ABSTRACT

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

Fakai Wang
Doctor of Philosophy, 2022

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, cost of medical resources. Thanks to the development of machine learning and computer vision in recent years, Artificial Intelligence (AI) can solve part of medical problems and improve people's health. The diagnosis of various diseases (such as Rheumatoid Arthritis, spine disorders, low bone mineral density or osteoporosis, liver cancer) relies on X-rays or Computed Tomography (CT). AI models could automatically analyze these radiography scans and make reliable diagnoses, which greatly increase efficiency, lower the cost, and improve the quality. Different organs and diseases have distinct characteristics, requiring customized algorithms and models. As the Computer Aided-Diagnosis (CAD) tools aim to assist doctors, the models must be in compliance with clinical practice. Therefore careful data curation and close collaboration with hospitals are essential to the success of AI

applications. In this dissertation, we investigate several CAD tasks and present the AI solutions. There are numerous advantages of applying AI (deep learning) in medical imaging tasks. The first advantage is efficiency. Reading radiography is tedious, requiring years' expertise. This is particularly true when radiologists need to localize tumors in hundreds of CT slices. Deep learning models could localize and identify the tumors within seconds, greatly reducing human labor. Secondly, it contributes to affordable health care. Rheumatoid Arthritis (RA) diagnosis requires orthopedists to visually check hand X-rays and classify the joint conditions, but experienced RA doctors are mostly in shortage. After gaining expertise via thousands of patient cases, machine interpretations are comparable to professional doctors'. With the low-cost deployment, high-quality and affordable medical diagnoses become more accessible. Thirdly, it leads to better disease screening and consequently better health management. The early detection of cancer and chronic diseases is critical to subsequent treatment. However, due to the low awareness and high expense, people lack adequate body screening. AI models enable opportunistic screening by reusing medical images originally taken for other purposes. For example, we develop deep learning models that utilize chest film to predict bone mineral density, which could alarm osteoporosis and help prevent bone fractures. Attention should be paid to AI limitations. Firstly, AI models lack explainability. Deep learning models store diagnosis knowledge and statistical patterns in their parameters, which are obscure to humans. The AI excellence depends on huge amounts of training data, while humans can learn and explain disease through logical reasoning. Secondly, uncertainty exists for rare diseases. If not exposed to rare cases, the models would yield uncertain outcomes. Thirdly, AI models are subject to standardized methods and clinical practice. Many diseases are

complex and rely on multiple examinations. What's more, the data curation process is based on established examinations, thus making the model susceptible to label errors. In summary, AI models for medical diagnosis are promising despite the challenges and issues. With the fast technological advancements and social development, AI will make more contributions to human health.

Table of Contents

Bibliography 139

Chapter 1

Introduction

As an important part of Computer-Aided Diagnosis (CAD), medical image analysis feeds in medical scanning and applies machine learning techniques to recognize clinically valuable clues automatically to guide doctors in diagnosis, treatment, operation planing. While medical scanning has been ubiquitously applied for the 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 the imaging analysis for improved efficiency and quality, but the process has been slowed down by the lack of adequate computer algorithms and models. With the emergency of advanced computing hardware, large scale image datasets, deep learning models, computer vision related areas have been revolutionized in the past 15 years. Deep learning models outperform human in many vision recognition tasks, and the same story goes for many medical imaging problems.

Precise and accurate as it requires, medical imaging tasks depend not only on computer vision technologies, but more importantly on quality data collection and medical 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 for 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 in terms of target types, including regression, classification, localization, segmentation, detection.

Before we delve into a particular task, we look at the general concepts, goals, 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 some incentives for each of the medical imaging problems. The readers can go to individual topics (Chapter 3,4,5) directly if interested. Chapter 2 gives more emphasis on development of computer vision and deep learning, as well as the technological aspects of medical imaging analysis. The readers can find these fundamental knowledge in many online resources, and we simply put them here for easy reference.

1.1 Medical Image analysis and computer vision

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

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

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

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

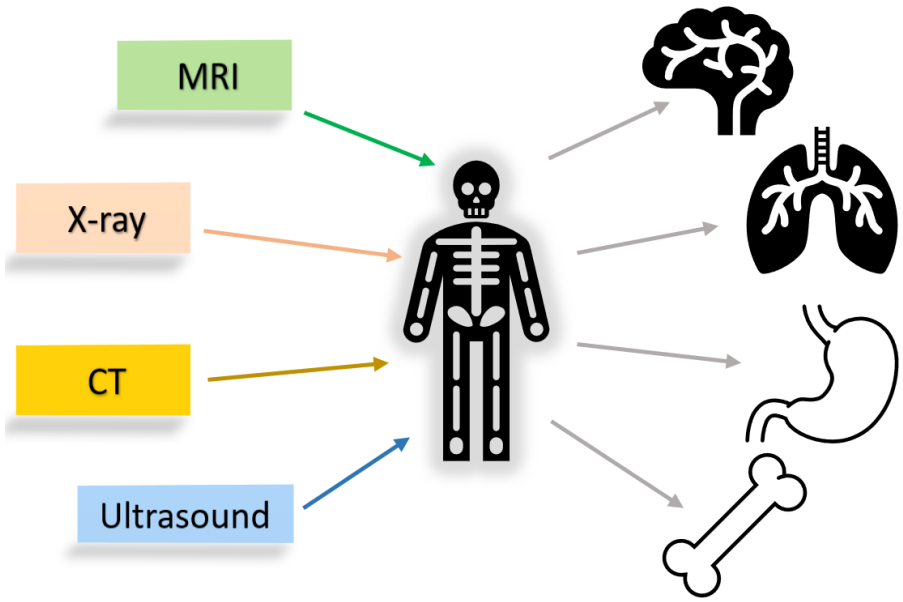

Figure 1.1: Medical imaging utilizes different scanning modalities on a variety of body parts. Many diseases manifest structural changes inside the body, which can be captured by Ultrasonography, X-ray, MRI, CT. In the diagnosis scenario, the scanning usually focuses on certain organs with specific imaging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

1.1.1 Medical image analysis tasks

Medical imaging has been widely used for the diagnosis and treatment of many fatal disease, such as cancer and Alzheimer's Disease (AD). Both cancer and AD would develop lesion or texture changes inside the organs for a period of time, before showing symptoms. Before the signs or symptoms appear, the initial recognition of these diseases are most likely through screening, including blood test, radiology scanning. A pathology study of organ tissues through biopsy is required to make final confirmation, in many cases.

According to the Global Cancer Statistics [1], an estimated of 19.3 million new cancer cases and almost 10.0 million cancer deaths occurred in 2020. The 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 earlier stage, making room for better progressive treatment. X-ray and CT images have been well adopted to screen many cancers, such as breast cancer, lung cancer, colorectal cancer, liver cancer, 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 cognitive degenerative disease, characterized by memory loss and brain changes. As neurons die and connections breaks 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 can be enough for doctors to tell whether Alzheimer's Disease exists. However, Alzheimer's Disease may affect the brain years before the symptom manifestation, 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 body checkups, patients take examinations such as chest X-ray, abdominal ultrasonography. If the technician finds any abnormal changes in the body, further tests would be arranged. Another example is Rheumatoid Arthritis (RA) classification, where the doctor needs to examine hand X-ray or foot X-ray to determine the severity of RA. These process all call for specially trained labors to

examine with great care. Computer vision workflows and machine learning models have high potential to improve both efficiency and quality.

1.1.2 Deep learning advancements in computer vision

Deep learning methodologies have dominated the computer vision research nowadays, and outperform humans in various benchmarks. In our society, many duties or jobs can be partly for wholly formulated as task-specific computer vision problems. For example, facial recognition in the cell phone unlocking function are based on computer vision and machine learning technologies, and the multiple process can be decomposed into visionary information extraction, feature registration, characteristic comparison and identity determination. Another example is the Optical Character Recognition, where machines can instantly transform text-containing images into sequence of words and sentences, ready for language translation and other instructions. In both the 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, convolutional layer have been proposed more than 30 years (before 1990), machine vision does not evolve into being widely-usable 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 the social media make it possible to create large scale image datasets, such as ImageNet (2008), MS COCO (2014). The fast advancement of General Purpose GPU and computational libraries (CUDA toolkit,

cuDNN library, Torch, Caffe, etc.) significantly reduce the technological difficulties for the wide adoption and research of deep learning. Harnessing all these hardware, software, and a thriving deep learning community, 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 several long-standing challenges. New layers (convolutional, dropout, ReLu, skip connection, etc.) have solved the computational problems in neural networks, such as the local pattern learning, gradient vanishing, overfitting. New loss functions (weight decay, Kullback–Leibler, Dice similarity, etc.) have enabled improved training process 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 largely solve computer vision tasks which can be decomposed into 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.

1.1.3 The formulation of medical imaging tasks

Medical image analysis has several differences from general computer vision. In medical imaging domain, the study is fixed in body parts, with limited types of imaging modalities which have standard scanning definitions. The main purpose of medical imaging

