracic spines, our method achieves the highest id. rates of 96.8% and 97.8%, respectively.

We note that Chen *et al.* [77] significantly outperforms the other baseline methods in the id. rate. The advantage can be attributed mainly to the adoption of the hard physical constraint imposed by the Markov modeling, which ensures the output to be anatomically plausible. Despite the performance gain, it has a noticeable tendency to achieve higher id. rate on the lumbar spine (*i.e.*, ranked 1st out of 10) but lower id. rate on the cervical spine (*i.e.*, ranked 7th out of 10). This is because their method employs the Markov model to trace vertebrae from one end of the spine (*i.e.*, lumbar) to the other end (*i.e.*, cervical).

The Markov model successfully regulates the consecutive vertebra indices, which leads to significant performance gain compared to previous methods without such regulation. However, the error can accumulate along with the number of Markov steps as the process goes toward the cervical end. In contrast, our method globally searches and identifies the vertebrae with the constraint of consecutive vertebra indices, which eliminates the directional bias caused by Markov model [77] and results in consistent performance in all spine regions.

4.5.4 Ablation Study

4.5.4.1 Effects of the Proposed Components

The spine rectification and anatomically constrained optimization are at the core of our method. In this section, we analyze their effects and behavior via an ablation study of the following alternative methods. The most naive alternative to the iterative optimization is to take the maximum location of individual 3-D activation map G_v as the center of the v-th vertebrae, denoted as *base* model. A slightly more sophisticated approach is to take the maximum location of individual 1-D activation signal Q_v as the center of the v-th vertebrae. Since this approach employs spine rectification, it is denoted as *base+rectify*. In the above two approaches, since there is no constraint applied, physically implausible vertebra orders that violate the anatomy can be produced. A more advanced variation is to take the locations of local maximums of $\hat{Q}$ as candidates of consecutive vertebrae and determine v_l following Equation 4.6. This approach ensures the consecutive order of the predicted vertebrae on top of spine rectification, thus referred to as *base+rectify+order*.

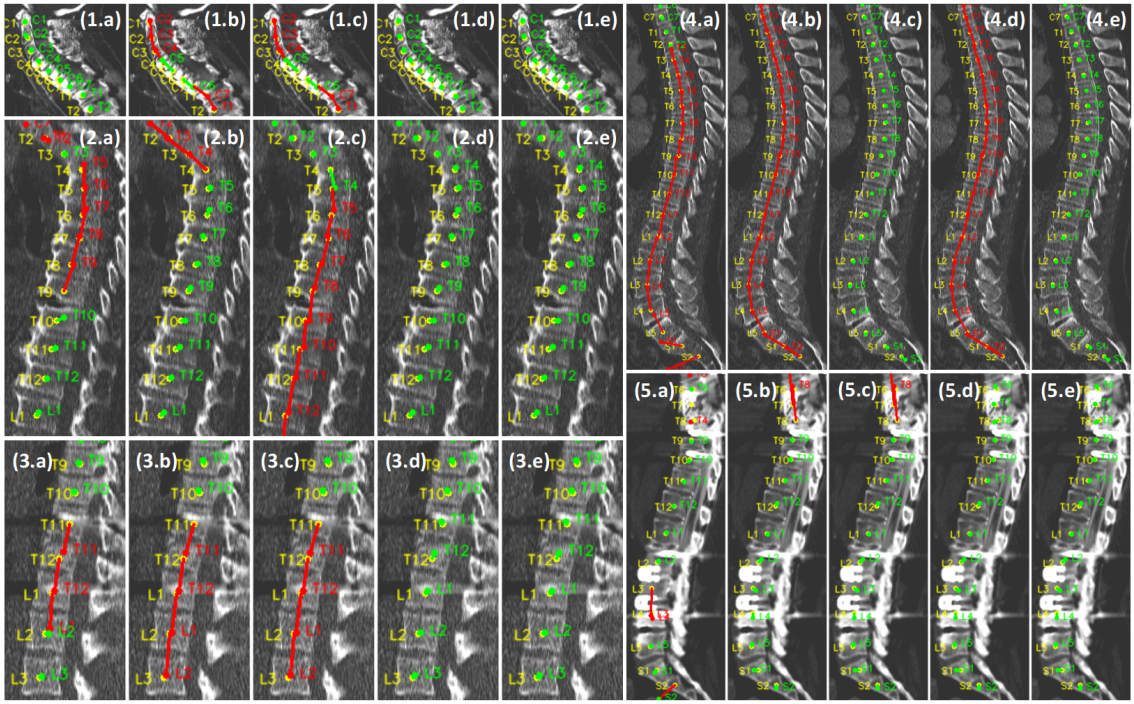

Figure 4.3: Visualization of five sets of final results by five methods. **Dataset (1-5)**: (1) CTs of the cervical spine, (2-4) CTs of the thoracic and lumbar spine, (5) CTs with a metal implant. **Methods (a-e)**: (a) base, (b) base+rect+order w.o. λ, (c) base+rect+order, (d) base+rect+optim w.o. λ, (e) base+rect+optim (ours). The ground-truth vertebra centers are marked by yellow dots and labels. The correct and incorrect predicted vertebra centers are marked in green and red colors, respectively. A line is drawn between the ground truth and predicted centers of the same vertebra for better visualization of the localization error.

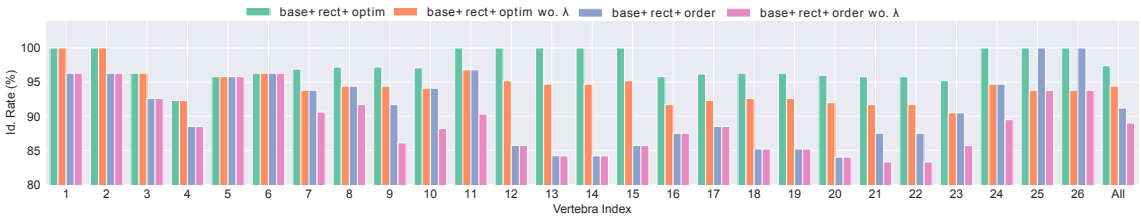

Figure 4.4: Comparisons of base+rect+order and base+rect+optim with and without using vertebrae weights. The identification rates (%) for each vertebra and their averages are reported.

Table 4.2: Results of the ablation study analyzing the effects of proposed components in our method and the use of vertebra weight λ.

Model	Cervical			Thoracic			Lumbar			All		
	Mean Error	Std	Id Rate	Mean Error	Std	Id Rate	Mean Error	Std	Id Rate	Mean Error	Std	Id Rate
base	2.24	1.25	96.8	2.53	1.53	76.0	2.67	1.66	75.2	2.46	1.48	81.9
base+rect	2.46	1.64	95.8	2.32	1.63	73.8	3.14	1.70	79.3	2.54	1.68	81.4
base+rect+order	2.55	1.83	94.2	2.31	1.54	89.4	3.19	1.71	91.0	2.57	1.70	91.2
base+rect+optim	2.40	1.18	96.8	2.35	1.28	97.8	3.19	1.69	97.2	2.55	1.40	97.4
base+rect+order wo. λ	2.54	1.84	93.7	2.33	1.25	87.2	3.15	1.69	86.9	2.57	1.58	89.0
base+rect+optim wo. λ	2.40	1.18	96.3	2.38	1.28	94.1	3.15	1.68	92.4	2.55	1.39	94.4

Note that this approach is equivalent to our method that stops after the *offset* operation in the first optimization iteration. Since our method employs both spine rectification and anatomically constrained optimization, it is referred to as *base+rectify+optim*.

The results of the ablation study are summarized in Table 4.2 and Fig. 4.4. Visualizations of illustrative image example results are shown in Fig. 4.3. The purpose of spine rectification is to enable applying anatomical constraints in the downstream processing. Therefore, employing spine rectification without imposing anatomical constraints does not bring any performance gain, as shown by the comparison between *base* and *base+rect*. By imposing an effective/meaningful constraint of the vertebra order, *base+rect+order* ensures physically plausible results and significantly improves the id. rate over *base+rect* from 81.4% to 91.2%. By employing the proposed anatomically constraint optimization, *base+rect+optim* is able to regulate the distance between predicted vertebrae while preserving the physically plausible vertebra order. As a result, the id. rate is further improved from 91.2% to 97.4%.

We also observe that while the overall id. rate improves significantly, the id. rate for the cervical region is consistently high using different methods. This is because the cervical vertebrae have a more distinct appearance and can be reliably recognized.

4.5.4.2 Effect of the Vertebra Weights λ

The vertebra weights λ also play an important role by encouraging the optimization to focus more on the vertebrae that can be reliably detected by the key point localization model. To analyze the contribution of the vertebra weights, we conduct an experiment to compare the performances of *base+rect+order* and our method with and without using vertebra weights. As summarized in Table 4.2, employing vertebra weights leads to improved performance on both *base+rect+order* and our method. In particular, the overall identification rate is improved from 89.0% and 94.4% to 91.2% and 97.4% on these two methods, respectively. The mean error is not affected much by employing the vertebra weights, which suggests that the vertebra weights have little effect on the accuracy of correctly identified vertebrae.

4.5.5 Analysis and Discussion of Failure Cases

In Fig. 4.5, we demonstrate three failure cases of our method. It shows that extreme pathology and/or low quality may degrade the performance of our method. In particular, the first case has severe vertebral compression fractures, which significantly reduces the height of the vertebrae as well as the space margins between them. The second case has low imaging quality, making it difficult to differentiate the boundary between vertebrae.

Consequently, we observe missed detection and false positive results in these two cases, respectively. In the last scenario, the vertebra centers are correctly located but the labels are off by one. The underlying cause of this failure case is the lack of distinct vertebrae that can be reliably recognized. In particular, the more distinct L5 and sacrum vertebrae are not in the field of view. The imaging appearance of T12 vertebrae (the lowest vertebra with rib) is affected by the metal implant.

4.6 Chapter Summary

In this chapter, we present a highly robust and accurate vertebra localization and identification approach. Based on thorough evaluations on a major public benchmark dataset (i.e., SpineWeb), we demonstrate that by rectifying the spine (via converting and effectively simplifying 3-D detection activation maps into 1-D detection signals) and jointly localizing all vertebrae following the anatomical constraint, our method achieves the new state-of-the-art performance and outperforms previous methods by significantly large quantitative margins. The effectiveness of each proposed algorithmic component has been validated using our ablation studies.

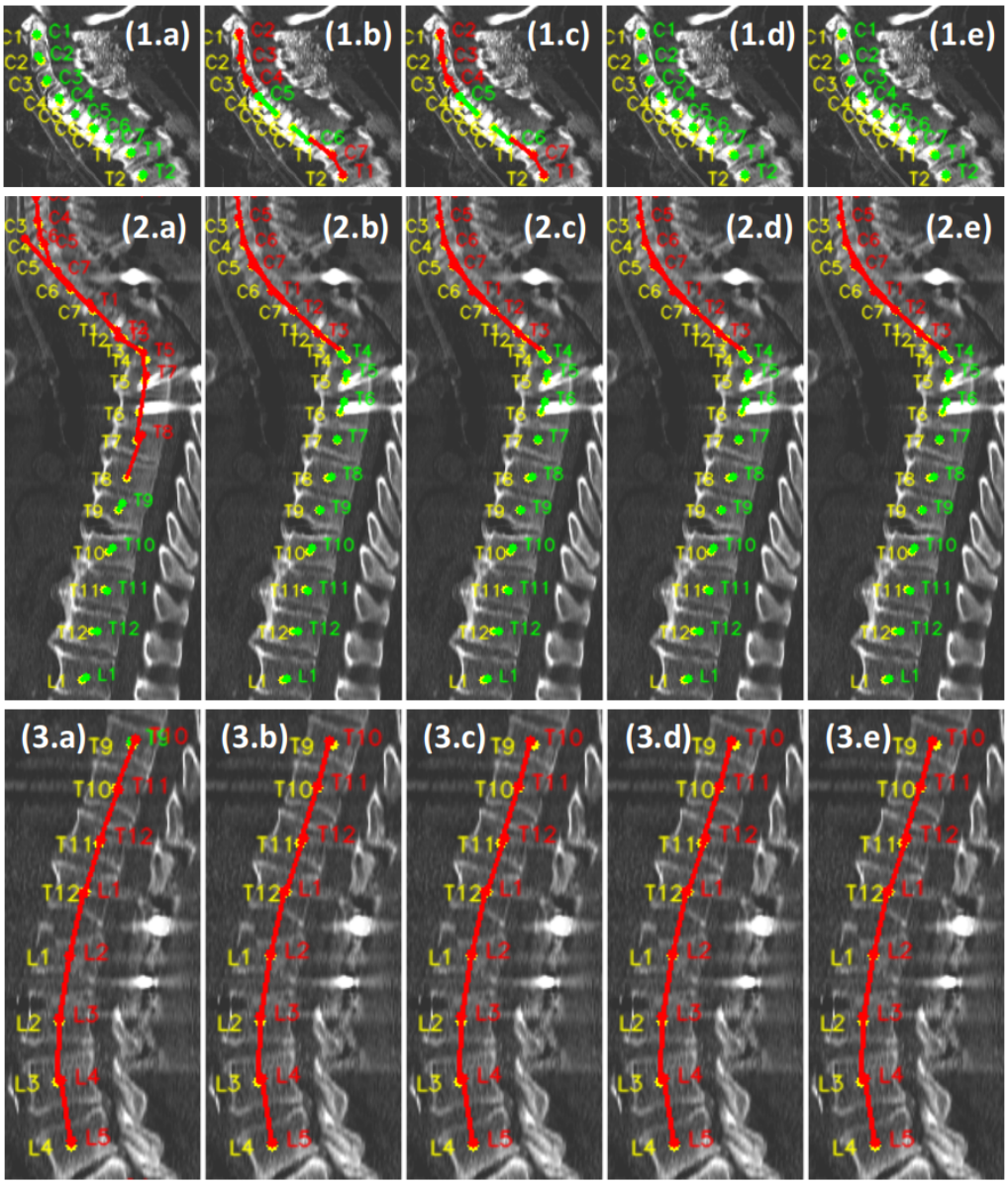

Figure 4.5: Examples of failure cases. The visualization scheme is the same as in Figure 4.3.

Chapter 5

Multi-sensitivity Segmentation with Context-aware Augmentation for Liver Tumor Detection in CT

5.1 Introduction

Liver cancer ranks 6th by incidence, but 3rd by mortality in 2020 [99]. Similar to other cancers, it is important to detect the tumors in the early stage and take prompt action. Ultrasound and non-contrast Computed Tomography (CT) are usually used in the screening setting, while contrast-enhanced CTs are used in the diagnosis scenario. According to LIRAD guidelines, contrast-enhanced CTs or MRIs are sufficient for identifying liver tumors in broad situations. Early detection and accurate diagnosis of liver cancer involve many challenges, such as low awareness, lack of examinations, missed detection of small tumors, and misclassification of lesion types.

In the opportunistic screening scenario(only the Non-Contrast/NC CT scanning), Small lesions usually have poor detection performance by radiologists, due to many factors such as the complex contexts of the liver, ambiguous appearances of small lesions, and low scanning quality. Reading CT scanning to detect potential lesions in the body

checkup scenario is costly since most patients have healthy livers. In the diagnosis scenario where the patient takes multi-phasic contrast-enhanced CT images (Non-Contrast/NC, Arterial-Phase/AP, Venous-Phase/VP, Delayed-Phase/DP) of the liver, it also costs great energy and time for the radiologist to compare, localize, determine the lesion within the large liver organ. Therefore, developing algorithms that can automatically segment, detect, and classify liver lesions in both the screening (only Non-Contrast/NC CT scanning) setting and diagnosis (contrast-enhanced four-phasic CT scanning) setting is of great interest.

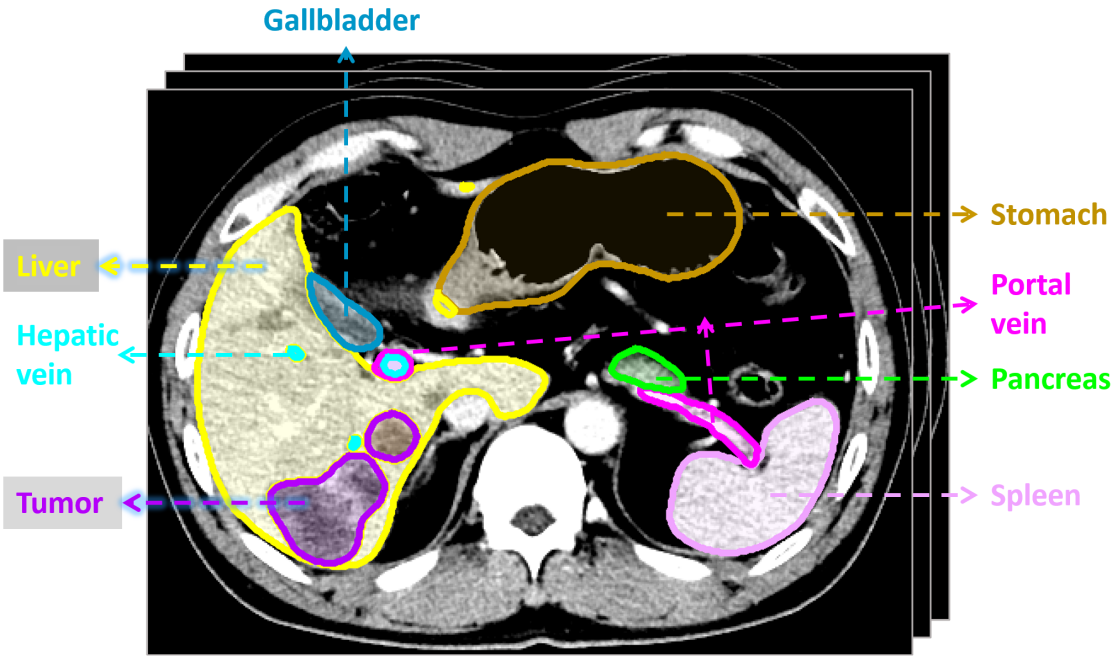

Figure 5.1: To train a better liver lesion segmentation model, we include 7 related organ segmenting tasks. We customize a CT labeling tool and adopt a semi-supervised learning approach to curate a large-scale multi-phase abdominal liver dataset, with high-quality masks.

Liver tumor detection via CT images faces challenges. First, the liver is the largest organ in the abdomen and there are many morphology variations in both the liver organ and the liver lesions, such as location, size, shape, intensity, or texture. The complex

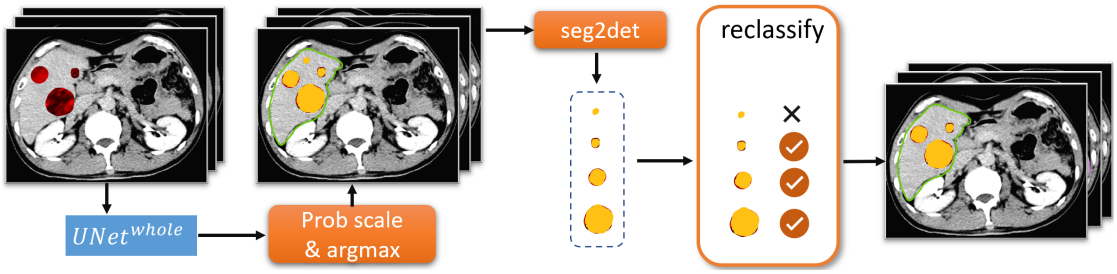

Figure 5.2: The proposed workflow for multi-sensitivity segmentation and lesion shuffling reclassification. The CT images go through the $UNet^{whole}$ model to get the 3D probability maps, and we adjust the lesion sensitivity to generate segmentation masks at different lesion sensitivities. We design the *seg2det* algorithm to generate precise lesion proposals and train a dedicated reclassification module to filter out *False Alarm*s.

vessels and ducts in the liver add to the difficulty of detecting and localizing lesions. Liver texture changes due to fat accumulation and fibrosis also create much ambiguity in lesion detection. There are many malignant and benign lesion types in the liver, and some types are hard to tell in the CT due to appearance similarities.

Second, there are no large-scale and high-quality liver CT datasets publicly available for deep-learning training. Due to the above-mentioned reasons, it requires a large number of patient cases of various liver conditions for deep learning models to learn the pattern. According to the LIRAD guideline, contrast-enhanced liver CT scanning images are needed for radiologists to determine the tumor types, and sometimes it requires pathology reports. The labeling part requires Careful and exact 3D masks for both organs and liver lesions, which is also hard to gain. Many works rely on independent datasets curated through collaboration between engineers and doctors. The Curation of a high-quality liver tumor dataset involves diagnosis-related privacy information and clinical practice in the medical center, thus making it difficult to open source.

Lastly, there are no standard evaluation procedures or metrics for comprehensive performance judgment in liver tumor detection tasks. Currently, many works employ

general segmentation metrics, which are volume and surface-based such as Dice, average symmetrical surface distance (ASD), etc. Some works employ object-detection metrics, such as recall/precision, Area Under Curve. However, these general metrics do not adequately evaluate the model performance, not covering the detecting ability of small or suspected malignant lesions which is most important in the real-world application. Therefore we need to adopt fine-level evaluation metrics for small lesions and for longitudinal comparisons.

To address these challenges, we curate a large-scale contrast-enhanced liver CT dataset with lesion segmentation masks, develop multi-sensitivity segmentation algorithms for better probabilistic inference, and employ a lesion-shuffle training scheme for better small-lesion detection. Our contributions have four aspects. Firstly, utilizing the masks in related public datasets, we design an automatic and intelligent labeling pipeline to curate a private dataset, such as multi-task learning, semi-supervised learning, and iterative mask labeling. Secondly, exploring the confidence maps from the 3D U-Net mdoels [100] [101], we design multi-sensitivity algorithms to generate detection results from segmentation, maximizing information utilization. Thirdly, we design the lesion-shuffle training scheme, which leads to more robust models differentiating true lesions from noises. Fourthly, we conduct both the lesion-level and patient-level evaluations in the screening as well as the diagnosing scenarios, with novel practical metrics.

5.2 Related work

Several organizations have proposed guidelines for liver cancer diagnosis via CT scanning [102] [103] [104]. The machine models must follow these medical principles, in order to produce scientific results. As for the scanning modalities, many aspects such as safety, availability, and accuracy come into consideration [105]. While ultrasound is commonly adopted for screening, CT, and MRI improve the detection of Hepatocellular Carcinoma (HCC), especially for patients with cirrhosis [103]. Dynamic and multiphase contrast-enhanced CT and MRI form the cornerstone in the diagnosis of HCC [106]. The *EASL Clinical Practice Guidelines* [107] provides advice for the clinical management of HCC. The Liver Imaging Reporting and Data System (LI-RADS) [108], integrated into the HCC clinical practice guidelines by the *AASLD* [109], provides standardization for HCC imaging in the contexts of screening, surveillance, diagnosis, and treatment response assessment.

Existing works attack the challenges from different aspects [110] [111], such as the organ contexts, lesion appearances, and data labeling. Many focus on the multiple-stage workflow [112] [113] to disentangle task complexity. The divide-and-conquer approach works well in many CT imaging tasks, such as lung, liver, pancreas, and colon. Some works address the lesion ambiguity challenge by multiple view [113], coarse to fine classification [114], multi-phasic CT [114]. Some work addresses the data scarcity challenge by data augmentation, Generative Adversarial Network (GAN) [115], semi-supervised learning [116].

The 2-stage paradigm works well in many CT-based lesion detection tasks [117]

[118] [119] [120] [121] [122] [123] [124]. As organs pack together closely in the abdomen, the boundaries can be vague due to the inter-organ texture similarity. The disproportionate volumes between organs and lesions also impede the model training. One form of the 2-stage pipeline [125] would localize the target organ first, and use a dedicated model to learn finer-level lesion patterns. This coarse-to-fine approach helps reduce false positive lesions and better distinguish lesion types. The two U-Net FCNs [121] is implemented in a cascading way, segmenting the liver and tumors in separate stages. In the coarse-to-fine pipeline by Yue [126], they employ the level-set method to further refine the tumor segmentation. The object-based post-processing [127] can also be employed in the second stage.

Machine models detect lesions based on pixel-level or region-level probabilities, with judging thresholds. Malignant nodules may be missed by an algorithm tuned for high precision, while excess false positive candidates from algorithms with high recall may overburden the lesion-diagnosis process unprepared for ambiguity [117]. Therefore it is critical to make the detecting threshold reasonable. The HPVD model [128] accepts any combination of contrast-phase inputs with adjustable sensitivity depending on the clinical purpose.

Although there have been several public CT datasets for liver tumor detection [129] [130], many shortcomings and challenges exist. The Medical Segmentation Decathlon (MSD) [129] collected from several medical centers, contains abdominal CT images covering multiple organs (pancreas, lung, prostate, liver, colon). The AbdomenCT-1K [130], reassembles data from existing datasets such as MSD and LiTS. However, these datasets are limited in terms of scale, disease distribution, and labeling standards. The clinical

diagnosis process follows strict routines, which is not addressed in current public datasets. Multi-phase CT images are required for accurate diagnosis in many scenarios, which is also absent.

Semi-supervised learning [131] can best utilize the knowledge in existing labeling to create new labeling [132] [133] [116]. In the pseudo-labeling and self-training settings [134], semi-supervised learning enhances the labeling gradually.

5.3 Methodology

The proposed workflow in Fig 5.2 consists of four major steps, which are whole segmentation, probability map adjustment, lesion detection generation, and lesion reclassification. In order to train more robust segmentation models for liver lesions, we include additional segmenting tasks of related organs in the first step. And the $UNet^{whole}$ model segments 6 types of liver lesions (**HCC**=*Hepatocellular Carcinoma*, **Cho**=*Cholangioma*, **Meta**=*Metastasis*, **Hem**=*Hemangioma*, **Cyst**, **Other**=*Focal nodular hyperplasia + implants + others*) and 7 organs (*liver, hepatic vessels, portal vein, gallbladder, stomach, spleen, pancreas*) in Fig 5.1. The performance benefits come from the *Multi-sensitivity segmentation* and the *lesion reclassification*.

5.3.1 Multi-sensitivity segmentation

We employ the nnUNet as the 3D UNet backbone, which takes in CT images ($\mathbf{I} \in \mathbb{R}^{Z \times Y \times X}$, $\mathbf{I}$ is one CT image) and generates probability confidence maps for 6 liver lesions

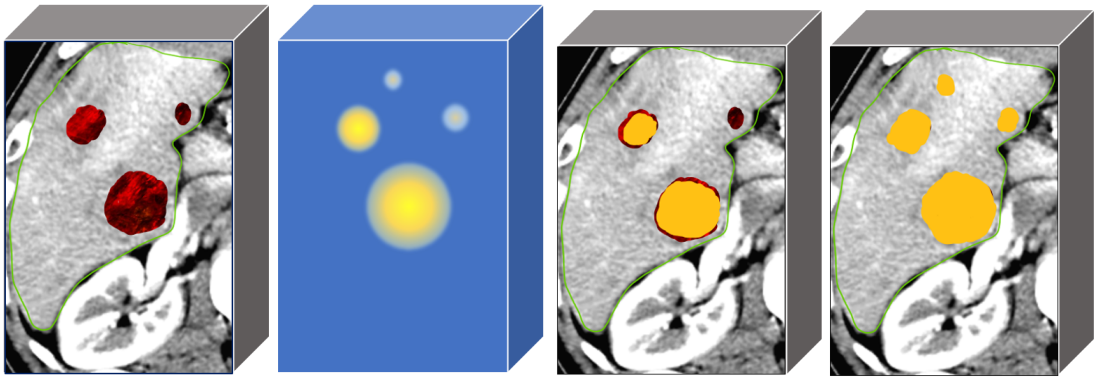

Figure 5.3: The illustration of lesion probability maps and segmentation results of varied sensitivities. The $UNet^{whole}$ generates separate confidence maps (2) for 6 lesion types. The small lesion representations in each 3D map are not stable. Increasing or decreasing the lesion sensitivity may lead to *False Alarm*s (4) or *Miss Detection*s (3).

and 7 organs, $\mathbf{Prob} \in \mathbb{R}^{13\times Z\times Y\times X}$,

$$\mathbf{Prob} \leftarrow UNet^{whole}(\mathbf{I}) \tag{5.1}$$

By default, the UNet simply applies the *argmax* function on **Prob** for the final 3D segmentation mask $\mathbf{Mask} \in \mathbb{R}^{Z\times Y\times X}$. For each voxel (z, y, x) in the 3D space, the *argmax* compares all 13 individual probability maps to find the most confident one and assigns the corresponding label as the voxel value $\mathbf{Mask}^{pred}_{z,y,x}$ in the segmentation mask,

$$\mathbf{Mask}^{pred}_{z,y,x} \leftarrow \underset{i}{argmax}(\mathbf{Prob}_{\mathbf{i,z,y,x}}), \quad i \in [1, 13] \tag{5.2}$$

The lesions (*labels*: 1 to 6) have much less volume compared to organs (*labels*: 7 to 13), subsequently leading to imbalanced prediction tendencies in the probability maps. Treating all 13 labels equally during segmentation map generation results in miss detection (**MD**, or **FN**=*False Negative*) of small or less obvious lesions in Fig 5.3 (3).