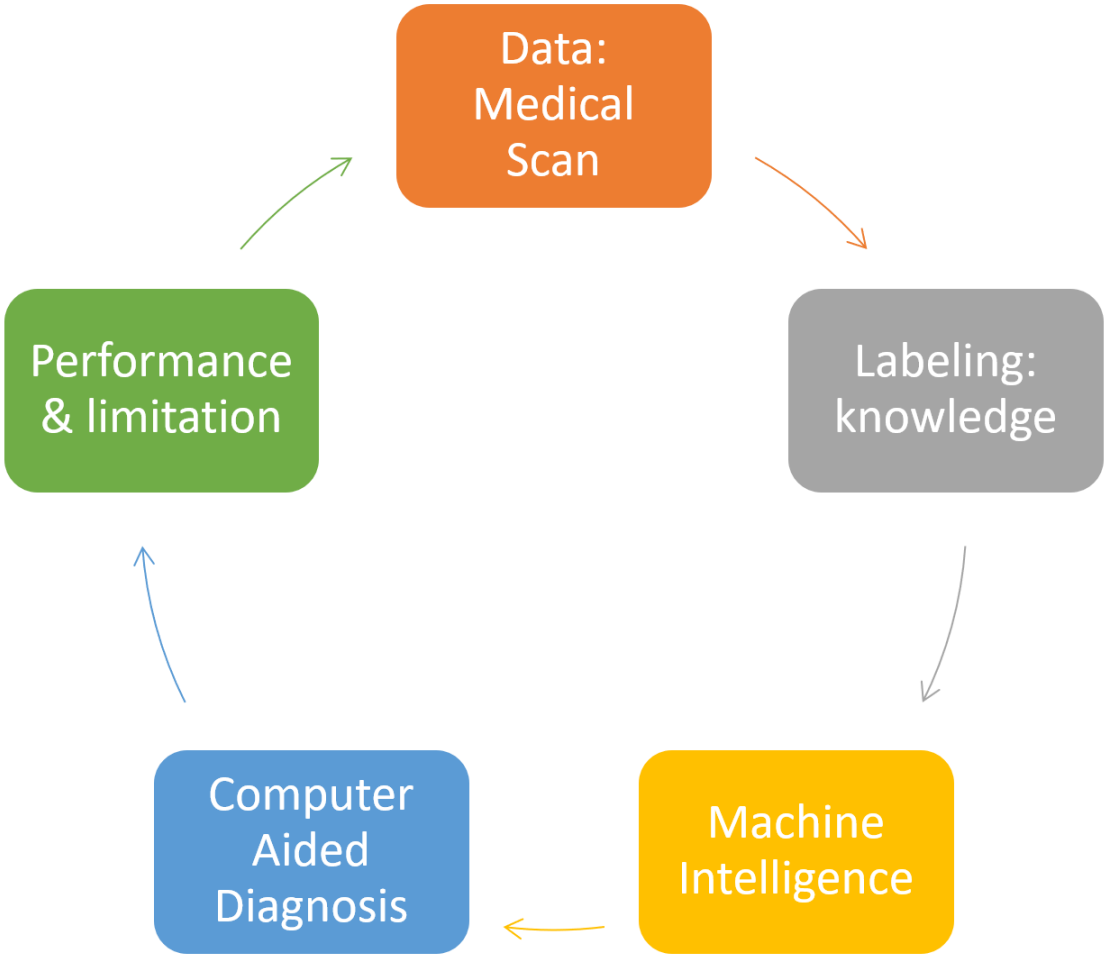

Figure 1.2: Procedures involved in medical imaging task. The data acquisition, labeling, model formulation, application integration, clinical verification all require knowledge and consideration in the medical domain.

is to learn the visual patterns of predefined lesions or diseases, while the purposes of general computer vision range from object classification, detection, segmentation to action recognition, behavior analysis. Though the task is much simpler in medical image analysis, the requirements of precision and interpretability are much higher. Cross-center 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 make the final decision.

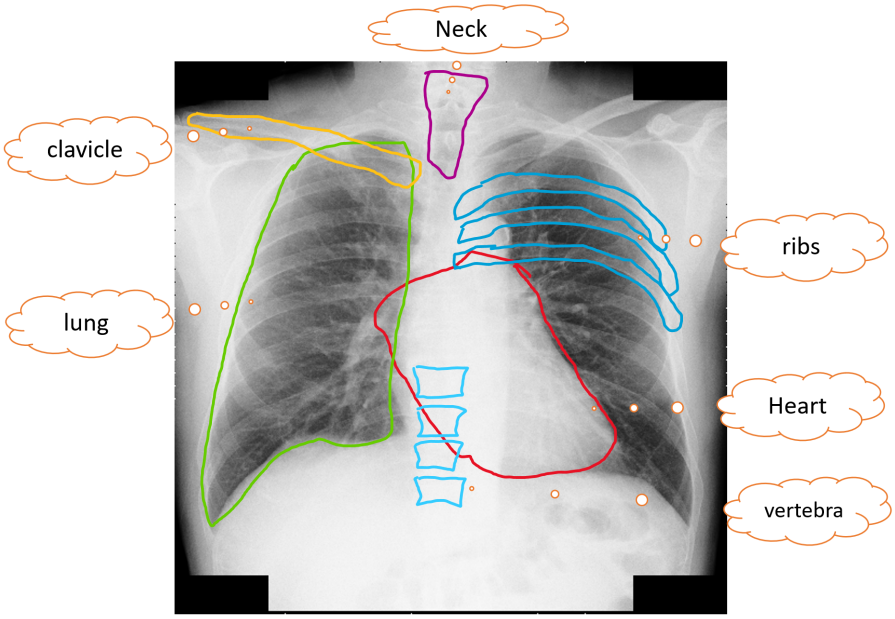

Figure 1.3: Chest X-ray covers multiple organs, bones, which can serve multiple purposes in the screening scenario. Radiologists can examine if any part goes wrong. Many anomaly can be spotted, such as rib fractures, cardiomegaly, compression fracture in spine, lung nodules.

Data collection and sharing regulations are of special concern for medical imaging tasks. Due 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 only some portion of hospital patients give permit of 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, the implementation. Based on private data and specifically defined task, deep learning models may not be directly compared 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 cover a large region of body 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 scenario, machine models are expected to not only detect 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, to tell the lesion types, to 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 the future disease development. In order to predict disease evolving, 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 general, environmental and genetic specification of the patient, 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 for labeling quality. As organs would have many associated diseases, medical experts with professional knowledge is need to determine lesion types. For some datasets, the labeling process should be guided by radiology and pathology reports, which provide insights from radiologists and histologists. Sometimes human eyes could not

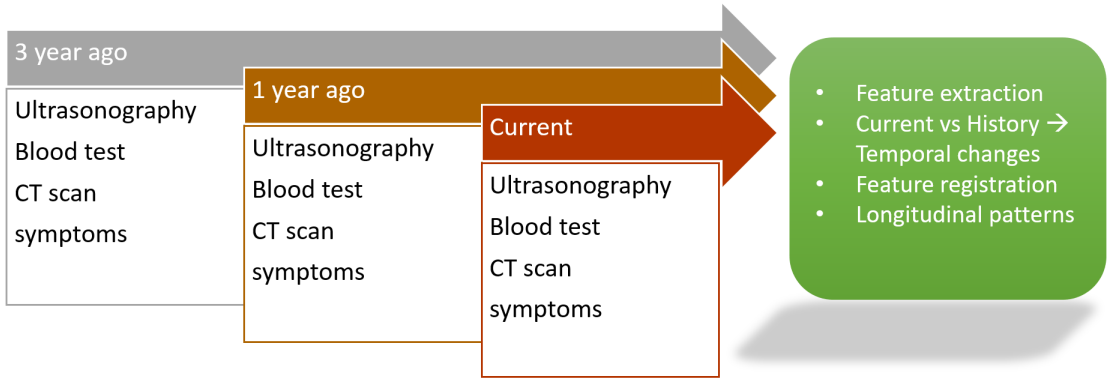

Figure 1.4: The longitudinal study can be designed in various ways. For some chronic diseases, periodic monitoring of body changes are necessary. For example, older patients with Hepatic B Virus are advised to take examination of the liver at some interval by the doctor, multiple modalities (including ultrasound, blood test, 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.

distinguish lesions by checking the medical images, but the representative visual pattern may actually exits. Given correctly labeled lesion masks or bounding boxes, deep learning models can pick up such subtle appearance nuances in the embedded features.

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

Different from other vision tasks, medical imaging models goes 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 data

from multiple medical centers in a large scale, models could not be considered fully representative. 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, who make the final decision.

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 tasks are all based on close collaboration between deep learning experts and experienced hospital doctors. The trained models in the experiments have been delivered to doctors for larger scale verification. The principles of different stages are consistent with real-world requirements. In each tasks, 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 metabolic disease widely affecting older people, characterized by extremely low Bone Mineral Density (BMD). In 2010 the adults 50 years and older in the US have an overall 10.3% prevalence of osteoporosis, and it is estimated that 10.2 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 heights, fatigue, but people tend to ignore these symptoms due to a lack of

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 device manufacturers. The vendor-dependent BMD scores are normalized into standard T-score range, which can be used to determine patients' bone density status.

awareness. As a silent disease, osteoporosis can cause serious injuries before the patients finally get aware. The bone become fragile when the mineral density is below certain threshold, which may easily lead to bone fractures.

In current clinical practice, Dual-energy X-ray Absorptiometry, DEXA or 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, BMD are far from 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 is still no clinically verified applications deployed in scale. What's more, many of the non-DXA methodologies suffer from critical drawbacks.

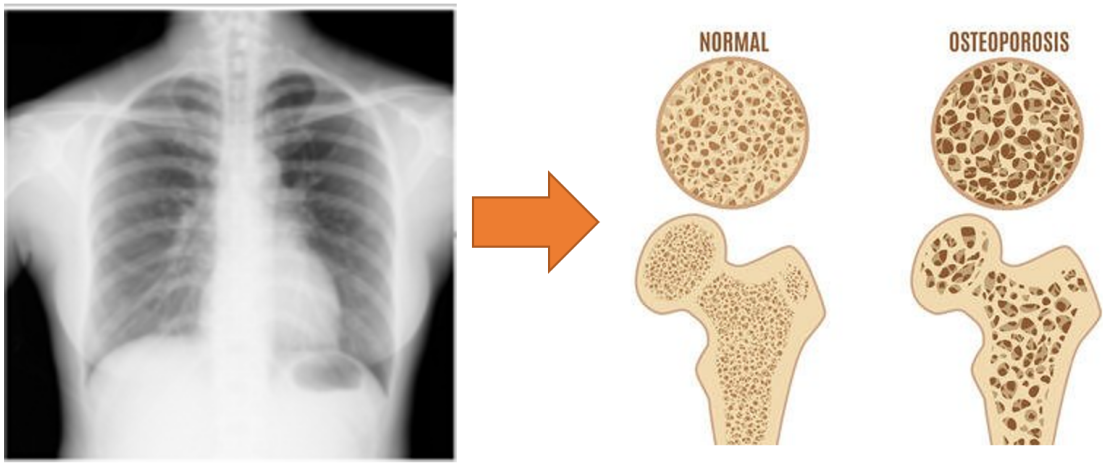

Figure 1.6: The purpose of chest X-ray BMD project is to investigate if we can predict bone density status from the plain film.

One effective way of improve the health in the whole society is to screen disease during opportunistic scanning. As the economic develops, normal body checkup 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 prediction 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 which 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 desired to help. In order to be more robust to all kinds of diseases, both the regional and the global information need to be harness to 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 the 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 region of interest (ROI) are utilized to predict BMD and calculate the loss. During inference, only the predicted BMD from the global feature are 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 achieve 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 web-UI 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 though web browsers.