By examining the individual confidence maps $\mathbf{Prob_i}$, $i \in [1, 13]$, we find there is much explorable information pertaining to lesion detection. For example, in Fig 5.3 (2) the lesion probability map has strong representations for large and middle tumors, but the small tumor has weak confidence. By setting $f = 4$ in Equation 5.3 before applying the *argmax* in Equation 5.2, the high lesion sensitivity segmentation represents weak confidence signals for small lesions much better in Fig 5.3 (4). The *Prob scale* process is defined as

$$\mathbf{Prob}[i, :, :, :] \leftarrow \mathbf{Prob}[i, :, :, :] \times f, \quad i \in [1, 6] \tag{5.3}$$

We develop the *seg2det* algorithm which precisely separates individual lesions from the $\mathbf{Mask} \in \mathbb{R}^{Z \times Y \times X}$. The lesions of one type all have the same label in $\mathbf{Mask}$, making it difficult to analyze individual lesions. There is no fixed shape for each lesion, and disentangling them requires analysis of the 3D boundary. The *seg2det* algorithm firstly separates all individual lesions in the 2D axial-view slices, then checks the lesion overlapping between neighboring slice pairs. If two lesions in neighboring slices have overlapping pixels in the 2D space, then they are considered from the same lesion in the 3D space.

$$\mathbf{Det}^{gt} \leftarrow seg2det(\mathbf{Mask}^{gt}) \tag{5.4}$$

$$\mathbf{Det}^{pred}_{f=factor} \leftarrow seg2det(\mathbf{Mask}^{pred}_{f=factor}) \tag{5.5}$$

$$\mathbf{TP}, \mathbf{FL}, \mathbf{FP}, \mathbf{FN} \leftarrow match(\mathbf{Det}^{gt}, \mathbf{Det}^{pred}) \tag{5.6}$$

$$\mathbf{Det}^{pred}_{same}, \mathbf{Det}^{pred}_{diff} \leftarrow match(\mathbf{Det}^{pred}_{f=1}, \mathbf{Det}^{pred}_{f=4}) \tag{5.7}$$

We apply the *seg2det* algorithm on both the *ground truth* $\mathbf{Mask}^{gt}$ and the *predicted* $\mathbf{Mask}^{pred}_{f=scale}$. The generated $\mathbf{Det}^{gt}$ and $\mathbf{Det}^{pred}_{f=factor}$ are *lesion detection sets*, with precise 3D locations and masks. We develop the *match* algorithm which compares two *lesion detection sets* and outputs the *overlapping relationship*. There can be three *matching* results for each lesion site in $\mathbf{Det}^{gt}$ and $\mathbf{Det}^{pred}$, *overlapped with the same label* (*i.e.*, **TP**=*True Positive*), *overlapped with different labels* (*i.e.*, **FL**=*False Label*), and *lacking correspondence* (*i.e.*, **FP**=*False Positive* or **FN**=*False Negative*).

5.3.2 Lesion reclassification

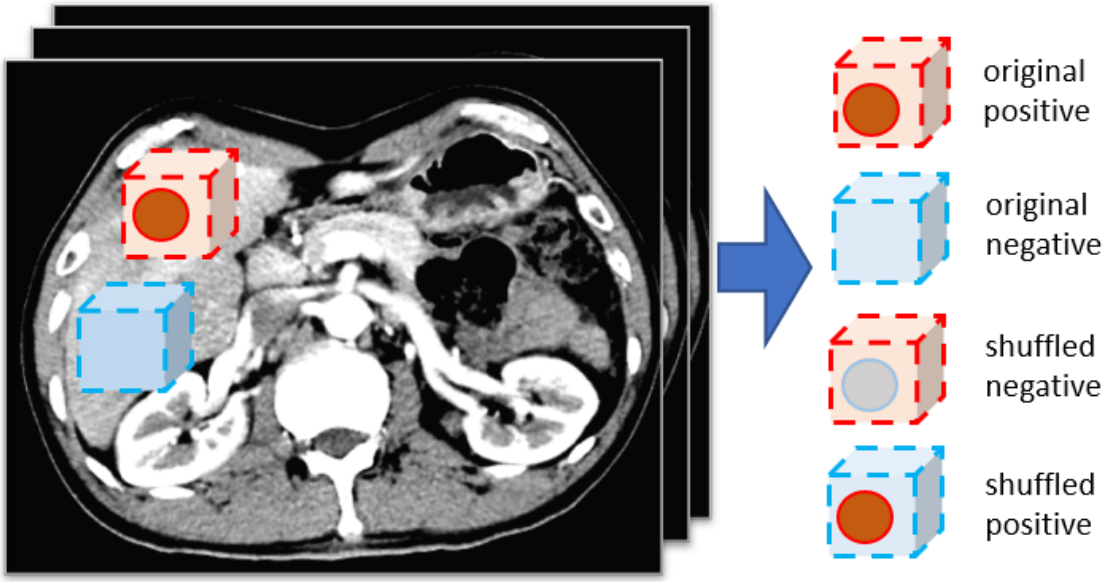

Figure 5.4: The lesion shuffling process is based on the lesion mask and detection results. The patches are randomly sampled around the lesion or in *lesion-free* areas in the liver. The shuffling exchanges a lesion with normal liver textures, creating augmentations for training or inferencing.

The drawback of high lesion sensitivity segmentation in Fig 5.3 (4) is more *false alarms* (*i.e.*, **FP**s) on suspicious liver nodules or normal textures. To effectively contrast normal liver textures with the **TP** and **FP** lesions in $\mathbf{Det}^{pred}$, we train a dedicated segmentation

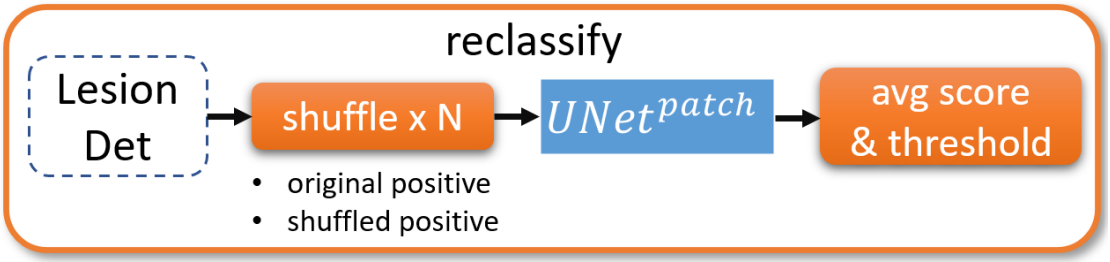

Figure 5.5: The *reclassification* process. Each lesion proposal is shuffled within the liver, to create N augmented patches for the $UNET^{patch}$ to segment. The average score is compared with a threshold to re-classify this lesion.

model $UNet^{patch}$ to classify voxels in the liver-lesion mixture patches. During inference, each lesion has multiple randomly augmented patches and the *lesion truthiness* is measured by the average classification score.

$$\mathbf{Det}^{reclassified} \leftarrow reclassify(\mathbf{Det}^{pred}) \tag{5.8}$$

We design the *shuffle* algorithm which creates 3D augmentation patches of the individual lesions. With the exact lesion mask and liver mask in the 3D space, we are able to crop patches in the lesion-free liver area and then randomly transplant the lesion into patches in Fig 5.4. This context-aware augmentation process effectively utilizes possible lesion surroundings within the same patient's liver. By additionally shuffling $\mathbf{Det}^{pred}_{FP}$ and $\mathbf{Det}^{pred}_{FN}$, the generated patches work as the hard-case augmentation.

$$\mathbf{I}_{patch}, \mathbf{Mask}_{patch} \leftarrow shuffle(\mathbf{Det}, \mathbf{I}) \tag{5.9}$$

During inference, the *shuffle* algorithm generates N original positive and shuffled positive patches. The $UNet^{patch}$ predicts on these N $\mathbf{I}_{patch}$ and compare with N corresponding

$\mathbf{Mask}_{patch}$ to get the average score for the lesion. If the score is below a *threshold*, the lesion will be discarded in the $Det^{reclassified}$. The **FP**s especially in high sensitivity segmentation are effectively reduced.

$$\mathbf{Mask}^{pred}_{patch,n} \leftarrow argmax(UNet^{patch}(\mathbf{I}_{patch,n})) \tag{5.10}$$

$$score \leftarrow \frac{1}{N}\Sigma_{n=1}^{N}(bin(\mathbf{Mask}^{pred}_{patch,n}, \mathbf{Mask}_{patch,n})) \tag{5.11}$$

5.3.3 Implementation

5.3.3.1 Large-scale multi-phase liver CT dataset

We utilize the MSD challenges, which contain different abdominal organ masks in separate datasets. We conduct cross-reference on them to generate 3D organ masks for all, from which we train one unified segmentation model. Then we generate the initial masks for organs and liver lesions in our independent dataset. We align CT images of four phases (*NC*, *AP*, *VP*, *DP*) for each patient with the DEEP algorithm. Two radiologists with more than 10 years of experience inspect and edit the initial masks.

5.3.3.2 The configurations for lesion shuffling

Large lesions usually have strong confidence in the probability maps $\mathbf{Prob}_i, i \in [1, 6]$, and **FN**s and **FP**s are of small or middle size. Therefore we fix the *patch size* for $UNet^{patch}$ model to be $8 \times 128 \times 128$ (approximately $4cm \times 10cm \times 10cm$), enough for

filtering out **FP** lesions. Lesions with a volume larger than $64cm^3$ in Det^{pred} will not go through the lesion reclassification process.

To generate as many realistic patches as possible for $UNet^{patch}$ training, we adopt four possible shuffling schemes in Fig 5.4. The *original positive* patches are randomly cropped around the lesion, and the randomly sampled *original negative* patches do not contain any lesion (*i.e.*, *lesion-free*). The *shuffled positive* patch transplants a lesion into *lesion-free* patch, and the *shuffled negative* patch replaces the lesion in *original positive* patch with normal liver texture. The shuffling process generates both CT images and ground-truth masks for individual patches. Each lesion has around 20 augmentation patches for each shuffling scheme.

In the reclassification process of Fig 5.5, we generate $N = 10$ augmented patches for each *detected lesion*. The *binarize* function in Equation 5.11 calculates the overlapping volume between $UNet^{whole}$-generated $\mathbf{Mask}_{patch,n}$ and the $UNet^{patch}$-generated $\mathbf{Mask}^{pred}_{patch,n}$. If the average *score* is smaller than $threshold = 0.5cm^3$, then this lesion is discarded (*i.e.*, reclassified as **FP**).

5.3.3.3 The multi-sensitivity detection

We set $f = 1.0$ and $f = 4.0$ as lesion probability factors for low sensitivity and high lesion sensitivity in Fig 5.2, which generate $Det^{reclassified}_{f=1}$ and $Det^{reclassified}_{f=4}$ respectively. We then apply the *match* in Equation 5.7 to generate the *common set* $Det^{reclassified}_{same}$ and *difference set* $Det^{reclassified}_{diff}$. The lesions in $Det^{reclassified}_{same}$ are more reliable, while the lesions in $Det^{reclassified}_{diff}$ are mostly from $Det^{reclassified}_{f=4}$ only.

Table 5.1: The test set lesion distribution. Lesions smaller than 0.5 cm^3 are excluded.

Volume	0.5∼2	2∼4	4∼8	8∼16	16∼64	>64	Total	Cases
HCC	22	28	36	47	60	66	259	216
Choln	0	0	1	2	7	3	13	13
Meta	69	27	32	18	24	22	192	34
Heman	1	2	1	0	3	5	12	8
Other	25	8	7	11	8	2	61	46
Cyst	51	7	8	1	2	0	69	46
Malig	91	55	69	67	91	91	464	261
Benign	77	17	16	12	13	7	142	90
All	168	72	85	79	104	98	606	280

Table 5.2: liver model comparison, main table

Dataset setting	Four phase						NC phase					
Model \Metrics	All		Malignancy				All		Malignancy			
	Accu	L.Accu	Preci	Sens	Spec	F	Accu	L.Accu	Preci	Sens	Spec	F
Sens1 w/o Patch	85.2%	92.4%	98.1%	97.7%	92.9%	97.9%	81.3%	88.2%	95.4%	95.0%	82.9%	95.2%
Sens1 w. Patch (no CA)	86.1%	93.1%	99.6%	96.2%	98.6%	97.9%	80.1%	87.0%	99.6%	89.3%	98.6%	94.2%
Sens1 w. Patch	87.0%	94.0%	99.6%	97.3%	98.6%	98.4%	81.6%	88.8%	98.8%	92.3%	95.7%	95.4%
Sens4 w. Patch	88.8%	95.2%	99.2%	99.2%	97.1%	99.2%	84.3%	92.4%	98.8%	97.3%	95.7%	98.0%
Sens1 w. Patch + Sens4 w. Patch	89.3%	95.6%	99.6%	99.2%	98.5%	99.4%	86.6%	93.8%	99.6%	97.5%	98.4%	98.5%

5.4 Experiments

5.4.1 Data Curation

5.4.1.1 Data collection

The private data comes from Chang Gung Memorial Hospital, Taiwan (2008 - 2018). We follow the Helsinki declaration with ethical permission number IRB-201800187B0 (Liver tumor detection through CT images). There are 1631 pathology-verified cases with four-phase CT images, radiology reports, and pathology reports, counting in both the cancerous (about 2/3) and the non-cancerous (about 1/3) cases. The malignant lesions are marked with bounding boxes by experts referring to the radiology and pathology reports [128]. More than half of the non-cancerous cases have some benign lesions in the liver.

The received CT images are in NIFTI format with patient information removed to

protect privacy. The CT machines include GE, SIEMENS, and TOSHIBA. The in-plane resolution ranges from 0.6 mm $\times$ 0.6 mm to 1.0 mm $\times$ 1.0 mm, and the slice thickness is 5.0 mm. Each axial slice size is 512 $\times$ 512, and the slice number varies from 35 to 78. The four-phase CT images are aligned by the *DEEDs* [135] algorithm.

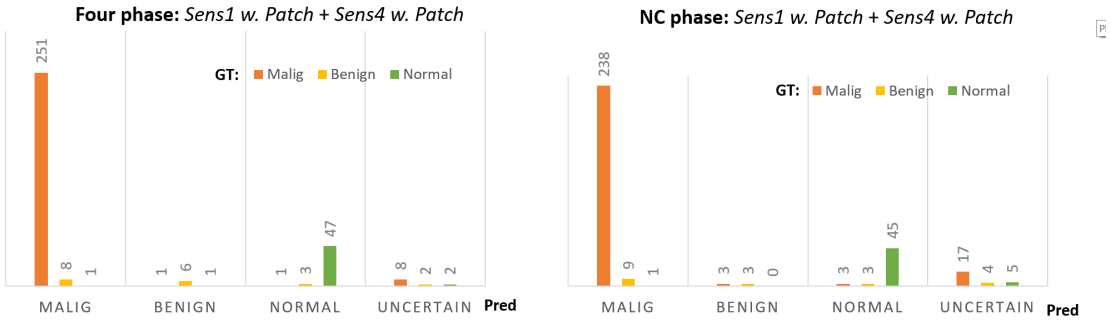

Figure 5.6: The patient-level classification on the test set (331 cases) of the proposed model. A small portion is labeled as *uncertain*, while the rest has high performances for malignancy prediction. The majority of the *uncertain* cases are actually abnormal.

5.4.1.2 Lesion statistics

Our data is from a regional liver medical center whose population is limited to East Asian Han Chinese and the majority of records have liver-related diseases. If multiple lesion types are present, the patient-level *main lesion* is defined by the highest priority (most severe) one. When multiple malignant lesions occur, it is defined by the largest volume type. The lesion has varied sizes/volumes, and the majority of non-HCC lesions are smaller than 16 cm^3. The 331 testing cases comprise 216, 12, 33, 6, 11, 2, and 51 cases for HCC, cholangioma, metastasis, hemangioma, other, cyst, normal (patient-level) respectively in Table 5.1.

5.4.2 Performance Metrics

To measure the lesion segmentation performance, we adopt the label-ignorant dice score, defined as $dice \leftarrow \frac{1}{N}\sum_{i\in[1,N]}\frac{Pred_i \bigcap Gt_i}{Pred_i \bigcup Gt_i}$ where $Pred_i$ and Gt_i are binary lesion masks of cases with lesions. To measure the tumor detection performance, the metrics consider both lesion-level and patient-level matching results. The lesion-level detection statistics include False Negative (*FN*, missed), False Positive (*FP*, not a real lesion), False Label (*FL*, detected but with wrong label), True Positive (*TP*, detected with correct label), and their derivatives (Precision, Recall). We calculate these lesion-level metrics in three grouping manners, which are lesion-specific, All-lesion (*All*), and Malignant-lesion (*Malig*). The *All* counts in all lesions and each type should match exactly, while the *Malig* treats malignant lesions as a single group for the precision and recall calculation. The Rough Recall (*R.Rec*) only concerns lesion overlapping, ignoring lesion types.

At the patient level, the metrics focus on *main lesion* matching. Each prediction or ground truth mask has exactly one main lesion tag under the *label assignment rules*, and the Accuracy (Accu) measures the lesion/label-specific correctness (7 classes). We further group main lesion types into four levels (Level3 = {HCC, cholangioma, metastasis}, Level2 = {hemangioma, other}, Level1 = {cyst}, Level0 = {normal}) to reflect the severity, and the Level Accuracy (*L.Accu*) measures the level-specific correctness. Alternatively, we can treat malignant types (HCC, cholangioma, metastasis) as one group and the left as another group, and adopt the Malignancy Precision, Sensitivity (Recall), Specificity, F-score ($= 2\frac{Precision \cdot Recall}{Precision + Recall}$) as the most concerning measurement.

5.4.3 The proposed model performance

The proposed model ($Sens1\ w.\ Patch + Sens4\ w.\ Patch$) combines results from the joint consensus of two models at different lesion-sensitivity levels. The consensus ratios are 97% in the four-phase setting and 92% in the NC-phase setting. More than half of the *uncertain cases* from the proposed model indeed have malignant lesions in the result. In Table 5.2, the proposed model achieves 89.3% and 86.6% lesion-specific accuracy, and the Malignancy sensitivity/specificity are 99.2%/98.5% and 97.5%/98.4% in the four-phase and NC-phase settings respectively.

5.4.4 The model variants

5.4.4.1 Baseline model

We take the nnUNet segmentation outputs and run the region computation and matching algorithms to get the Baseline ($Sens1\ w/o\ Patch$) results in Table 5.2. In the four-phase setting, it reaches 85.2% lesion-specific accuracy and 92.4% level-specific accuracy. Compared with previous works on the same data source, our Baseline model achieves the HCC F-score of 0.862 (Table 5.3), which is much higher than Yuankai *et al.*'s result of 0.763 [132]. This proves the effectiveness of the segmentation-to-detection transformation workflow in this paper.

5.4.4.2 Context-aware lesion augmentation model

During inference, multiple augmented CT image croppings are fed into the lesion model and we only extract the lesion area in the generated segmentation by comparing it with the correspondingly augmented mask croppings. The re-classification of the target lesion is determined by voting. If more than half of the predictions exceed a volume threshold ($0.5cm^3$), then it is considered a true lesion, otherwise re-classified as the liver. The resulting context-aware lesion augmentation model ($Sens1\ w.\ Patch$) has 1.8% and 0.3% detection accuracy increases compared to the $Sens1\ w/o\ Patch$ in the four-phase and NC-phase settings respectively in Figure 5.2. The $Sens1w.Patch$ effectively reduced *FP*s, where the specificity improves 5.7% and 12.8% in the two settings respectively.

Table 5.3: Lesion-level performance comparisons of model variants in the four-phase setting.

GT \Pred	Sens1 w/o Patch				Sens1 w. Patch(no context aug)				Sens1 w. Patch				Sens4 w. Patch			
	FN	FP	Preci	Recal	FN	FP	Preci	Recal	FN	FP	Preci	Recal	FN	FP	Preci	Recal
HCC	16	7	84.1%	88.4%	26	1	89.4%	84.9%	17	1	87.0%	88.0%	12	8	84.6%	89.9%
Choln	0	3	50.0%	61.5%	0	1	57.1%	61.5%	0	1	57.1%	61.5%	0	1	57.1%	61.5%
Meta	24	8	85.9%	77.5%	40	5	86.2%	68.8%	24	7	86.5%	77.5%	13	8	85.5%	82.3%
Heman	0	2	60.0%	46.2%	2	0	60.0%	50.0%	1	0	66.7%	50.0%	1	1	54.5%	50.0%
Other	6	14	52.6%	49.2%	19	2	66.7%	36.7%	11	3	67.6%	41.7%	9	8	62.5%	41.7%
Cyst	0	5	87.5%	88.9%	7	1	94.4%	82.3%	3	2	91.8%	88.9%	5	4	88.9%	88.9%
All	46	39	80.5%	79.5%	94	10	86.0%	73.5%	56	14	84.9%	78.8%	40	30	82.5%	81.0%
Malig	40	18	95.6%	88.7%	66	7	98.1%	83.1%	41	9	97.8%	88.5%	25	17	95.9%	91.8%

Table 5.4: Lesion-level performance comparisons of model variants in NC-phase setting.

GT \Pred	Sens1 w/o Patch				Sens1 w. Patch(no context aug)				Sens1 w. Patch				Sens4 w. Patch			
	FN	FP	Preci	Recal	FN	FP	Preci	Recal	FN	FP	Preci	Recal	FN	FP	Preci	Recal
HCC	30	27	72.5%	83.9%	49	2	79.8%	77.3%	34	6	77.6%	82.7%	24	15	73.6%	86.7%
Choln	0	2	44.4%	30.8%	0	1	44.4%	30.8%	0	1	44.4%	30.8%	0	1	44.4%	30.8%
Meta	39	7	88.5%	60.0%	63	2	93.8%	50.0%	39	3	93.1%	60.0%	26	10	85.1%	62.6%
Heman	2	0	100%	33.3%	3	0	100%	33.3%	2	0	100%	33.3%	1	0	100%	33.3%
Other	12	8	59.0%	39.7%	21	1	73.9%	29.8%	15	2	69.0%	34.5%	12	4	70.0%	35.0%
Cyst	2	3	81.6%	95.4%	5	3	82.1%	90.2%	3	3	81.3%	93.8%	3	5	83.3%	93.8%
All	85	47	76.1%	71.2%	141	9	82.3%	63.5%	93	15	80.8%	70.0%	66	35	77.0%	72.6%
Malig	69	36	91.0%	80.8%	112	5	98.5%	72.3%	73	10	97.3%	80.1%	50	26	93.2%	85.3%

5.4.4.3 High sensitivity model

The $Sens_1\ w.\ patch$ model reduces *FP*s at the cost of *FN* increase. High sensitivity (**f**=4) effectively brings down the *FN*s, at the cost of more *FP*s. With the help of the context-aware lesion re-classification module, $Sens_4\ w.\ patch$ model keeps a low number of *FN*s and an acceptable *FP*s. In Table 5.2, $Sens_4\ w.\ patch$ model has 1.8% and 2.7% advantage of lesion-specific accuracy over the $Sens_1\ w.\ patch$ model in the Four-phase and NC-phase settings respectively. The benefit becomes even larger for the level-specific accuracy. These comparisons clearly prove the combined power of high lesion sensitivity scaling and context-aware reclassification.

5.4.5 Performance comparisons

The baseline ($Sens_1\ w/o\ patch$) may have limited usability because of its low specificity in the opportunistic screening setting, while the $Sens_1\ w.\ patch$ effectively avoids this drawback. Comparing their patient-level malignancy prediction in Table 5.2, the sensitivity slightly decreases, but the precision and specificity get substantially enhanced, especially in the NC phase setting (3.4% and 12.8% increase). The advantage of increasing lesion sensitivity can be clearly demonstrated by comparing $Sens_4\ w.\ patch$ against $Sens_1\ w.\ patch$ for malignancy detection in the NC setting. $Sens_4\ w.\ patch$ boosts the malignancy sensitivity by 5% while keeps precision and specificity unchanged. Thus the high sensitivity and context-aware augmentation are complementary to achieve the best single-model performance.

More specific proof can be found at the lesion level in Table 5.35.4. The $Sens_1\ w.\ patch$

model reduces *FP* lesions against $Sens_1\ w/o\ patch$ by 64% and 68% in the four-phase and NC-phase settings respectively, proving the power of context-aware lesion re-classification. The $Sens_4\ w.\ patch$ model reduces *FN* lesions against the $Sens_1\ w.\ patch$ by 28.6% and 29% in the two settings respectively, demonstrating the value of high-sensitivity lesion segmentation. Although $Sens_1\ w.\ patch\ (no\ context\ aug)$ model could reduce the *FP*s, it increases the *FN*s substantially, since many lesions are missed due to the lack of adequate liver context learning during the patch model training. The high-sensitivity context-aware lesion augmentation ($Sens_4\ w.\ patch$) model eventually increases the malignant lesion recall (3.1%, 4.5%) and precision (0.3%, 2.2%) in both settings.

$Sens_4\ w.\ patch$ discovers more lesions but increases the *FP*s, the drawback of which is not well presented due to the tumor-focused distribution in our data. It can be problematic in opportunistic settings where most patients do not have liver diseases. The proposed ($Sens_1\ w.\ patch$+$Sens_4\ w.\ patch$) model compares multi-sensitivity context-aware lesion detection results and makes the prediction based on consensus, emulating the varied judging strictness of radiologists. By jointly referring to multi-sensitivity patient-level results, the proposed model achieves both low *FP*s and low *FN*s.

Table 5.5: In the proposed workflow, the lesion level performance of the consensus portion (96% and 92%) by the $Sens_4\ w.\ patch$.

GT \Pred	Four phase: Sens4 w. Patch (319 cases)						
	FN	FP	FL	TP	Preci	Recal	R.Rec
HCC	11	6	13	223	85.1%	90.3%	95.5%
Choln	0	1	5	8	61.5%	61.5%	100.0%
Meta	13	8	15	130	85.5%	82.3%	91.8%
Heman	1	1	5	6	60.0%	50.0%	91.7%
Other	8	6	24	25	65.8%	43.9%	86.0%
Cyst	5	4	2	56	88.9%	88.9%	92.1%
All	38	26	64	448	83.3%	81.5%	93.1%
Malig	24	15	9	385	96.3%	92.1%	94.3%

GT \Pred	NC phase: Sens4 w. Patch (305 cases)						
	FN	FP	FL	TP	Preci	Recal	R.Rec
HCC	18	9	9	207	77.2%	88.5%	92.3%
Choln	0	1	9	4	50.0%	30.8%	100.0%
Meta	16	10	27	89	84.8%	67.4%	87.9%
Heman	1	0	5	3	100.0%	33.3%	88.9%
Other	10	4	22	19	67.9%	37.3%	80.4%
Cyst	2	4	1	57	83.8%	95.0%	96.7%
All	47	28	73	379	79.0%	76.0%	90.6%
Malig	34	20	11	334	94.4%	88.1%	91.0%

Table 5.6: The patient level performance of the proposed model on the consensus portion.