1.2.2 Spine vertebra localization and identification via CT

The human spine is an important structure for the body motion. The sequential vertebra and surrounding mussels are responsible for a variety of gestures and action support. The spine also protect the spinal cord, which is part of the central nervous system. The spine diseases include injuries, infections, aging related bone changes. Orthopedists need to first localize and identify the vertebra in medical image scans for the further diagnosis. Due the vertebra similarities in terms of bone structure and contexts, machine learning models tend to assign wrong identities. 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 the precise segmentation and improved detection of vertebra, doctors can conduct examination and diagnosis more efficiently.

The movable part of the human spine contains 24 vertebra, including 7 cervical vertebra (C1 - C7), 12 thoracic vertebra (T1 - T12), 5 lumbar vertebra (L1 - L5). The sacrum and coccyx vertebra are fused 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 vertebra, such as C7, T10, L5. But the other vertebra are not easily recognized because of the appearance similarities with neighboring vertebra.

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. Small field of view contains less landmark structures, which can be used to anchor the vertebra sequence. Some patients have metal implantation or severe spine curvature, which impede correct localization. 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 vertebra 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. Afterwards, we transform the 3D vertebra centroid probability maps into 1D signal, retaining the spatial distances between vertebra. In the straightened signal line, vertebra centroids and discs are represented as peaks and valleys. A final sequence of vertebra 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 small field of view or there exist extreme spine curvatures which would cause ambiguities. Some people may have abnormal number of vertebra in the spine, and this would also cause algorithm failure.

1.2.3 Liver tumor segmentation and detection from CT images

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

As a large organ in terms of both scale of functionalities, the liver diseases develop in various and complex ways, and it can go into several different stages of texture changes before the malignant tumors actually form. Common liver texture changes include fatty liver, fibrosis, cirrhosis. There are many types of lesions in the liver, and the most common ones include cysts, hemangioma, focal nodular hyperplasia (FNH). There are mainly three types of malignant tumors in liver, namely Hepatic Cellular Carcinoma (HCC), Cholangio, 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. Theses all add up to the detection difficulties for liver lesions.

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

images for training and validation purposes. As the liver tumor labeling needs professional knowledge and pathology support, close collaboration with liver experts is a must. Due to the privacy regulations and data sharing restrictions, there is no public liver CT datasets with 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, customized labeling tools for the curation. At the end we obtain 1633 patient cases, with four-phase CT images having segmentation masks of 7 organs and 6 lesions.

We develop the 2-stage liver tumor detection workflow which both high sensitivity and specificity for malignancy detection. We adopt the 3D U-Net for the organ segmentation and lesion segmentation in the first place, which gives us the probability maps for individual classes. Doubling the lesion 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 the context-aware lesion augmentation, which randomly combine the lesion patch with non-lesion patch of the save liver. In this way, the lesion model learns better ability of distinguishing tree lesion from suspect textures. The segmentation result of the first model and the lesion reclassification result of the second model are combined to form a consensus, which yield best result.

We experiment and test our proposed working pipeline in both the four-phase setting and the non-contrast setting. In the diagnosis scenario, doctors would use all four CT phases to identify the liver tumor types, guiding the further treatment. In the non-contrast

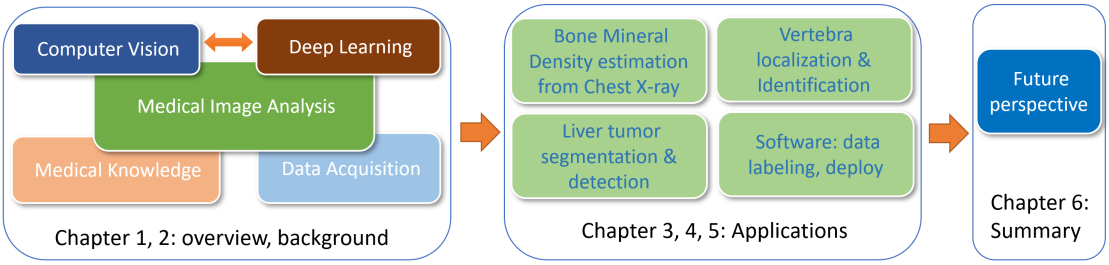

Figure 1.7: The arrangement of all the chapters. The first part gives the reader the necessary background, including medical image analysis and related technologies. The second part contains three projects, applying deep learning models to solve real-world medical tasks. The last part is the summary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

or opportunistic setting, patient go for a regular body checkup or some preliminary examination, with the purpose of determine the malignancy existence. Testing on 331 cases more than half of which have some liver disease, the proposed pipeline achieve 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 technology foundations for medical image analysis. Medical image analysis tasks originate from clinical reality and can be reformulated as computer vision tasks. Computer vision technologies 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 RGB computer vision tasks in recent 10 years,

and medical imaging tasks are also significantly improved. Given enough labeled data, many radiology tasks have been fully solved under some circumstances. 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 affect more than half of the population over 65 years old. Low bone mineral density may cause serious injuries such as fragile fractures for 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 largest 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 data collection, deep learning models, experiments, result analysis to verify the assumptions.

In Chapter 4, we look at the vertebra localization and identification problem. As a major bone structure in the human anatomy, spine vertebra bear the upper body weights and function as important mechanism for body actions. When the patient takes CT scan of the lateral spine for the doctor to examine, the first step is to localize and identify vertebra. 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 vertebra. With the centroid and identity of all vertebra 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 in the

clinical settings, which could help with early detection of liver cancer and also reduce the radiologists' workload substantially. However, there are many challenges due to the 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 task, investigate the challenges, and put forward our solutions. The whole pipeline of automatic liver tumor detection are discussed, including data curation, workflow design, 2-stage reliable prediction, and result analysis.

In Chapter 6, we discuss the medical image analysis development opportunities and challenges. Computer vision and deep learning methodology has revolutionize vision related task, and it is no surprise medical image analysis would play more important roles in the all related clinical aspects. Machine backed automation will surpass human inspections of radiography images in many scenarios, and the challenges mainly lie in data-sharing regulations, diagnostic procedures, clinical integration of computer-aided diagnosis. We should be confident that deep learning technologies 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 backgrounds of medical imaging analysis. We will start from 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 the medical imaging tasks. Thirdly, We overview computer vision techniques and milestones. Lastly, we look into deep learning developments, with a focus on models, datasets related to medical imaging tasks in this thesis. Now let us start from common types of medical images.

2.1.0.1 Medical ultrasonography

Medical ultrasonography creates images of internal body structures to measure characteristics such as distance, velocities, lesion presence. Even though it avoids the use of ionizing radiation, it has many shortcomings. Ultrasonography requires specially trained operator to actively inspect and cooperate with the patient, and it also suffer from

noises or obscurity from bone or air. Ultrasound examination can be used as screening tool for many diseases, such as liver tumors, stomach lesions. For diagnosis purpose, accurate imaging modalities such as CT or MRI are used more.

2.1.0.2 X-ray image

X-ray is the most common diagnostic imaging modality. The X-ray machine sends out electromagnetic waves which passes through the body, and the film receptor on the other side captures the radiation signals. Dense parts such as bone, some lesions would absorb the X-ray thus rendering the corresponding regions darker in the X-ray film. Conditions such as bone fractures, tumor formation can be detected if the anomaly is noticeable, making X-ray scanning a good screening tool.

2.1.0.3 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. CT service become widely available in last 30 years, and about half CT scanning in the US are contrast-enhanced ones which shows the lesions or organs with more clarity. However, a chest or abdominal CT can have as many as 100 times body radiation absorption compared to chest X-ray. Low-dose CT or organ-focused CT may be preferred especially for sensitive population.

2.1.0.4 Magnetic Resonance Imaging

MRI produce 3D voxel scanning of the body, without radiation. Instead, it utilize magnetic fields and a special computer to take high-resolution pictures of body part, which could show bones as well as soft tissues. Having higher resolution and better visual qualities for soft tissues, MRI has advantages for examining subtle changes in brain or soft tissues. The disadvantage is the high cost and the complex operation.

2.1.1 Medical imaging analysis as Computer vision tasks

As medical imaging equipment is widely adopted in the world, radiology becomes a crucial means for disease diagnosis and therapy. X-ray, CT, MRI are increasingly used to screen organs, bones, and soft tissues as devices become more accurate and affordable. The scanning process usually requires professionally trained operators, who know the steps and cautions. Afterward, radiologists with years of experience 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 are deployed in hospitals for screening or diagnosis purposes, it boosts workflow efficiency and lowers medical cost, promoting affordable and high-quality health care.

Medical image analysis can perform better than radiologists 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, 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 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 in a desirable way for osteoporosis screening.

Considerations are needed to formulate medical image analysis as computer vision tasks. Medical image analysis has many similarities and differences with RGB computer vision problems. The task scopes include classification, segmentation, and detection. Many computer vision methodologies also apply to medical image tasks, such as semi-

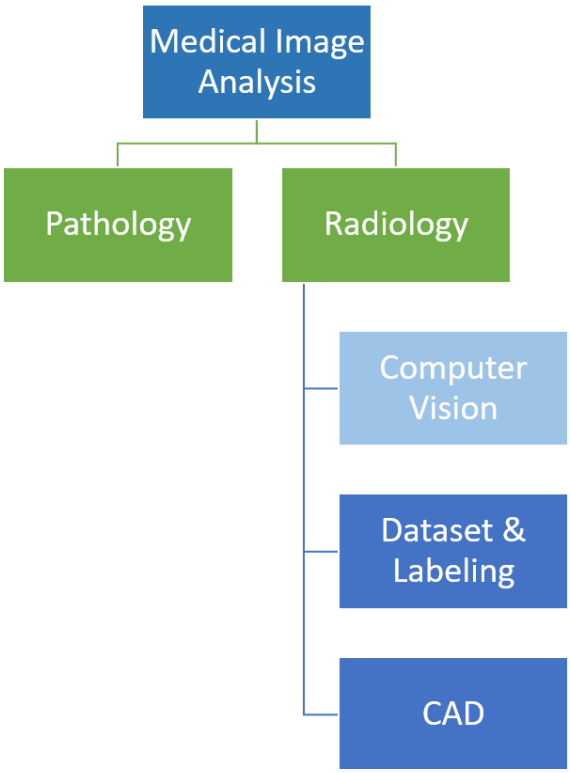

Figure 2.1: Medical Imaging Analysis depends on radiology scanning and may rely on the pathology report. Machine learning application in medical imaging analysis further requires critical factors such as data collection, labeling, model development.

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 an RGB CV task would lie in large and complex relationships in the space. Based on the fundamental differences in patterns and purposes, pixel or voxel segmentation which can capture local nuances is more important and useful for medical image tasks. Segmentation masks not only contain organ statistics but also provides flexible and detailed lesion detection results.

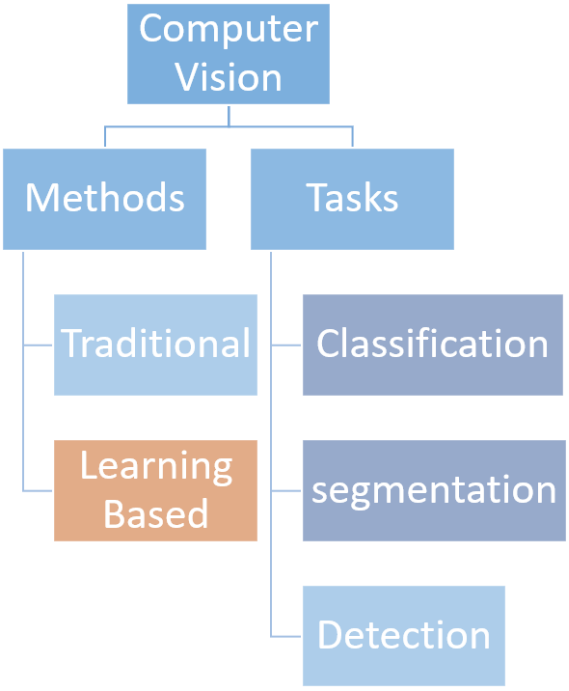

Figure 2.2: Computer vision tasks and methods. The traditional methods contain all the non-learning methodologies, such as hand crafted feature matching, linear regression, SVM, decision tree. The learning based methods upgrade model by adjusting parameters iteratively in an automatic way, until finding out best fitting solution.

2.1.2 The computer vision techniques (Selected)

Computer vision methods mainly consist of traditional ones and deep learning models. The traditional machine learning in CV has developed for tens of years, but is soon replaced by deep learning models in recent 20 years. Traditional pattern recognition methods include hand crafted feature matching, neural network, support vector machines, principle component analysis, decision tree. Depending on the task purposes, these methods may have different usages in image classification, registration, clustering. But they also suffer from many limitations, including small image dimension, simple target pattern, too much hand engineering, little robustness, limited generalization. Deep learning methodology mimic the biologic computation process, and implement solutions with

parametric layered architectures. Relying on convolutional layers, ReLu layers, dropout layer, skip connections, and optimized parameter updating scheme, learning based models are adaptive and resilient.

Deep neural network is inspired by biological neuron functioning in the vision system. In the vision perception system of humans or other mammals, the retina transform the light into neuronal signals, which can be further transmitted into the visual cortex. It still remains a mystery how the brain processes the 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, and these knowledge accumulate in the visual system. Humans have more advanced visual recognition ability than other animals, since humans learns more about logical concepts beyond object appearances. It is also sensible to assume that our brain functions in 'sequential concept layers', and the initial layers are in charge of points, color, distances, lines, corners. The intermediate layers process the shapes, speed, contours. The advanced layers recognize more complex objects, action, and associate 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 task, linear mapping layers could not capture texture patterns due to 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, which captures the visual elements in cascaded manners. When regularization techniques such as dropout

came out, very deep convolutional neural network architectures (AlexNet, VGG, etc.) become possible and popular.

2.1.3 Deep learning models, techniques, datasets (Selected)

In the last 10 years, the deep learning methodology has replaced most traditional machine learning methods in many computer vision tasks. Deep learning models such as AlexNet, VGGNet, ResNet have achieve higher and higher accuracy on the ImageNet challenge. Thanks to the architecture platforms (Caffe, Tensorflow, PyTorch, etc.) and open source movement of deep learning models, an entry-level researcher could train or inference complex CV tasks in a very short time. The flourishing developments of deep learning draws attention from all areas, producing increasingly large and high-quality datasets. Deep learning building blocks are summarized in Figure 2.3.

Regularization layers improve the parameter updating and the gradient calculation process in the model training. Non-linear activation functions such as ReLu and its variatns increase the modeling capacity without easy saturation in Sigmoid or Tanh functions. The dropout layer forces the model to become robust by randomly suppressing some neuron outputs during the forward pass. The weight decay reduces the complexity of a model and prevent overfitting by adding weight regularization term in the loss calculation.

New network architectures help solve fundamental computer vision tasks such as image classification, segmentation, object detection, 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.

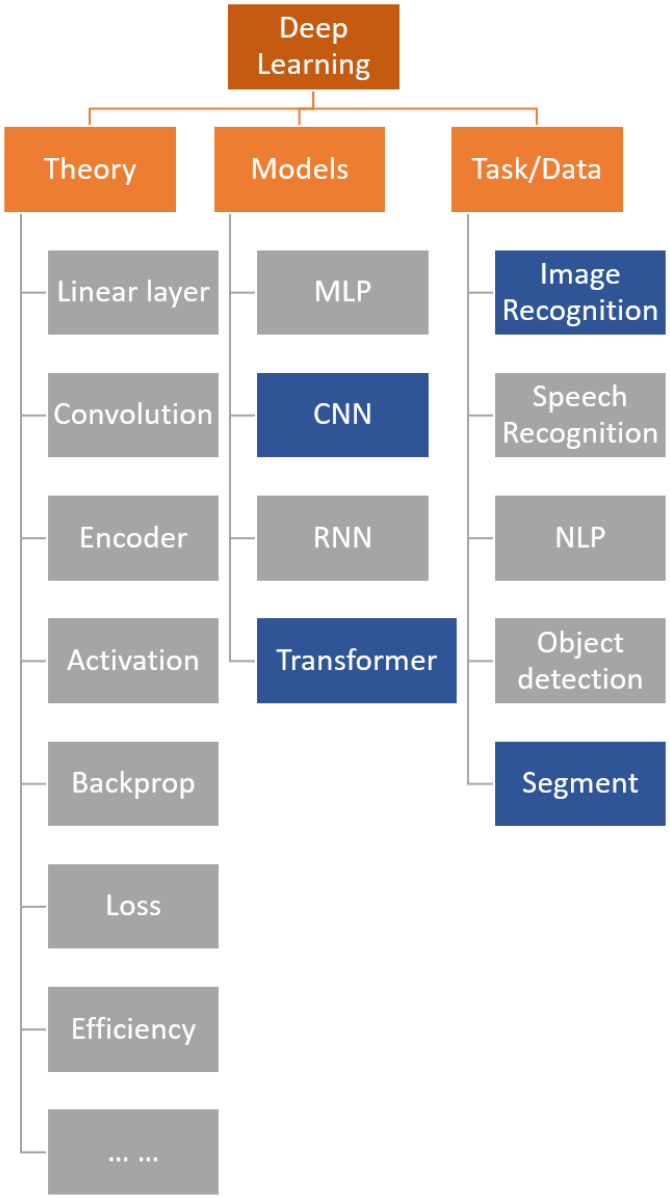

Figure 2.3: Deep learning models are ubiquitously adopted in computer vision, natural language processing, speech recognition tasks. Although tasks differ by scopes, goals, datasets, they share the similar underlying deep learning components and techniques.

After passing though layers of mapping and transformation, the encoder layers learn a good feature embedding. The decoder layers extract the embedded information and combine with corresponding shortcut high-resolution intermediate information of the encoding layers, to generate desired outputs. Encoder-decoder models have been used

in image denoising, object or semantic segmentation, variational auto-encoding.

New visual datasets and benchmarks promote the technology communication and the 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 pretrained on ImageNet contains 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. The transfer learning not only reduces training time, but also substantially improves the performance especially under limited datasets.

2.1.4 The development of computer vision in medical image analysis

Artificial Intelligence (AI) generally includes all man-made machinery, algorithms, methodology that can handle complex tasks requiring memory and reasoning. Human have long sought after automatic and intelligent machines that could accomplish professional tasks to facilitate production and to improve quality. As a key aspect of AI application, Computer Vision (CV) deals with recognizing image patterns, such as classification, detection, segmentation of defined object, activity, and more targets. Computer vision develops along side with AI advancement in recent decades, from initial conceptualization in the 1960s, to established traditional machine learning methodology in the 1990s, and to Deep Learning (DL) era in the 2010s.

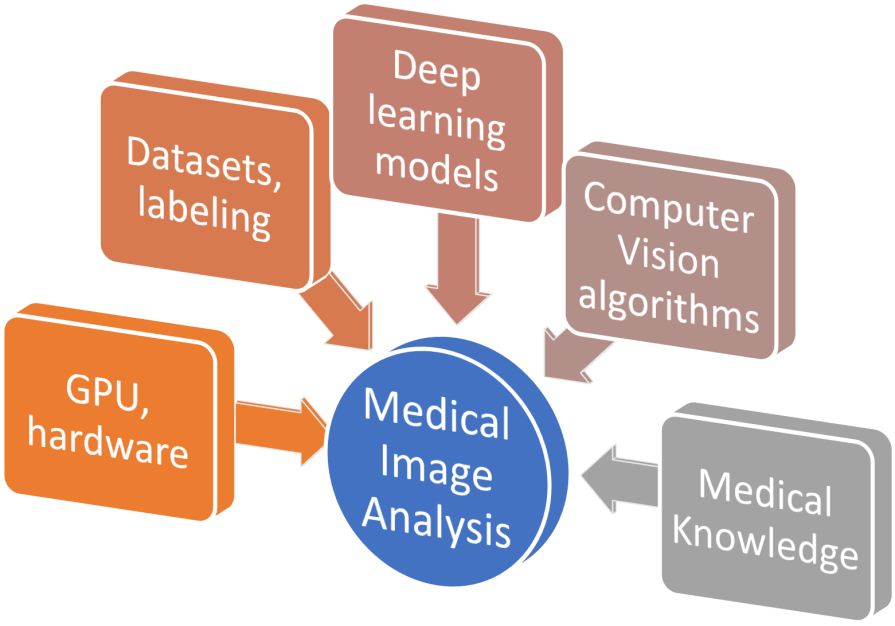

Figure 2.4: Medical image analysis factors. Researches and applications rely on medical understanding of the diseases, hardware, deep learning platform, machine learning algorithms, and models.