GT \Pred	Four phase: Sens1 w. Patch + Sens4 w. Patch (319 cases)						
	FN	FP	FL	TP	Preci	Sens	Spec
HCC	1	0	6	201	92.6%	96.6%	100.0%
Choln	0	1	4	8	61.5%	66.7%	99.7%
Meta	0	0	10	23	76.7%	69.7%	100.0%
Heman	0	0	2	4	100.0%	66.7%	100.0%
Other	3	0	6	1	50.0%	10.0%	100.0%
Cyst	0	1	0	1	50.0%	100.0%	99.7%
All	4	2	28	238	88.8%	88.1%	95.9%
Malig	1	1	1	251	99.6%	99.2%	98.5%

GT \Pred	NC phase: Sens1 w. Patch + Sens4 w. Patch (305 cases)						
	FN	FP	FL	TP	Preci	Sens	Spec
HCC	3	0	7	193	89.4%	95.1%	100.0%
Choln	0	1	8	4	57.1%	33.3%	99.7%
Meta	0	0	10	19	76.0%	65.5%	100.0%
Heman	0	0	2	2	100.0%	50.0%	100.0%
Other	3	0	7	0	0.0%	0.0%	100.0%
Cyst	0	0	0	1	50.0%	100.0%	100.0%
All	6	1	34	219	86.2%	84.6%	97.8%
Malig	3	1	3	238	99.6%	97.5%	98.4%

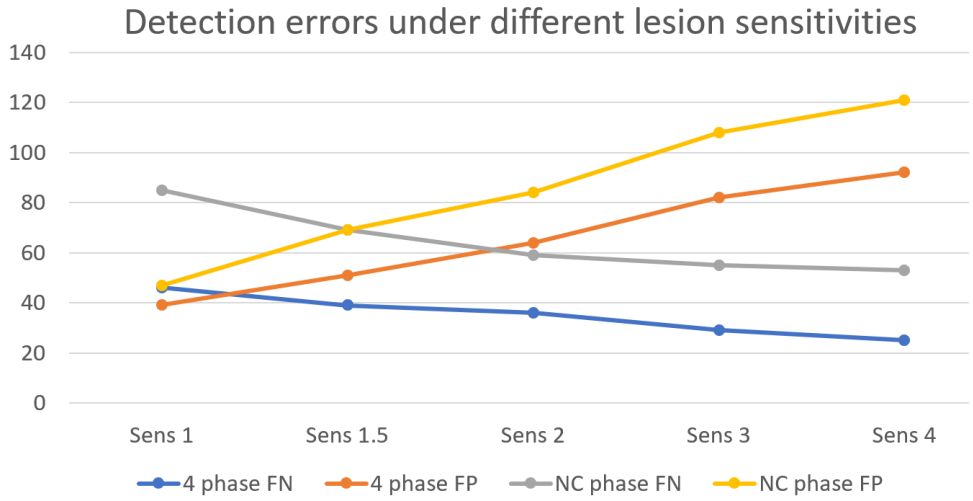

Figure 5.7: The effect of lesion sensitivity scaling factor f. Compare lesion *FN*s and *FP*s using $Sens_f\ w/o\ patch$ model with $f = 1, 1.5, 2, 3, 4$ respectively in both the four-phase and NC-phase settings.

5.5 Ablation study

5.5.1 The adjustment of sensitivity scaling factor

Increasing the lesion factor f in Equation 5.3 reduces the *FN*s while producing more *FP*s. We compare the scaling f effect on the $Sens_f\ w/o\ patch$ model in Figure 5.7. The *FN* reduction (primary goal) slows down beyond $f = 3$, while the *FP*s keeps increasing (side effect) all the time. Therefore the appropriate f would be around 3 or 4. Under the 4-phase setting and f=4, we have α_{FN}=-7 and α_{FP}=17 roughly.

5.5.2 Context-aware augmentation

There are many implementation nuances such as lesion cropping range, negative patch cropping, combination manner, and augmentation strength. Particularly interested in the *negative context*, we compare the context-ignorant augmentation ($Sens_1\ w.\ patch\ (no\ context\ aug)$) with context-aware augmentation ($Sens_1\ w.\ patch$) in Table 5.35.4. For the training and testing preparation, the context-ignorant model lacks the combination of non-lesion regions. It can control the *FP*s at the cost of increasing *FN*s. On the other hand, the context-aware model achieves both low *FP*s and low *FN*s, which proves to be more robust.

5.5.3 The consensus of Multi-sensitivity models

The proposed ($Sens_1\ w.\ patch + Sens_4\ w.\ patch$) workflow divides the results into the *consensus* and the *uncertain* portions. We show the consensus portion performances in Table 5.5,5.6. At the lesion-level (NC-phase), the consensus portion has considerable improvements in malignancy precision (94.4% vs. 93.2%) and recall (88.1% vs. 85.3%) compared to the corresponding results of $Sens_4\ w.\ patch$ in Table 5.4. The patient-level statistics (**FN, FP, FL, TP**) are comparable in the four-phase and NC-phase settings. Though the HCC, cholangioma, metastasis seem to have higher *FL*s, many misclassifications happen within the malignant group. At the patient level, it achieves similar malignancy precision and specificity in the NC-phase setting compared to in the four-phase setting.

5.6 Discussion

5.6.1 Dataset limitations

Our data is from a regional medical center for liver diseases, where the tumor chances are much higher, and only 15% cases are free from liver lesions in the testing set. In the opportunistic setting, the majority of patients would not have malignant liver tumors. Thus the performances in this paper only apply to similar data distributions and may vary substantially on other datasets. However, the principles of lesion sensitivity adjustment, region computation, and context-aware lesion augmentation would generalize to all tumor detection tasks.

5.6.2 The performance upper bound and evaluation applicability

Due to the complex conditions and various textures in the liver, many ambiguous lesions would not be classified into a certain type by appearance. Therefore changing the lesion definition or labeling protocol would dramatically influence the outcome. What's more, since lesions by nature are growing biological tissues that could morph in different scanning settings, there is no perfect boundary in many lesion categorization scenarios. In short, the deep learning models could only approach the performance upper bound preset by dataset characteristics, lesion categorization rules, labeling protocols, and evaluation metrics. Any changes would render the results from different works less comparable.

5.7 Chapter Summary

In this paper, we design deep learning models and workflows to segment and detect 6 types of liver lesions, which works for both contrast-enhanced CTs and non-contrast CTs. The proposed Multi-sensitivity context-aware augmentation model could detect malignant lesions with high precision and recall rates. We develop the region computation and matching procedures for a fully automatic working pipeline. We show the benefits of the proposed method with comprehensive experiments and results, with lesion-level and patient-level analysis supporting the technological soundness. The proposed model holds great clinical potential in both opportunistic screening and diagnosis settings.

Chapter 6

Conclusions and Future Perspectives

We have covered both the computational basis and application principles of medical imaging analysis in previous chapters. Now we will summarize the critical aspects and discuss some concerns. From experiments and reflections, we can grow and create better systems. In the end, we will look into the future, discussing technological development, industry evolution, societal preparations, and medical service trends.

6.1 Dissertation summary

6.1.1 Deep learning for medical image analysis

Deep learning has become the *de facto* backbone for computer vision problems, thanks to the foundations laid by neural network research and high-performance computing. The convolutional neural network can extract vision patterns with high fidelity and efficiency, and the deep-layer architecture enables complex concept learning. Architecture variants, loss functions, efficiency-aware design, robust parameter optimizations, and novel training schemes all together push the deep learning capability forward in computer vision tasks.

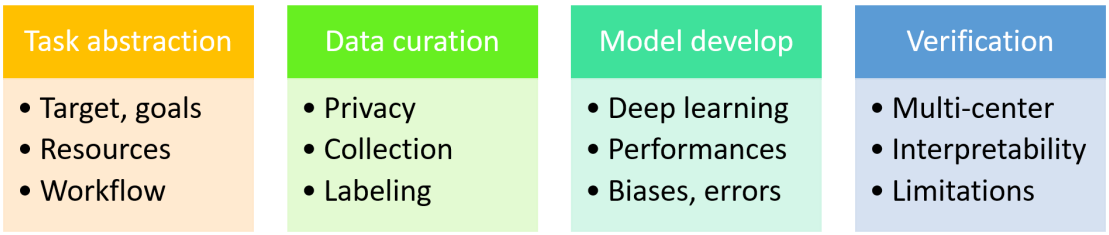

Figure 6.1: The formulation of medical image analysis tasks. There are multiple aspects to consider, including task abstraction, data curation, model development, and verification.

As an important branch of computer vision, medical imaging analysis benefits from deep learning technologies. Combining anatomy constraints and medical knowledge with scanning images, medical imaging tasks nowadays can be solved with engineering elegance and clinical satisfaction. Deep learning models for medical image analysis have been applied in many clinical settings, diagnosing diseases such as cancer, Alzheimer's Disease, and Covid-19. With careful design and verification, it is foreseeable to become more important in computer-aided diagnosis.

6.1.2 The formulation of medical imaging tasks

There are many factors to consider in the medical imaging domain. Similar to other computer vision tasks, medical imaging relies on task abstraction, datasets, model selection, training, and testing. The differences are also obvious, such as privacy concerns, data sharing and labeling, and clinical interpretability. A mindful formulation calls for close collaboration between researchers, engineers, and doctors. Successful deployments depend on efforts from multiple parties, who conduct multiple rounds of verification.

Broadly speaking, medical imaging tasks include not only radiography and MRI but also ultrasonography, histopathology, and photoplethysmogram. Radiography and MRI

images generate a 2D or 3D projection of internal body parts, with high resolution and accuracy, making them the first choice for disease screening and diagnosis. Depending on the composition of the interested body parts, non-contrast or contrast-enhanced radiography/MRI can be configured to realistically present the internal organ and lesion morphology, which serve as important clinical clues. For example, non-contrast CT scanning has been adopted to monitor lung changes in patients infected with Covid-19.

There may be different challenges for each task, but the common one is data collection and sharing. Medical data collection is restricted by the government and medical centers, and the patients must be asked for a permit. Researchers must follow the Helsinki Declaration and acquire ethical permission for the data collection and modeling. For collaboration purposes, data may be transferred to engineers out of the medical center. Before the patient data leave the hospital's database, thorough data cleaning to eliminate personal information must be conducted, to protect privacy. The collected data should be protected and limited only to approved parties, to prevent unauthorized distribution.

Due to the limited number of data samples, the trained models would unavoidably introduce bias and discrimination. The sample distribution would not truthfully represent the real distribution in society for several reasons. Firstly, medical centers usually cover a regional population, bearing specific ethnic characteristics. And many diseases are directly influenced by living environments and habits. Secondly, the sample collection rate for patients may vary because of disease severity, economic status, and willingness to share. Therefore the final samples tend to represent patients with severe diseases. Thirdly, the training of convolutional neural networks encourages models to predict conservatively, biased toward classes with fewer samples. Class-specific augmentation, weight adjustment,

or post-calibration is usually needed to make corrections. To mitigate the bias and discrimination issues, explicit documentation of the sample distribution and model prediction threats are necessary, to promote clinical interpretability. Before the real application, the model also needs to go through verified cross-center verification.

6.1.3 Medical imaging applications in this dissertation

Bone Mineral Density estimation from chest X-ray images. We collaborate with a large regional medical center to collect paired chest X-rays with DXA measurements. After studying the task characteristics and identifying the challenges, we propose the attentive multi-ROI workflow, which yields predictions with small Mean Absolute Error and high correlation with ground truth (DXA scores). The proposed pipeline consists of multiple steps, including landmark localization, local patch cropping, feature extraction, and transformer-based feature fusion. Given chest X-ray images, we train models to predict the BMD of four lumbar vertebrae independently with the four-fold cross-validation scheme. We package all the modules into a docker image and develop an HTTP server for user interaction through web browsers.

Spine vertebra localization and identification via CT images. Vertebra detection in CT scans has several challenges, namely narrow field of view, large spine curvature, metal implantation, vertebra fracture, and scanning noise. Previous methods rely on techniques such as feature fusion, message passing, joint prediction, and 2-stage classification. None of them fully address these challenges explicitly. We propose the 3D probability transformation and anatomy-constrained optimization, which models the physical characteristics and

outputs the most possible detection result. We achieve state-of-the-art performance on a challenging public benchmark, which reduces nearly 50% identification error compared to the second best.

Liver tumor segmentation and detection in CT scanning. Liver cancer ranks 3rd in cancer mortality, and it has increasing trends in many countries. One way to mitigate the society of cancer is screening and detecting cancer at its early stages. Low-dose CT scanning is being used in many body-checking scenes, and it can be used to examine liver health status. One challenge is the lack of lesion-level ground truth segmentation masks. We collaborate with a liver-disease medical center to curate large-scale four-phase datasets. We customize a 3D CT labeling tool to efficiently create and inspect CT masks for the liver tumor detection task. We develop the multi-sensitivity context-aware lesion augmentation pipeline, which can effectively reduce False Negatives and False Positives. Our fully automatic workflow reaches high sensitivity and specificity for malignancy detection at the patient level.

6.2 Future perspectives

6.2.1 Prospective directions of medical image analysis

Deep learning models have achieved super-human abilities in recognizing everyday objects. Given enough training samples for distinguishable patterns of predefined classes, the convolutional neural network can reliably learn the representative features. Deep learning architectures can be adjusted accordingly for the best-fitting result. Training techniques such as augmentation, transfer learning, semi-supervised learning, unsupervised

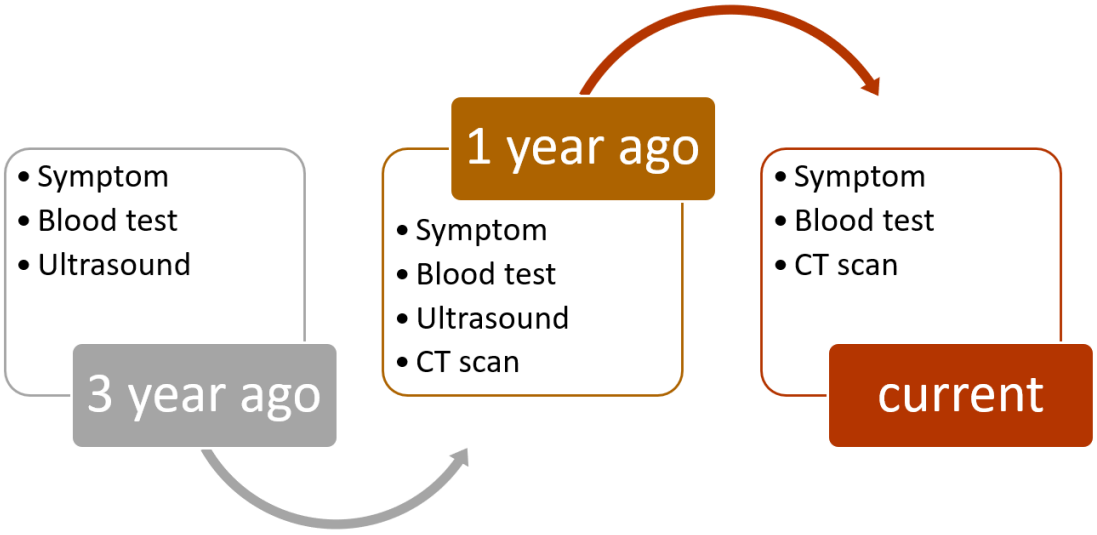

Figure 6.2: The longitudinal study involves health data from different periods.

learning, and adversarial learning have been successfully adopted in computer vision tasks. It is reasonable to assume that deep learning models could achieve a similar degree of performance given enough qualified training samples, utilizing radiology images for defined disease diagnosis in any single organ.