Same to the critical factors for any AI applications, computer vision advancements build on high-performance computing, high-quality datasets, improved machine learning algorithms. Before General-purpose computing on graphics processing units (GPGPU) was introduced at the beginning of 21st century, machine learning algorithms run in the CPU only, which usually has 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 camera, 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, pretrained models. What's more, many institutions are willing to share data collection with label, which makes it possible for anyone to build AI models.

2.2 Applying deep learning in medical image analysis

2.2.1 Deep learning methodology for medical image analysis

Medical image analysis play a crucial role in many disease diagnosis. Ultra sound, X-ray, CT, and MRI take the scanning image of some body portion, to monitor abnormal changes inside. Traditionally, medical specialists or radiologists need to spend time and patience to inspect, despite the high labor cost.However, human judgements unavoidably introduce 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, reliability.

Deep learning has been applied in Computed-Aided Diagnosis of many diseases. For example, Chen-I Hsieh *et al.* [4] presents an automated tool to identify fractures, predict BMD, and evaluate fracture risk using plain radiography. Kang *et al.* [5] proposes a semi-supervised self-training algorithm to train a BMD regression model. Hoo-Chang *et al.* [6] investigate the CNN architectures and datasets for medical imaging tasks. Deep

learning models have shown efficacy in lesion detection and classification for various diseases [7]. One example is lung lesion detection, which draws large amount of social attention during the COVID19 outbreak. COVID19 has many symptoms, and one of them shows on 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 large amount of human-labeled CT scans, deep learning models could help doctors identify lung parenchyma changes.

2.2.2 Medical Datasets

In recent years, the medical imaging community has created many datasets and challenges, which significantly speed up the research activities. For example, Xiaosong *et al.* [8] 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.* [9] organizes the MSD challenge, comprised of different targets, modalities and challenging characteristics. The MSD challenge provide datasets for different organs, such as brain, heart, hippocampus, liver, lung, pancreas, prostate, colon, hepatic vessels, spleen. With the availability of high-quality labeled datasets, researchers can propose a variety of models and compare against each other to boost the technology improvement.

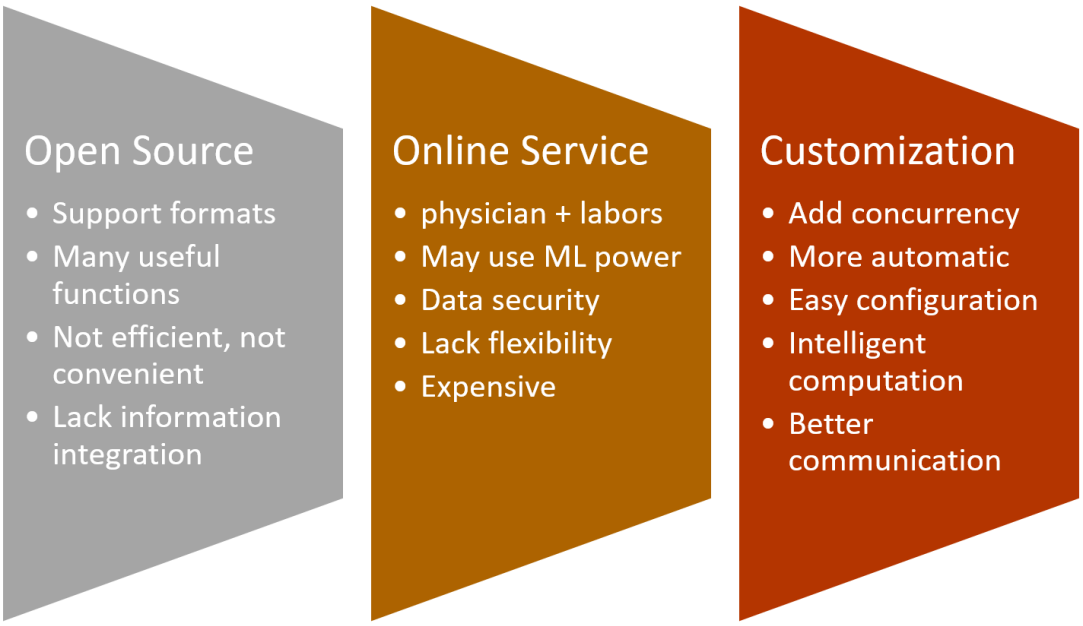

Figure 2.5: There are multiple choices for the labeling tool. We can utilize open source desktop software, which is free and reliable. We can also pay for online labeling service to label the medical images, which has advantages and drawback. When we are working on a large quantity of images, the most effective solution may be customization of all the required functions and features.

2.2.3 Data labeling

One of the centric problem is data labeling in medical image analysis. There can be 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 board. For a particular labeling task in an organ, the committee should define lesion types, characteristics, visual appearances, exceptions. As there can be many rare or exceptional cases in any organ, clear 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 supports main stream medical image formats (DICOM, NIfTI), and have integrated many useful functions (draw, erase, zoom, selection, multiple/3D view, statistics, change contrast). However, open source software lacks many modern features, such as file management, auto loading, auto contrast, label-color configuration, concurrent display for multi-phase CTs, intelligent mask editing.

Besides traditional desktop software, online labeling services for medical image are also popular. Online services (egg: iMerit - Medical Imaging) provide both professional physician guidance and trained labeling labors. The experienced teams would develop labeling guidance and distribute the labeling work to individuals. Sometimes machine auto labeling combined with human correction can effectively boost the efficiency. However, the labeling services also suffer from several drawbacks, such as data security concern, expensive prices, labeling errors.

To ensure high-quality labeling results, medical imaging task may need careful design in all the relevant factors. Due to the differences of targets and medical images, the labeling steps and actions may vary. It is best to customize the labeling tool for best workflow integration. If the it needs communications for review and discussion of labeling, there should be convenient functionalities, which enable comments exchange and navigation. When a patient case has multiple images or masks, the software should load and display simultaneously. The tool should also be able to load medical reports (radiology, pathology) automatically, for improved efficiency.

2.2.4 Limitations

Although it has been proved that deep learning models have outperformed human in many vision benchmark datasets, medical image analysis still has many challenges or limitations. First of all, due to data sharing difficulties and medical procedure standards across the global hospitals, it is hard to develop an 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 depend on many factors. 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 professional grades and so on. The labeling quality may have substantial variations even between doctors, which add up to the difficulty of obtaining a large and uniform dataset. Thirdly, computer-aided protocols and regulations falls far behind the technology development. Human doctors are trained years in medical school and hospitals, and they must get professional certificates before working in diagnosis or therapy. As an emerging role, machine learning models have not be well defined in terms of application scopes, steps, limitations, liabilities. Without adequate discussion and legislation, machine learning models would not be fully trusted or deployed in large scale.

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 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 the 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 [10]. Osteoporosis increases the risk of fragility fractures, which are associated with disability, fatality, reduced life quality, and financial burden to the family and the society. While with an early diagnosis and treatment, osteoporosis can be prevented or managed, osteoporosis is often under-diagnosed and under-treated among the population at risk [11]. More than half of insufficiency fractures occur in individuals who have never been screened for osteoporosis [12]. 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 [13–16]. 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.

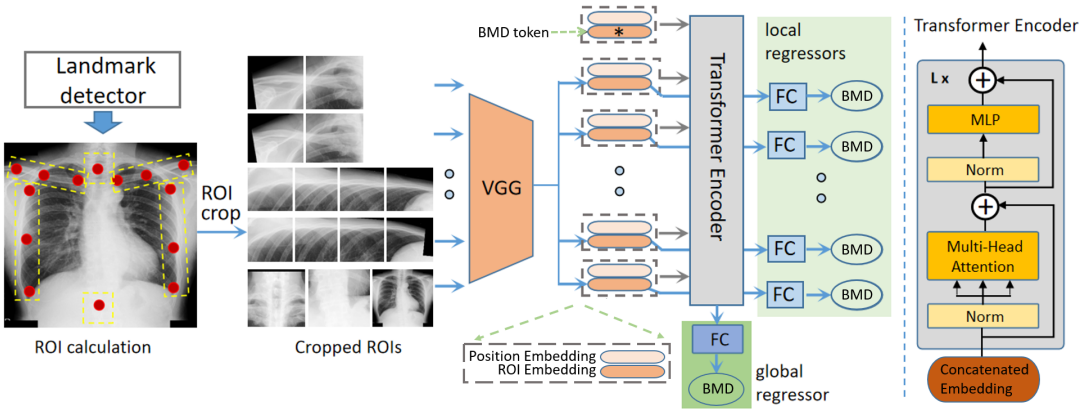

Figure 3.1: The proposed working pipeline. The landmark detector locates key bone points (red dots) on CXR images, then we crop and normalize 14 ROIs. These modalities go through a shared feature extractor (VGG16). The global regressor (dark green) works on the global feature generated by the Transformer Encoder [18] (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 [17]. 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) [19]. *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 the 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 [20] [12]. 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 abdomen or chest can be re-

used without additional radiation exposure or cost [21] [22] [23] [24] [25]. The *Hounsfield Units* of the QCT scans correlate well with DXA BMD scores for low BMD diagnosis [22] [26]. QCT has the advantage of the three-dimensional assessment of the structural and geometric properties of the examined bone [27], but QCT is not widely available. Secondly, the Quantitative Ultra Sound (QUS) based techniques have advantages of safety, low cost, operating flexibility [27] [28] [29]. 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 [28].

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 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 have been verified in osteoporosis screening [30]. In this paper 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 [31] [32] [33] [34], 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 [35] [36] [18]. Some papers exploit textual relationship by enhancing spatial feature encodings [35] or through channel-wise feature recalibration [37]. CCNet [38] proposes the criss-cross attention module to harvest the contextual information on the criss-cross path. LR-Net [39] presents the local relation layer (Local Relation Network) that adaptively determines aggregation weights based on the compositional relationships. GCNet [40] unifies the simplified non-local network [35] and SENet [37] into a general framework for global context modeling.