The longitudinal characteristics of human disease have drawn more attention in medical imaging research. Attacking disease research as an integrated process of diagnosis and prevention, the relevant practitioners should pay attention not only to the current medical status of the patients but also to the manifestation patterns in the longer term. For example, radiology scans of the same patient at different intervals may record the disease-changing patterns. By monitoring and comparison, medical care providers could identify emerging anomalies and suggest follow-ups. Machine learning models could sense subtle or suspected changes, in a much better way. Another example is using historical medical scans to predict disease progression. Every patient has unique genes, habits, and conditions, which influence lesion development. Customized medical care

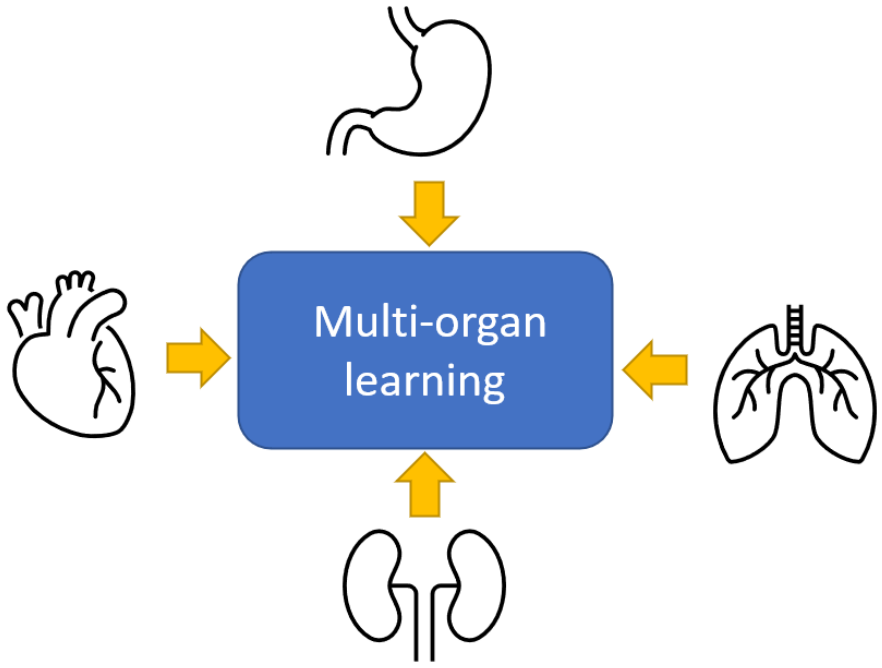

Figure 6.3: Many diseases show symptoms simultaneously at related organs. The collective information may benefit machine learning models.

requires doctors to know patient health records and provide precise treatment. Based on both the general and the personalized disease evolving patterns, it is possible to make customized treatment plans, where deep learning models could play an important role in pattern learning and prediction.

The multi-organ joint analysis would contribute to the ability of medical image analysis. Multi-organ analysis has been adopted in anatomical models in recent years [136]. The human organs are not only spatially connected but also closely related in functionality and disease manifestation. Radiologists and doctors often need to consider the inter-relations for better diagnosis and therapy in a holistic manner, taking the human body as a complex system. Some disorders and diseases may show signs in multiple organs simultaneously. For example, the human digestive system consists of the gastrointestinal tract and the accessory organs of digestion, which include the tongue, salivary glands,

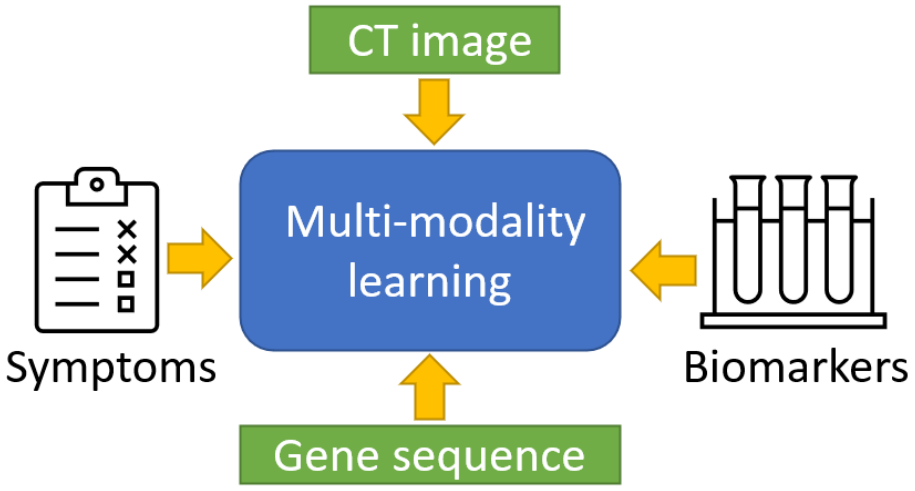

Figure 6.4: Some diseases not only have visual changes in the organ but also bring in the bio-marker anomaly. In clinical reality, patients would get multiple examinations to make the best decision. Therefore machine learning models need to learn multi-modality patterns comprehensively to understand the disease.

pancreas, liver, and gallbladder. The malfunction of the liver may lead to the abnormal status of the gallbladder, subsequently increasing the burdens of the intestines. When the small intestine does not function well, it would put pressure on the colorectal tract. So the disorder or lesions in the colorectal part may be a result of liver disorder, and the joint analysis of these organs is necessary. When more comprehensive medical imaging data is available in the future, deep learning models for multi-organ joint analysis are expected to improve our understanding of many diseases.

Multi-modality analysis contributes to a better understanding of diseases. Disease diagnosis and treatment are complex and comprehensive processes, which depend on multiple sources of medical information. Many diseases involve changes in different aspects simultaneously, such as the blood, the organ, the skin appearance, the genes, and so on. Radiology analysis on the organ alone sometimes would not capture the body condition accurately, so it needs multi-modality learning of related sources to understand

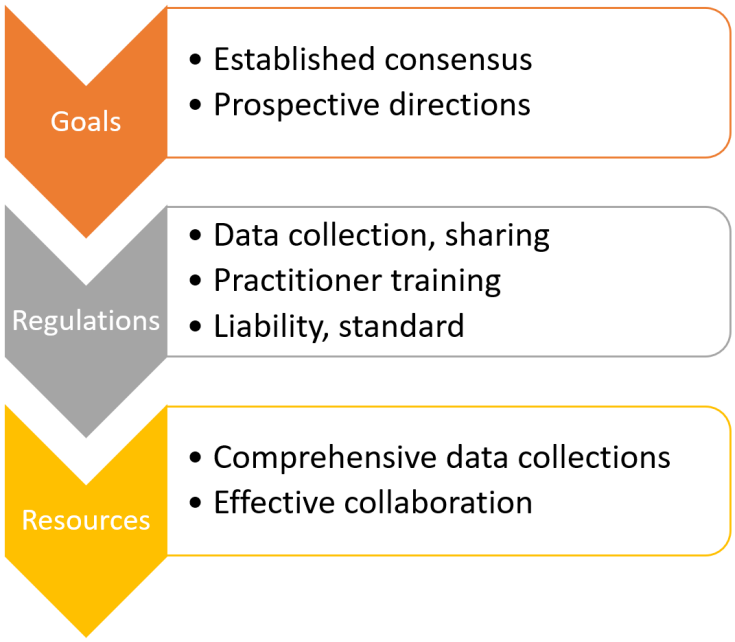

Figure 6.5: The future perspective for the medical imaging community. Although artificial intelligence is promising and expected to revolutionize the medical care industry, there are challenges and risks ahead. As these changes are associated with many parties and interests in society, it would be a long process.

and monitor diseases in a more comprehensive way.

6.2.2 The road ahead for the medical image analysis community

Now we have covered both future applications and challenges in medical image analysis, we should think about how to prepare for the future. Though medical care is an essential part of any society, medical standards, quality, affordability, and regulations have large variations across the world. The medical imaging community should evolve towards established consensus and prospective directions.

Through research and discussion, new laws should be made to rectify practices for data collection, sharing, and utilization to protect privacy and promote the medical imaging industry. Medical centers and practitioners should be trained to harness computer-

aided diagnosis and therapy facilities. Through continuous integration and coordination, the quality and efficiency of related applications are expected to improve. Broad and detailed elaboration on the liabilities of computer-aided diagnosis is needed. Humans could make mistakes under some circumstances, and a machine-learning model is no exception. Strict standards and verification processes would reduce the risk and promote reliability.

There are many opportunities for medical imaging research in the longitudinal study, multi-organ joint analysis, and multi-modality analysis. In clinical practice, the treatment for many diseases already involves comprehensive analysis, but the medical image analysis community has not seen enough related studies due to several reasons. First and foremost, there is not enough high-quality data collection. Collecting such kind of data relies on both clear medical goals and clinical support. Secondly, comprehensive disease understanding is essential to design and implementing machine learning models for multi-source tasks. Therefore it calls for close and effective collaboration between doctors and researchers, which is well-enforced in practice for various reasons.

6.2.3 The limitation of deep learning applications

How much will Artificial Intelligence replace human labor in medical image analysis tasks? This question is associated with technology development, social perception, economic factors, government regulations, and ethical concerns. From a technical perspective, it is promising to predict that machine models would well surpass humans in many medical imaging tasks, which are defined by established knowledge and practices. However,

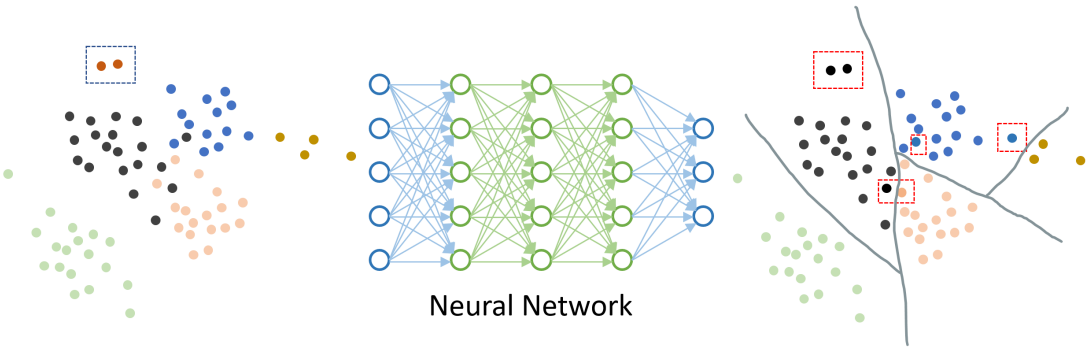

Figure 6.6: The neural network learns to classify data points in its representative space. The points on the left are the input data, where the colors represent the ground truth labels. The points on the right are the output, where the colors represent the predicted labels. The red cases are not learned due to data scarcity, one brown case is misclassified as green, due to data unbalance. Three cases (pink, black, black) are misclassified due to lying across the decision boundaries.

machine learning models have their limitations, resulting from data collection, labeling, and training. It can learn and distinguish common lesions and symptoms in radiology images quite well because of the large number of training samples. But many rare diseases do not have enough training samples, and the model could not learn the corresponding patterns.

Knowing that machine learning models could miss or misclassify some rare diseases or symptoms, it needs caution to deploy computer-aided diagnosis systems. In the screening scenarios, the main target is the suspected illness. Depending on the model's sensitivity and specificity for particular diseases, computer-aided diagnosis could replace human labor in many cases. In diagnosis scenarios, deep learning model predictions can serve as an efficient reference. However, the doctor usually needs to conduct multiple examinations to make the final judgment. Only human has the extensive knowledge and logical reasoning to specify rare cases, identify new discoveries, and make the sensible judgment in a situation of conflict.

Bibliography

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

[2] Alzheimer's Association. 2016 alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 12(4):459–509, 2016.

[3] Nicole Wright, Anne Looker, Kenneth Saag, Jeffrey Curtis, Elizabeth Delzell, Susan Randall, and Bess Dawson-Hughes. The recent prevalence of osteoporosis and low bone mass in the united states based on bone mineral density at the femoral neck or lumbar spine. *Journal of Bone and Mineral Research*, 29, 11 2014.

[4] Medical ultrasound. `https://en.wikipedia.org/wiki/Medical_ultrasound`. Accessed: 2023-02-09.

[5] X-ray. `https://en.wikipedia.org/wiki/X-ray`. Accessed: 2023-02-09.

[6] Ct scan. `https://en.wikipedia.org/wiki/CT_scan`. Accessed: 2023-02-09.

[7] Magnetic resonance imaging. `https://en.wikipedia.org/wiki/Magnetic_resonance_imaging`. Accessed: 2023-02-09.

[8] Visual system. `https://en.wikipedia.org/wiki/Visual_system`. Accessed: 2023-02-09.

[9] Yann LeCun and Yoshua Bengio. *Convolutional Networks for Images, Speech, and Time Series*, page 255–258. MIT Press, Cambridge, MA, USA, 1998.

[10] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989.

[11] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[12] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[15] Chen-I Hsieh, Kang Zheng, Chihung Lin, Ling Mei, Le Lu, Weijian Li, Fang-Ping Chen, Yirui Wang, Xiaoyun Zhou, Fakai Wang, Guotong Xie, Jing Xiao, Shun Miao, and Chang-Fu Kuo. Automated bone mineral density prediction and fracture risk assessment using plain radiographs via deep learning. *Nature Communications*, 12:5472, 09 2021.

[16] Kang Zheng, Yirui Wang, Xiao-Yun Zhou, Fakai Wang, Le Lu, Chihung Lin, Lingyun Huang, Guotong Xie, Jing Xiao, Chang-Fu Kuo, and Shun Miao. Semi-supervised learning for bone mineral density estimation in hip x-ray images. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors,

Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, pages 33–42, Cham, 2021. Springer International Publishing.