Inspired by the *Transformer* success in language tasks [41] [42] [43] [44], emerging works employ the *Transformer* modules to replace or facilitate the convolutional layers for visual tasks [18] [45] [46] [47]. The *Transformer Encoder* learns the global relationship through repetitive layers of *Multi-Head Self-attention* and *Multi-Layer Perception* operations. Attention Augmentation [36] augments convolutional operators with a self-attention mechanism by concatenating convolutional feature maps with a set of self-attention feature maps. The iGPT [48] train the transformer model [43] 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 5.1 consists of chest landmark detection, bone ROI cropping, local texture extraction, global feature fusion, BMD regression. We get the key landmarks for the chest from a Graph Convolution Network (GCN) model [49]. 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

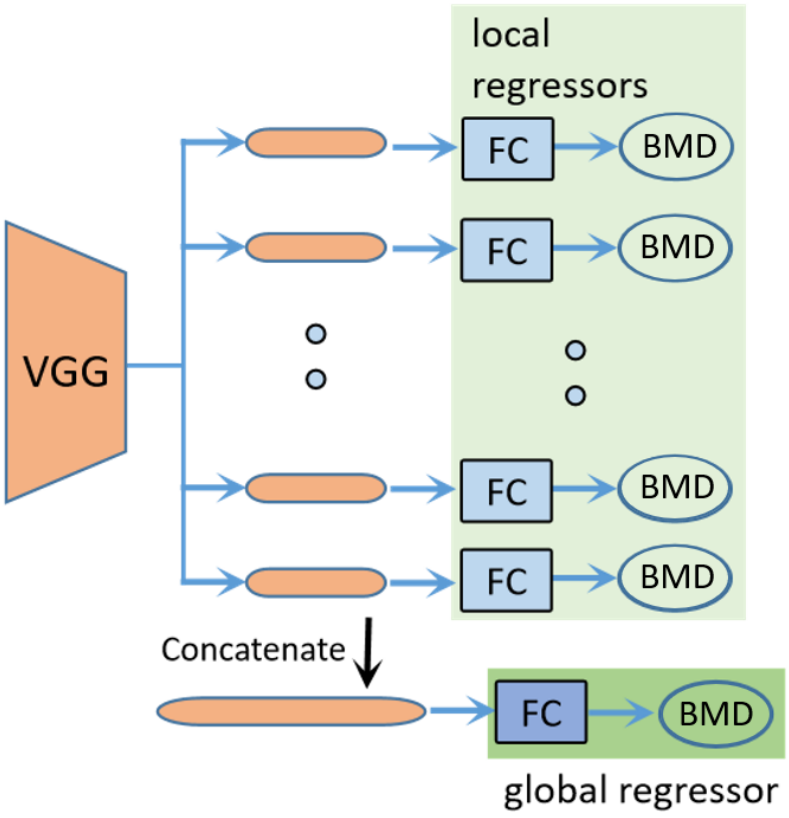

Figure 3.2: The plain fusion process in the Multi-ROI model. Individual feature vectors are concatenated as one before going through the global regressor.

Encoder refines individual representations to enable inter-regional feature interaction. At the end of the pipeline, local regressors and global regressor 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 the metabolic nature, the model should focus on the most effective regions. It is also unclear if the

combination of distinct bone patterns are 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, 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, T12 vertebra.

We utilize the Graph Convolution Network (GCN) based Deep Adaptive Graph (DAG) [49] to automatically detect critical landmarks in the chest. We identify 16 landmarks in Figure 5.1, 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 keypoints. Given the keypoints 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 5.1, there are 2, 2, 4, 4, 1, 1 croppings for left clavicle, right clavicle, left ribcage, right ribcage, cervical, lumbar respectively. This 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.

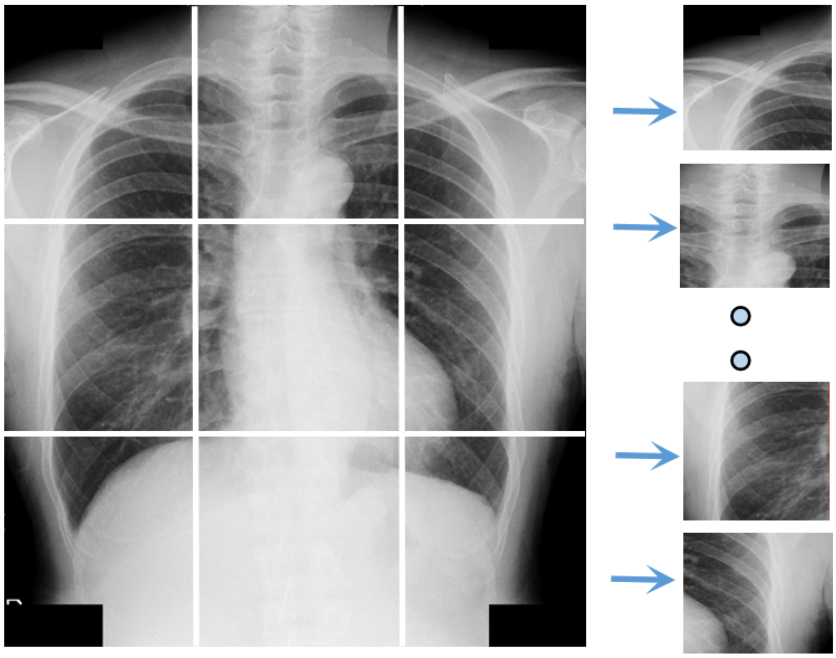

Figure 3.3: Patch generation of Multi-Patch model. The CXR image is split into *3*x*3* non-overlapping regions.

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 [19] [20]. 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* [41], which has been commonly applied in NLP and vision tasks [50] [44] [18] [51], 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 [41]. Similar to *BERT*'s [CLS] token [42], 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 5.1) 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 [18]. 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 [41] [18] 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 modules uses constant latent vector size D=512 through all layers, the same size as VGG16 feature. The Layer Norm (LN) and residual connections are applied before

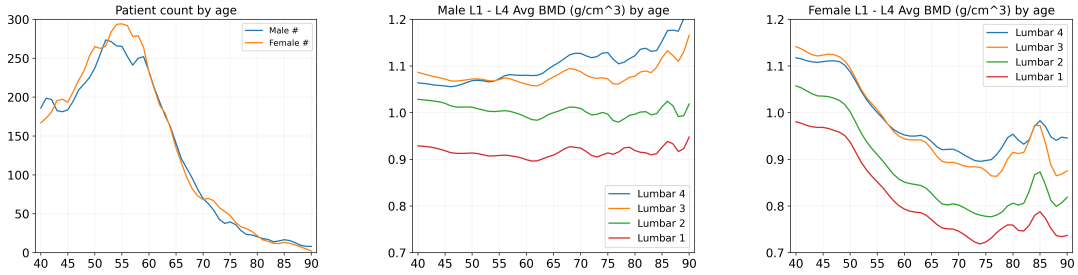

Figure 3.4: DXA BMD averages across ages for both genders. Different lumbar vertebrae (L1, L2, L3, L4) are drawn separately.

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 5.1. 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, 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, 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-val, MAE, SD, AUC represent Pearson correlation coefficient, Mean Absolute Error, Standard Deviation, 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	AUC
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	0.96
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	0.965
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	0.966
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	0.966
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**	**0.968**

3.5 Experiments

3.5.1 Data collection

The data comes from Chang Gung Research Database [52], Chang Gung Memorial Hospital, Taiwan. We follow the Helsinki declaration with ethical permission number IRB-202100564B0 (The correlation between 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 [30].

3.5.2 Experiment Setup

Dataset. We collected 13719 frontal view 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 are conducted in four separate experiments. For a particular lumbar BMD model, it is trained using 4-fold cross-validation, with 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.

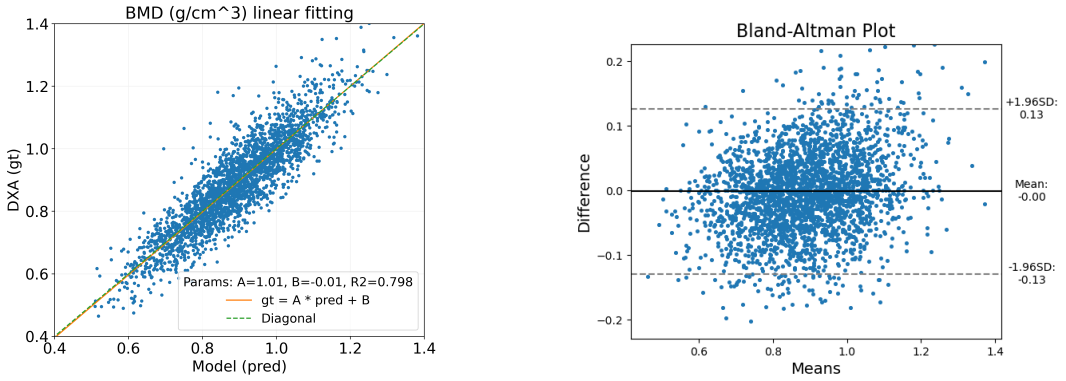

Figure 3.5: Illustrate the proposed model results on Lumbar 1 BMD, each prediction is compared with its paired DXA BMD value. In the linear fitting line (left), **A** and **B** represent **slop** and **y-intercept**, **R2** is the coefficient of determination. In the Bland-Altman plot (right), the horizontal 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 [19]. In Figure 3.4 (1), the data counts

peak 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 the 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 -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, 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 [19] [20]. 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 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 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 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, 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**.