[17] Hoo-Chang Shin, Holger R. Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel J. Mollura, and Ronald M. Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, 35:1285–1298, 2016.

[18] Dakai Jin, Adam P. Harrison, Ling Zhang, Ke Yan, Yirui Wang, Jinzheng Cai, Shun Miao, and Le Lu. Chapter 14 - artificial intelligence in radiology. In Lei Xing, Maryellen L. Giger, and James K. Min, editors, *Artificial Intelligence in Medicine*, pages 265–289. Academic Press, 2021.

[19] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *arXiv:1705.02315*, 05 2017.

[20] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, AnnetteKopp-Schneider, Bennett A. Landman, Geert J. S. Litjens, Bjoern H. Menze, Olaf Ronneberger, Ronald M.Summers, Bram van Ginneken, Michel Bilello, Patrick Bilic, Patrick Ferdinand Christ, Richard K. G. Do, Marc J. Gollub, Stephan Heckers, Henkjan J. Huisman, William R. Jarnagin, Maureen McHugo, Sandy Napel, Jennifer S. Goli Pernicka, Kawal S. Rhode, Catalina Tobon-Gomez, Eugene Vorontsov, James Alastair Meakin, Sébastien Ourselin, Manuel Wiesenfarth, Pablo Arbeláez, Byeonguk Bae, Sihong Chen, Laura Alexandra Daza, Jian-Jun Feng, Baochun He, Fabian Isensee, Yuanfeng Ji, Fucang Jia, Namkug Kim, Ildoo Kim, Dorit Merhof, Akshay Pai, Beomhee Park, Mathias Perslev, Ramin Rezaiifar, Oliver Rippel, Ignacio Sarasua, Wei Shen, Jaemin Son, Christian Wachinger, Liansheng Wang, Yan Wang, Yingda Xia, Daguang Xu, Zhanwei

Xu, Yefeng Zheng, Amber L. Simpson, Lena Maier-Hein, and Manuel Jorge Cardoso. The medical segmentation decathlon. *Nature Communications*, 13, 2022.

[21] Pyqt. https://en.wikipedia.org/wiki/PyQt. Accessed: 2023-02-09.

[22] Flask. https://flask.palletsprojects.com/en/2.2.x/. Accessed: 2023-02-09.

[23] Jinja. https://jinja.palletsprojects.com/en/3.1.x/. Accessed: 2023-02-09.

[24] Tümay Sözen, Lale Özışık, and Nursel Çalık Başaran. An overview and management of osteoporosis. *European journal of rheumatology*, 4(1):46, 2017.

[25] E Michael Lewiecki, Deane Leader, Richard Weiss, and Setareh A Williams. Challenges in osteoporosis awareness and management: results from a survey of US postmenopausal women. *Journal of Drug Assessment*, 8(1):25–31, 2019.

[26] Andrew D Smith. Screening of bone density at CT: an overlooked opportunity, 2019.

[27] Xiaoguang Cheng, Kaiping Zhao, Xiaojuan Zha, Xia Du, Yongli Li, Shuang Chen, Yan Wu, Shaolin Li, Yong Lu, Yuqin Zhang, et al. Opportunistic Screening Using Low-Dose CT and the Prevalence of Osteoporosis in China: A Nationwide, Multicenter Study. *Journal of Bone and Mineral Research*, 2020.

[28] Noa Dagan, Eldad Elnekave, Noam Barda, Orna Bregman-Amitai, Amir Bar, Mila Orlovsky, Eitan Bachmat, and Ran D Balicer. Automated opportunistic osteoporotic fracture risk assessment using computed tomography scans to aid in FRAX underutilization. *Nature Medicine*, 26(1):77–82, 2020.

[29] Samuel Jang, Peter M Graffy, Timothy J Ziemlewicz, Scott J Lee, Ronald M Summers, and Perry J Pickhardt. Opportunistic osteoporosis screening at routine abdominal and thoracic CT:

normative L1 trabecular attenuation values in more than 20 000 adults. *Radiology*, 291(2):360–367, 2019.

[30] Perry J Pickhardt, Peter M Graffy, Ryan Zea, Scott J Lee, Jiamin Liu, Veit Sandfort, and Ronald M Summers. Automated abdominal CT imaging biomarkers for opportunistic prediction of future major osteoporotic fractures in asymptomatic adults. *Radiology*, 297(1):64–72, 2020.

[31] Fakai Wang, Kang Zheng, Yirui Wang, Xiaoyun Zhou, Le Lu, Jing Xiao, Min Wu, Chang-Fu Kuo, and Shun Miao. Opportunistic screening of osteoporosis using plain film chest x-ray. In Islem Rekik, Ehsan Adeli, Sang Hyun Park, and Julia Schnabel, editors, *Predictive Intelligence in Medicine*, pages 138–146, Cham, 2021. Springer International Publishing.

[32] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[33] World Health Organization et al. *Assessment of fracture risk and its application to screening for postmenopausal osteoporosis: report of a WHO study group [meeting held in Rome from 22 to 25 June 1992]*. World Health Organization, 1994.

[34] Juliet E Compston, A. L. Cooper, Cyrus Cooper, Neil Gittoes, Celia L. Gregson, Nicholas C. Harvey, Sally Hope, John A. Kanis, Eugene V. McCloskey, Kenneth E. S. Poole, D. M. Reid, Peter Selby, F. Thompson, Anne Thurston, and Norma Vine. Uk clinical guideline for the prevention and treatment of osteoporosis. *Archives of Osteoporosis*, 12, 2017.

[35] Stephanie Boutroy, Mary L. Bouxsein, Françoise Munoz, and Pierre Dominique Delmas. In vivo assessment of trabecular bone microarchitecture by high-resolution peripheral quantitative

computed tomography. *The Journal of clinical endocrinology and metabolism*, 90 12:6508–15, 2005.

[36] Joseph J. Schreiber, Paul A. Anderson, Humberto G. Rosas, Avery L. Buchholz, and Anthony G. Au. Hounsfield units for assessing bone mineral density and strength: a tool for osteoporosis management. *The Journal of bone and joint surgery. American volume*, 93 11:1057–63, 2011.

[37] Judith E. Adams. Quantitative computed tomography. *European journal of radiology*, 71 3:415–24, 2009.

[38] Sungjoon Lee, Chun Kee Chung, Songchol Oh, and Sung Bae Park. Correlation between bone mineral density measured by dual-energy x-ray absorptiometry and hounsfield units measured by diagnostic ct in lumbar spine. *Journal of Korean Neurosurgical Society*, 54:384 – 389, 2013.

[39] Yan-Lin Li, Kin Hoi Wong, Martin Wai-Ming Law, Benjamin Xin-Hao Fang, Vince Wing Hang Lau, Vince Varut Vardhanabuti, Victor Kam-Ho Lee, Andrew Kai-Chun Cheng, Wai yin Ho, and Wendy Wai Man Lam. Opportunistic screening for osteoporosis in abdominal computed tomography for chinese population. *Archives of Osteoporosis*, 13:1–7, 2018.

[40] Elena Alacreu, David Moratal, and Estanislao Arana. Opportunistic screening for osteoporosis by routine ct in southern europe. *Osteoporosis International*, 28:983–990, 2016.

[41] Bonny Specker and Eckhard Schoenau. Quantitative bone analysis in children: current methods and recommendations. *The Journal of pediatrics*, 146 6:726–31, 2005.

[42] Baroncelli and I. Giampiero. Quantitative ultrasound methods to assess bone mineral status in children: Technical characteristics, performance, and clinical application. *Pediatric Research*, 63(3):220, 2008.

[43] Paola Pisani, Maria Daniela Renna, Francesco Conversano, Ernesto Casciaro, Maurizio Muratore, Eugenio Quarta, Marco Di Paola, and

Sergio Casciaro. Screening and early diagnosis of osteoporosis through x-ray and ultrasound based techniques. *World journal of radiology*, 5 11:398–410, 2013.

[44] Kang Zheng, Yirui Wang, Xiao-Yun Zhou, Fakai Wang, Le Lu, Chihung Lin, Lingyun Huang, Guotong Xie, Jing Xiao, Chang-Fu Kuo, and Shun Miao. Semi-supervised learning for bone mineral density estimation in hip x-ray images. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 33–42, Cham, 2021. Springer International Publishing.

[45] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[46] Hoo-Chang Shin, Le Lu, Lauren Kim, Ari Seff, Jianhua Yao, and Ronald M. Summers. Interleaved text/image deep mining on a large-scale radiology database. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1090–1099, 2015.

[47] Ke Yan, Xiaosong Wang, Le Lu, and Ronald M. Summers. Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of Medical Imaging*, 5, 2018.

[48] Rikiya Yamashita, Mizuho Nishio, Richard K. G. Do, and Kaori Togashi. Convolutional neural networks: an overview and application in radiology. *Insights into Imaging*, 9:611 – 629, 2018.

[49] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7794–7803, Los Alamitos, CA, USA, jun 2018. IEEE Computer Society.

[50] Irwan Bello, Barret Zoph, Quoc Le, Ashish Vaswani, and Jonathon Shlens. Attention augmented convolutional networks. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3285–3294, 2019.

[51] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[52] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[53] Han Hu, Zheng Zhang, Zhenda Xie, and Stephen Ching-Feng Lin. Local relation networks for image recognition. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3463–3472, 2019.

[54] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 1971–1980, 2019.

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[56] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[57] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[58] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020.

[59] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3588–3597, 2018.

[60] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, Zhaohui Yang, Yiman Zhang, and Dacheng Tao. A survey on visual transformer. *ArXiv*, abs/2012.12556, 2020.

[61] Salman H. Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *arXiv preprint arXiv:2101.01169*, 2021.

[62] Mark Chen, Alec Radford, Jeff Wu, Heewoo Jun, Prafulla Dhariwal, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *ICML*, 2020.

[63] Weijian Li, Yuhang Lu, Kang Zheng, Haofu Liao, Chihung Lin, Jiebo Luo, Chi-Tung Cheng, Jing Xiao, Le Lu, Chang-Fu Kuo, et al. Structured landmark detection via topology-adapting deep graph learning. *arXiv preprint arXiv:2004.08190*, 2020.

[64] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi

Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[65] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. *arXiv preprint arXiv:2103.15808*, 2021.

[66] Ming-Shao Tsai, Meng-Hung Lin, Chuan-Pin Lee, Yao-Hsu Yang, Wen-Cheng Chen, Geng-He Chang, Yao-Te Tsai, Pau-Chung Chen, and Ying Huang Tsai. Chang gung research database: A multi-institutional database consisting of original medical records. *Biomedical Journal*, 40:263 – 269, 2017.

[67] Mirella Lopez Picazo, Alba Baro, Luis Del Rio, Silvana Digregorio, Yves Martelli, Jordi Romera, Martin Stephofer, Miguel Angel Gonzlez Ballester, and Ludovic Humbert. 3-d subject-specific shape and density estimation of the lumbar spine from a single anteroposterior dxa image including assessment of cortical and trabecular bone. *IEEE Transactions on Medical Imaging*, PP:1–1, 06 2018.

[68] Sami P. Väänänen, Lorenzo Grassi, Gunnar Flivik, Jukka S. Jurvelin, and Hanna Isaksson. Generation of 3d shape, density, cortical thickness and finite element mesh of proximal femur from a dxa image. *Medical Image Analysis*, 24(1):125–134, 2015.

[69] Omar Ahmad, Krishna Ramamurthi, Kevin E Wilson, Klaus Engelke, Richard L Prince, and Russell H Taylor. Volumetric dxa (vxa): A new method to extract 3d information from multiple in vivo dxa images. *Journal of Bone and Mineral Research*, 25(12):2744–2751, 2010.

[70] Ludovic Humbert, Yves Martelli, Roger Fonollà, Martin Steghöfer, Silvana Di Gregorio, Jorge Malouf, Jordi Romera, and Luis Miguel Del Río Barquero. 3d-dxa: Assessing the femoral shape, the trabecular macrostructure and the cortex in 3d from dxa images. *IEEE Transactions on Medical Imaging*, 36(1):27–39, 2017.

[71] Barbara Campolina Silva, William D. Leslie, Heinrich Resch, Olivier Lamy, Olga M. Lesnyak, Neil Binkley, Eugene V. McCloskey, John A. Kanis, and John P. Bilezikian. Trabecular bone score: A noninvasive analytical method based upon the dxa image. *Journal of Bone and Mineral Research*, 29, 2014.

[72] Valérie Bousson, Catherine Bergot, Bruno Sutter, Pierre Levitz, Bernard Cortet, and the Scientific Committee of the Grio. Trabecular bone score (tbs): available knowledge, clinical relevance, and future prospects. *Osteoporosis International*, 23:1489–1501, 2011.

[73] Nicholas C. Harvey, Claus C. Glüer, Neil Binkley, E. V. McCloskey, Ml. Brandi, Cyrus Cooper, David Kendler, O. Lamy, Andrea Laslop, Bruno Muzzi Camargos, Jacques Reginster, René Rizzoli, and John A. Kanis. Trabecular bone score (tbs) as a new complementary approach for osteoporosis evaluation in clinical practice. *Bone*, 78:216–24, 2015.

[74] MICCAI CHALLENGES. CSI 2014 Vertebra Localization and Identification Dataset. `http://spineweb.digitalimaginggroup.ca/Index.php?n=Main.Datasets#Dataset_3.3A_Vertebrae_Localization_and_Identification`, 2014.

[75] H. Liao, A. Mesfin, and J. Luo. Joint vertebrae identification and localization in spinal ct images by combining short- and long-range contextual information. *IEEE Transactions on Medical Imaging*, 37(5):1266–1275, 2018.

[76] Dong Yang, Tao Xiong, Daguang Xu, Qiangui Huang, David Liu, S. Kevin Zhou, Zhoubing Xu, Jin Park, Mingqing Chen, Trac Tran, Sang Chin, Dimitris Metaxas, and Dorin Comaniciu. Automatic

vertebra labeling in large-scale 3d ct using deep image-to-image network with message passing and sparsity regularization. pages 633–644, 05 2017.

[77] Yizhi Chen, Yunhe Gao, Kang Li, Liang Zhao, and Jun Zhao. Vertebrae identification and localization utilizing fully convolutional networks and a hidden markov model. *IEEE TMI*, 07 2019.

[78] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[79] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4732, 2016.

[80] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3476–3483, 2013.

[81] J. Lv, X. Shao, J. Xing, C. Cheng, and X. Zhou. A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3691–3700, 2017.