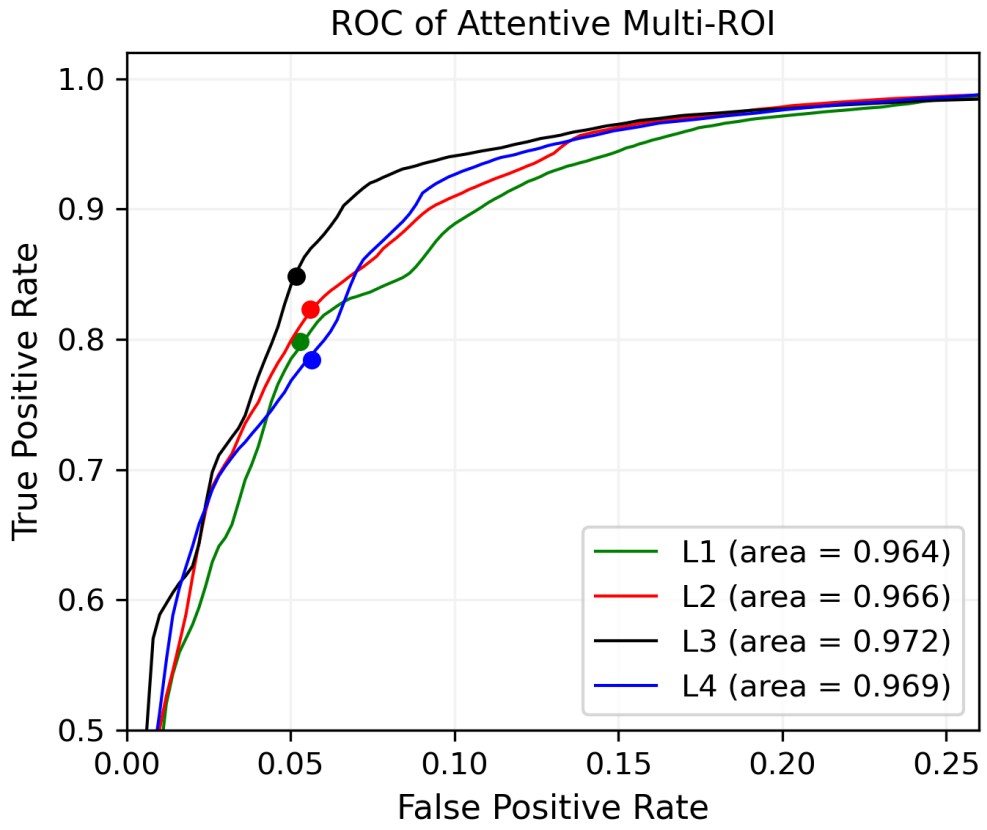

Figure 3.6: The Receiver Operating Characteristic (ROC) Area Under Curve (AUC) of the proposed *Attentive Multi-ROI* model for osteoporosis classification in L1, L2, L3, L4 experiments (only upper left AUC). The colored dots operate on **Flex** T-score thresholds.

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 [30], 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 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 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	Sens	Spec	Sens	Spec
-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

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 [53] [54] 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

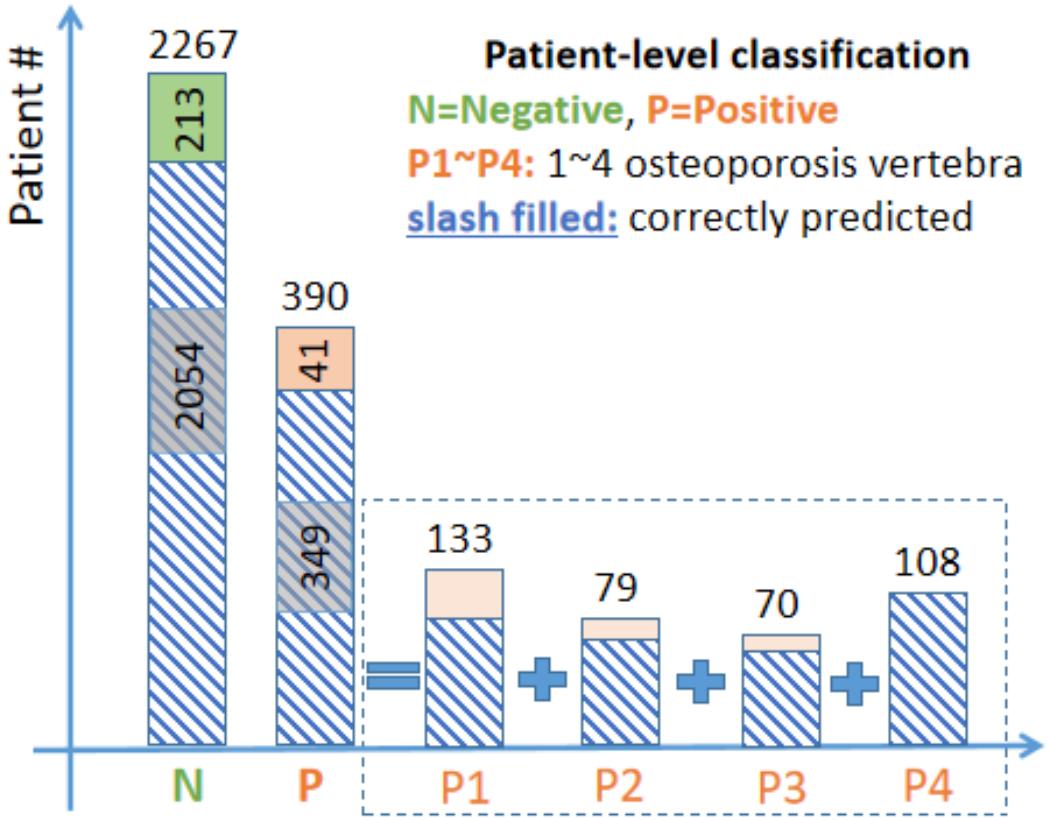

Figure 3.7: Patients with four lumbar records (2657 totally) in the testing set are assigned as normal (Negative) or osteoporosis (Positive). The positive cohort can be further decomposed into four bins, according to the number of osteoporosis vertebrae. The shadowing parts are true negatives and true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

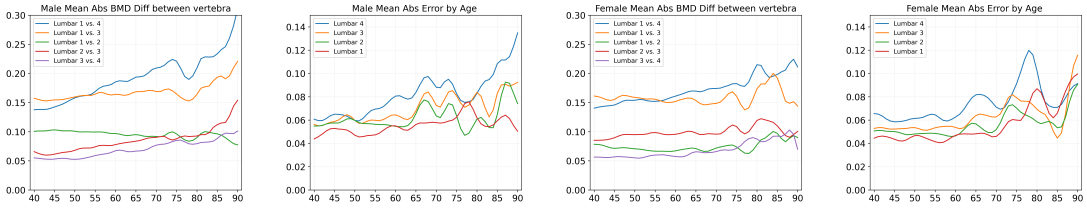

Figure 3.8: Cross-vertebra BMD differences (1)(3). Measure the mean of the absolute difference between vertebra pairs, using DXA BMD (g/cm^3). The mean of absolute prediction error (2)(4). The proposed model (Attentive Multi-ROI) predictions have a satisfactory error range.

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 dimension 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, size, and positions in scanning. As a result, N=3 fits the the best (used as default).

3.6.3 Determine the proper T-score thresholds

Different lumbar vertebrae have distinct BMD distribution 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 the 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%, 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 the 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 [-1.96SD, +1.96SD] range in 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. 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 Standard Deviation (SD) of absolute errors from 0.066 to 0.05, it also has a substantially smaller error variances. 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. Comparing against the proposed model, the Base, the MultiPatch, the AttMultiPatch, 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 on 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 $[0.05, 0.10]$ 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 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,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 [55] [56]. Prior works extracted 3D geometric and structural measurements from area-DXA scans to mitigate this shortcoming [55] [56] [57] [58]. Metrics such as the trabecular bone score (TBS, based on lumbar DXA scanning) are developed to provide bone microarchitecture and skeletal information [59] [60] [61]. These DXA augmentations shine insights for accuracy correction and quality assessment. Though the 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 the cases with implants or bone fractures, which are less frequent in the opportunistic setting. 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 texture 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 [19] [20]. 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 is much greater than DXA, opportunistic screening

using CXR may increase the 'eligible' screening population by multifolds. 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 [20]. This study points out the possibility of a new multi-step osteoporosis screening strategy, incurring no addition 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 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

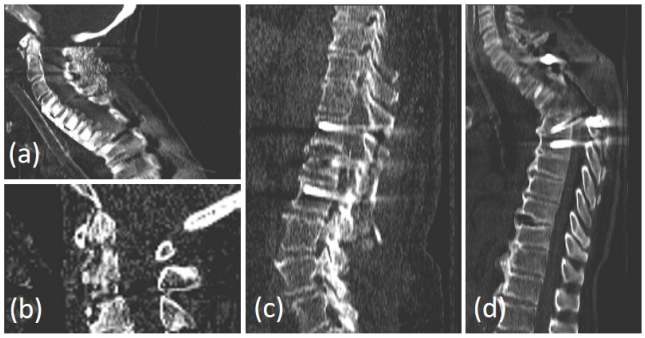

Figure 4.1: Example spine CT images from the SpineWeb benchmark dataset demonstrating the challenges. (a) small field of fiew, (b) low image quality, (c,d) metal implants and severe compression fracture.

being evaluated on a major public benchmark of 302 highly pathological CT images, the proposed method reports the 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 [16]. 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 [62]. 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 [63–65]. 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 [63], or explicitly using an information aggregation layer considering the spatial distribution prior of the vertebrae [64]. The anatomical prior has also been used to post-process the neural network output [65]. While steady performance improvements are observed in these works, the anatomical knowledge is still not fully utilized. In particular, anatomy-inspired network architectures like Bi-RNN [63] rely on the network to learn the anatomical prior without the guaranteed respect to the prior. Building the anatomical knowledge into a network layer [64] or the optimization target [65] 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 [63] and message passing [64]) 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 [63], 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 [64] 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 [65] to jointly label the vertebrae by formulating a global objective function. However, the Markov modeling of vertebra labels employed in [65] 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 the anatomical knowledge. In particular, a key point localization U-Net [66] 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 global message passing 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 [65] 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 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 [67], coordinate-based [68, 69] and graph-based methods [70,71]. Specialized methods focusing on vertebra localization and identification have also been extensively studied to optimize the performance by exploiting the a prior knowledge of the spine anatomy.

Early works rely on hand-crafted low-level image features and/or a priori knowledge. Glocker *et al.* [72] propose to use regression forests and probabilistic graphical models to handle arbitrary field-of-view CT scans. They [73] further transform the sparse centroid annotations into dense probabilistic labels for classifier training. Zhan *et al.* [74] 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.