[82] Shizhan Zhu, Cheng Li, C. C. Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4998–5006, 2015.

[83] W. Yu, X. Liang, K. Gong, C. Jiang, N. Xiao, and L. Lin. Layout-graph reasoning for fashion landmark detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2932–2940, 2019.

[84] Ben Glocker, J. Feulner, Antonio Criminisi, D. R. Haynor, and E. Konukoglu. Automatic localization and identification of vertebrae in arbitrary field-of-view ct scans. In *MICCAI 2012*, pages 590–598. Springer.

[85] Ben Glocker, Darko Zikic, Ender Konukoglu, David R. Haynor, and Antonio Criminisi. Vertebrae localization in pathological spine ct via dense classification from sparse annotations. In *MICCAI 2013*, pages 262–270. Springer.

[86] Yiqiang Zhan, Dewan Maneesh, Martin Harder, and Xiang Sean Zhou. Robust mr spine detection using hierarchical learning and local articulated model. In *International conference on medical image computing and computer-assisted intervention*, pages 141–148. Springer, 2012.

[87] Hao Chen, Chiyao Shen, Jing Qin, Dong Ni, Lin Shi, Jack C. Y. Cheng, and Pheng-Ann Heng. Automatic localization and identification of vertebrae in spine ct via a joint learning model with deep neural networks. In *MICCAI 2015*, pages 515–522. Springer.

[88] Amin Suzani, Alexander Seitel, Yuan Liu, Sidney Fels, Robert N Rohling, and Purang Abolmaesumi. Fast automatic vertebrae detection and localization in pathological ct scans-a deep learning approach. In *International conference on medical image computing and computer-assisted intervention*, pages 678–686. Springer, 2015.

[89] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *CoRR*, abs/1411.4038, 2014.

[90] Chunli Qin, D. Yao, Han Zhuang, H. Wang, Yong-Hong Shi, and Z. Song. Residual block-based multi-label classification and localization network with integral regression for vertebrae labeling. *ArXiv*, abs/2001.00170, 2020.

[91] Roman Jakubicek, Jiri Chmelik, Jiří Jan, Petr Ourednicek, Lukas Lambert, and Giampaolo Gavelli. Learning–based vertebra localization and labeling in 3d ct data of possibly incomplete and pathological spines. *Computer Methods and Programs in Biomedicine*, 183:105081, 09 2019.

[92] James McCouat and B. Glocker. Vertebrae detection and localization in ct with two-stage cnns and dense annotations. *ArXiv*, abs/1910.05911, 2019.

[93] Rhydian Windsor, Amir Jamaludin, Timor Kadir, and Andrew Zisserman. A convolutional approach to vertebrae detection and labelling in whole spine mri, 2020.

[94] T. van Sonsbeek, P. Danaei, D. Behnami, M. H. Jafari, P. Asgharzadeh, R. Rohling, and P. Abolmaesumi. End-to-end vertebra localization and level detection in weakly labelled 3d spinal mr using cascaded neural networks. In *ISBI*, pages 1178–1182, 2019.

[95] Anjany Sekuboyina, Markus Rempfler, Jan Kukacka, Giles Tetteh, Alexander Valentinitsch, Jan S. Kirschke, and Bjoern H. Menze. Btrfly net: Vertebrae labelling with energy-based adversarial learning of local spine prior. *CoRR*, abs/1804.01307, 2018.

[96] Stuart Russell and Peter Norvig. Artificial intelligence: a modern approach. 2002.

[97] Fabian Isensee, Paul F Jäger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. Automated design of deep learning methods for biomedical image segmentation. *arXiv preprint arXiv:1904.08128*, 2019.

[98] Fabian Isensee. nnU-Net implementation github. `https://github.com/MIC-DKFZ/nnUNet`, 2020.

[99] Harriet Rumgay, Melina Arnold, Jacques Ferlay, Olufunmilayo Lesi, Citadel J. Cabasag, Jérôme Vignat, Mathieu Laversanne, Katherine A. McGlynn, and Isabelle Soerjomataram. Global burden of primary liver cancer in 2020 and predictions to 2040. *Journal of Hepatology*, 2022.

[100] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. *ArXiv*, abs/1606.06650, 2016.

[101] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 2020.

[102] et al Masao Omata. Asia–pacific clinical practice guidelines on the management of hepatocellular carcinoma: a 2017 update. *Hepatology International*, 11(4):317–370, July 2017.

[103] Nam Yu, Vinika Chaudhari, Steven Raman, Charles Lassman, Myron Tong, Ronald Busuttil, and David Lu. Ct and mri improve detection of hepatocellular carcinoma, compared with ultrasound alone, in patients with cirrhosis. *Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association*, 9:161–7, 10 2010.

[104] Cher Heng Tan, Su-Chong Low, and Choon Hua Thng. Apasl and aasld consensus guidelines on imaging diagnosis of hepatocellular carcinoma: A review. *International journal of hepatology*, 2011:519783, 01 2011.

[105] Lewis Rowland Roberts, Claude B. Sirlin, Feras Zaiem, Jehad Almasri, Larry J. Prokop, Julie K. Heimbach, Mohammad Hassan Murad, and Khaled Mohammed. Imaging for the diagnosis of hepatocellular carcinoma: A systematic review and meta-analysis. *Hepatology*, 67, 2018.

[106] Tiffany Hennedige. Advances in computed tomography and magnetic resonance imaging of hepatocellular carcinoma. *World Journal of Gastroenterology*, 22:205, 01 2016.

[107] Easl clinical practice guidelines: Management of hepatocellular carcinoma. *Journal of Hepatology*, 69(1):182–236, 2018.

[108] et al. Chernyak, Victoria. Liver imaging reporting and data system (li-rads) version 2018: Imaging of hepatocellular carcinoma in at-risk patients. *Radiology*, 289(3):816–830, 2018. PMID: 30251931.

[109] Jorge A. Marrero, Laura M. Kulik, Claude B. Sirlin, Andrew X. Zhu, Richard S. Finn, Michael M. Abecassis, Lewis R. Roberts, and Julie K. Heimbach. Diagnosis, staging, and management of hepatocellular carcinoma: 2018 practice guidance by the american association for the study of liver diseases. *Hepatology*, 68(2):723–750, 2018.

[110] Khaled Bousabarah, Brian Letzen, Jonathan Tefera, Lynn Savic, Isabel Schobert, Todd Schlachter, Lawrence Staib, Martin Kocher, Julius Chapiro, and MingDe Lin. Automated detection and delineation of hepatocellular carcinoma on multiphasic contrast-enhanced mri using deep learning. *Abdominal Radiology*, 46, 01 2021.

[111] Qiangguo Jin, Zhao-Peng Meng, Changming Sun, Leyi Wei, and Ran Su. Ra-unet: A hybrid deep attention-aware network to extract liver and tumor in ct scans. *Frontiers in Bioengineering and Biotechnology*, 8, 2020.

[112] Miriam Bellver, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Xavier Giró i Nieto, Jordi Torres, and Luc Van Gool. Detection-aided liver lesion segmentation using deep learning. *ArXiv*, abs/1711.11069, 2017.

[113] Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Geert Litjens, Paul Gerke, Colin Jacobs, Sarah J. van Riel, Mathilde Marie Winkler Wille, Matiullah Naqibullah, Clara I. Sánchez, and Bram van Ginneken. Pulmonary nodule detection in ct images: False positive reduction using multi-view convolutional networks. *IEEE Transactions on Medical Imaging*, 35(5):1160–1169, 2016.

[114] Jiarong Zhou, Wenzhe Wang, Biwen Lei, Wenhao Ge, Yu Huang, Linshi Zhang, Yingcai Yan, Dongkai Zhou, Yuan Ding, Jian Wu, and Weilin Wang. Automatic detection and classification of focal liver lesions based on deep convolutional neural networks: A preliminary study. *Frontiers in Oncology*, 10, 2021.

[115] Maayan Frid-Adar, Eyal Klang, Michal Marianne Amitai, Jacob Goldberger, and Hayit Greenspan. Synthetic data augmentation using gan for improved liver lesion classification. *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 289–293, 2018.

[116] Feng Shi, Bojiang Chen, Qiqi Cao, Ying Wei, Qing Zhou, Rui Zhang, Yaojie Zhou, Wenjie Yang, Xiang Wang, Rongrong Fan, Fan Yang, Yanbo Chen, Weimin Li, Yaozong Gao, and Dinggang

Shen. Semi-supervised deep transfer learning for benign-malignant diagnosis of pulmonary nodules in chest ct images. *IEEE Transactions on Medical Imaging*, 41(4):771–781, 2022.

[117] Onur Ozdemir, Rebecca L. Russell, and Andrew A. Berlin. A 3d probabilistic deep learning system for detection and diagnosis of lung cancer using low-dose ct scans. *IEEE Transactions on Medical Imaging*, 39:1419–1429, 2020.

[118] Hongtao Xie, Dongbao Yang, Nannan Sun, Zhineng Chen, and Yongdong Zhang. Automated pulmonary nodule detection in ct images using deep convolutional neural networks. *Pattern Recognition*, 85:109–119, 2019.

[119] Xiao Han. Automatic liver lesion segmentation using a deep convolutional neural network method. *ArXiv*, abs/1704.07239, 2017.

[120] Patrick Ferdinand Christ, Mohamed Ezzeldin A. Elshaer, Florian Ettlinger, Sunil Tatavarty, Marc Bickel, Patrick Bilic, Markus Rempfler, Marco Armbruster, Felix O. Hofmann, Melvin D'Anastasi, Wieland H. Sommer, Seyed-Ahmad Ahmadi, and Bjoern H. Menze. Automatic liver and lesion segmentation in ct using cascaded fully convolutional neural networks and 3d conditional random fields. *ArXiv*, abs/1610.02177, 2016.

[121] Omar Ibrahim Alirr. Deep learning and level set approach for liver and tumor segmentation from ct scans. *Journal of Applied Clinical Medical Physics*, 21:200 – 209, 2020.

[122] Koichiro Yasaka, Hiroyuki Akai, Osamu Abe, and Shigeru Kiryu. Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced ct: A preliminary study. *Radiology*, 286 3:887–896, 2018.

[123] Samy A. Azer. Deep learning with convolutional neural networks for identification of liver masses and hepatocellular carcinoma: A systematic review. *World Journal of Gastrointestinal Oncology*, 11:1218 – 1230, 2019.

[124] Grzegorz Chlebus, Andrea Schenk, Jan Hendrik Moltz, Bram van Ginneken, Horst K. Hahn, and Hans Meine. Automatic liver tumor segmentation in ct with fully convolutional neural networks and object-based postprocessing. *Scientific Reports*, 8, 2018.

[125] Lu Meng, Qianqian Zhang, and Sihang Bu. Two-stage liver and tumor segmentation algorithm based on convolutional neural network. *Diagnostics*, 11(10), 2021.

[126] Yue Zhang, Benxiang Jiang, Jiong Wu, Dongcen Ji, Yilong Liu, Yifan Chen, Ed. X. Wu, and Xiaoying Tang. Deep learning initialized and gradient enhanced level-set based segmentation for liver tumor from ct images. *IEEE Access*, 8:76056–76068, 2020.

[127] Grzegorz Chlebus, Andrea Schenk, Jan Moltz, Bram Ginneken, Hans Meine, and Horst Hahn. Automatic liver tumor segmentation in ct with fully convolutional neural networks and object-based postprocessing. *Scientific Reports*, 8, 10 2018.

[128] Chi-Tung Cheng, Jinzheng Cai, Wei Teng, Youjing Zheng, Yu-Ting Huang, Yu-Chao Wang, Chien-Wei Peng, Youbao Tang, Wei-Chen Lee, Ta-Sen Yeh, Jing Xiao, Le Lu, Chien-Hung Liao, and Adam P. Harrison. A flexible three-dimensional heterophase computed tomography hepatocellular carcinoma detection algorithm for generalizable and practical screening. *Hepatology Communications*, 6(10):2901–2913, 2022.

[129] Amber L. Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram van Ginneken, Annette Kopp-Schneider, Bennett A. Landman, Geert J. S. Litjens, Bjoern H. Menze, Olaf Ronneberger, Ronald M. Summers, Patrick Bilic, Patrick Ferdinand Christ, Richard Kinh Gian Do, Marc J. Gollub, Jennifer Golia-Pernicka, Stephan Heckers, William R. Jarnagin, Maureen McHugo, Sandy Napel, Eugene Vorontsov, Lena Maier-Hein, and M. Jorge Cardoso. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *ArXiv*, abs/1902.09063, 2019.

[130] Jun Ma, Yao Zhang, Song Gu, Yichi Zhang, Cheng Zhu, Qiyuan Wang, Xin Liu, Xingle An, Cheng Ge, Shucheng Cao, Qi Zhang, Shangqing Liu, Yunpeng Wang, Yuhui Li, Congcong Wang, Jian He, and Xiaoping Yang. Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:6695–6714, 2022.

[131] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors. *Semi-Supervised Learning*. The MIT Press, 2006.

[132] Yuankai Huo, Jinzheng Cai, Chi-Tung Cheng, Ashwin Raju, Ke Yan, Bennett A. Landman, Jing Xiao, Le Lu, Chien-Hung Liao, and Adam P. Harrison. Harvesting, detecting, and characterizing liver lesions from large-scale multi-phase ct data via deep dynamic texture learning. *ArXiv*, abs/2006.15691, 2020.

[133] Ke Yan, Jinzheng Cai, Youjing Zheng, Adam P. Harrison, Dakai Jin, Youbao Tang, Yuxing Tang, Lingyun Huang, Jing Xiao, and Le Lu. Learning from multiple datasets with heterogeneous and partial labels for universal lesion detection in ct. *IEEE Transactions on Medical Imaging*, 40(10):2759–2770, 2021.

[134] Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–20, 2022.

[135] Mattias Paul Heinrich, Mark Jenkinson, Bartlomiej W. Papież, Sir Michael Brady, and Julia A. Schnabel. Towards realtime multimodal fusion for image-guided interventions using self-similarities. In *MICCAI 2013*, pages 187–194, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[136] Juan Cerrolaza, Mirella López Picazo, Ludovic Humbert, Yoshinobu Sato, Daniel Rueckert, Miguel Ángel González Ballester, and Marius George Linguraru. Computational anatomy for multi-organ analysis in medical imaging: A review. *Medical Image Analysis*, 56, 05 2019.