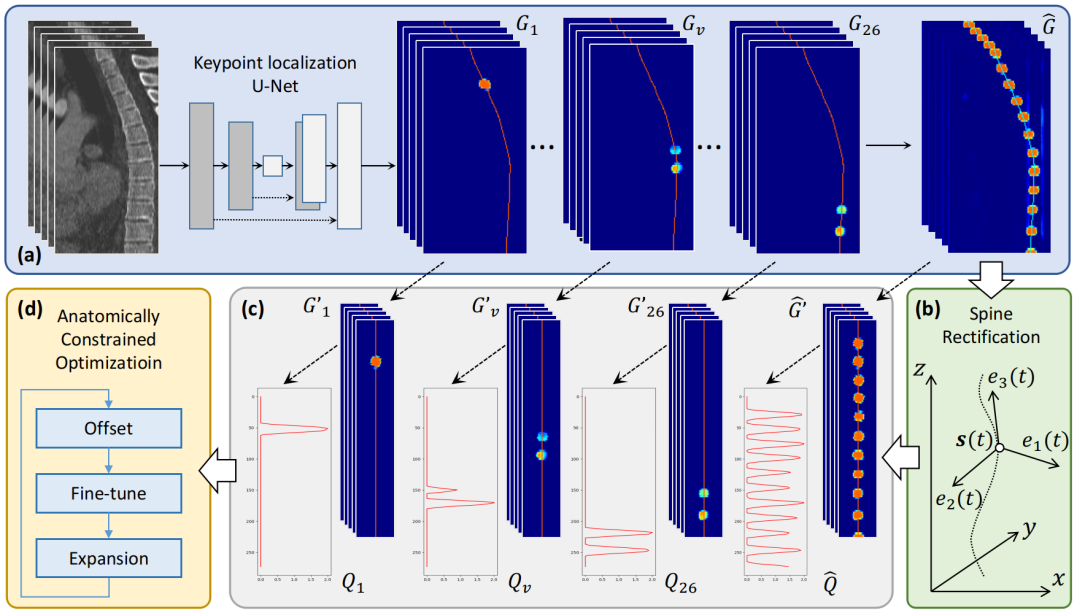

Figure 4.2: Overview of the proposed system. (a) 26 vertebra activation maps $\{G_v\}$ produce by the key point localization nnU-Net, and the all vertebrae activation map $\hat{G}$ produced by aggregating $\{G_v\}$. The centerline of the spine is marked in the activation maps. (b) Spine rectification 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 applied on the 1-D activation signals.

Deep neural networks are employed to detect spine vertebrae and achieve substantially improved performance. A few publications [75, 76] employ convolutional neural network (CNN) to directly detect the vertebra centers. Fully convolutional network (FCN) [77] has also been adopted for the vertebra center detection task [63, 64, 78]. 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 [79, 80] 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 classify them into the respective vertebrae [81, 82].

Many techniques have studied the use of a prior knowledge of spine anatomy to facilitate vertebra localization and identification [63–65, 74, 75, 78, 83]. Domain expert knowledge is used to categorize vertebrae into anchor and bundle sets and treat them differently [74]. Markov modeling is adopted to label vertebrae by preserving the consecutive order [65]. Various attempts have been made to automatically learn the knowledge in a data-driven manner [63, 64, 78, 83]. Bi-directional recurrent neural network (RNN) is adopted to enable the model to capture the spatial relations of predictions in different regions [63, 78]. A message passing mechanism is used to exploit the prior distribution of the distance between vertebrae to regulate the prediction [64]. Adversarial learning has also been employed to encode and impose the anatomical prior [83]. The multi-stage methods [79–82] 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 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 wrong vertebrae, or activate on multiple vertebrae, as shown in Fig. 5.1(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 [75]. Instead of directly processing the activation map channels to obtain vertebra centers, we propose an 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 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. 5.1(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. 5.1(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 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. 5.1(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:

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

$Q_{v_l+i}(k_i)$ 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 [74], 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 weights 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 [84]:

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

The fine-tune 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 [62]. The dataset consists of 302 CT scans with vertebra center annotations. This dataset is commonly considered challenging and representative for 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 [63–65, 72, 75], 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*. Identification rate measures the percentage of vertebrae that are successfully

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
Glocker *et al.* [73]	6.81	10.0	88.8	17.6	22.3	61.8	13.1	12.5	79.9	13.2	17.8
McCouat *et al.* [80]	3.93	5.27	90.6	6.61	7.40	79.8	5.39	8.70	92.0	5.60	7.10
Jakubicek *et al.* [79]	4.21	-	-	5.34	-	-	6.64	-	-	5.08	3.95
Chen *et al.* [75]	5.12	8.22	91.8	11.4	16.5	76.4	8.42	8.62	88.1	8.82	13.0
Sekuboy *et al.* [83]	5.90	5.50	89.9	6.80	5.90	86.2	5.80	6.60	91.4	6.20	4.10
Yang *et al.* [64]	5.60	4.00	92.0	9.20	7.90	81.0	11.0	10.8	83.0	8.60	7.80
Liao *et al.* [63]	4.48	4.56	95.1	7.78	10.2	84.0	5.61	7.68	92.2	6.47	8.56
Qin *et al.* [78]	**2.20**	5.60	90.8	3.40	6.50	86.7	2.90	4.30	89.7	2.90	5.80
Chen *et al.* [65]	2.50	3.66	89.5	2.63	3.25	95.3	**2.19**	1.82	**100**	2.56	3.15
Ours	2.40	**1.18**	**96.8**	**2.35**	**1.28**	**97.8**	3.19	**1.69**	97.2	**2.55**	**1.40**

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 [85] [86]. 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 [72], multi-stage methods [79,80], techniques with data-driven anatomical prior [63, 64, 78, 83] and methods with anatomy inspired architectures [65, 75]. 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.* [65], 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.* [65]. 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.* [65] 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 lumbar spine (*i.e.*, ranked 1st out of 10) but lower id. rate on cervical spine

(*i.e.*, ranked 7th out of 10). This is because their method employs 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 regulate 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 [65] 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 the 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

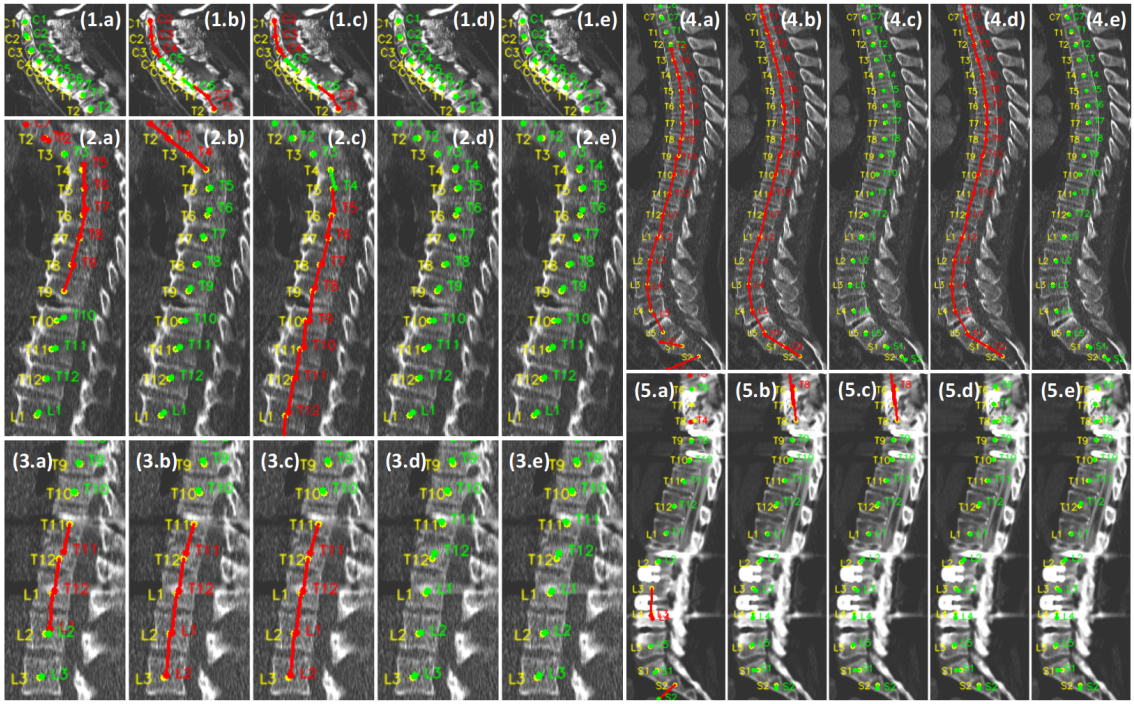

Figure 4.3: Visualization of five sets of final results by five methods. **Dataset (1-5)**: (1) CTs of cervical spine, (2-4) CTs of thoracic and lumbar spine, (5) CTs with metal implant. **Methods (a-e)**: (a) base, (b) base+rect+order w.o. λ, (c) base+rect+order, (d) base+rect+optim w.o. λ, (e) base+rect+optim (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.

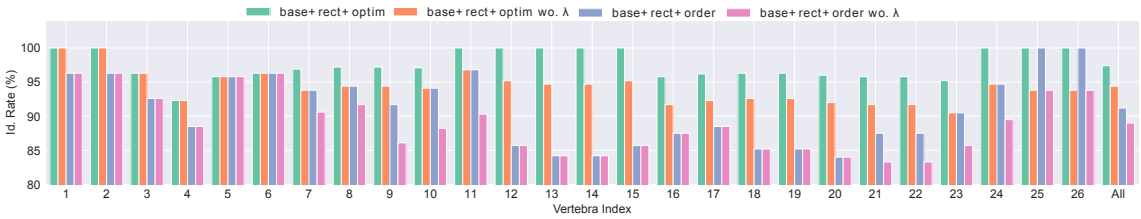

Figure 4.4: Comparisons of base+rect+order and base+rect+optim with and without using vertebrae weights. The identification rates (%) for each vertebra and their averages are reported.

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	I R
base	2.24	1.25	96.8	2.53	1.53	76.0	2.67	1.66	75.2	2.46	1.48	8
base+rect	2.46	1.64	95.8	2.32	1.63	73.8	3.14	1.70	79.3	2.54	1.68	8
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	9
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	9
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	8
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	9

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*. 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 the 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 labels are off by one. The underlying cause of this failure case is the lack of distinct vertebra 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 Conclusion & Discussion

In this paper, 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.

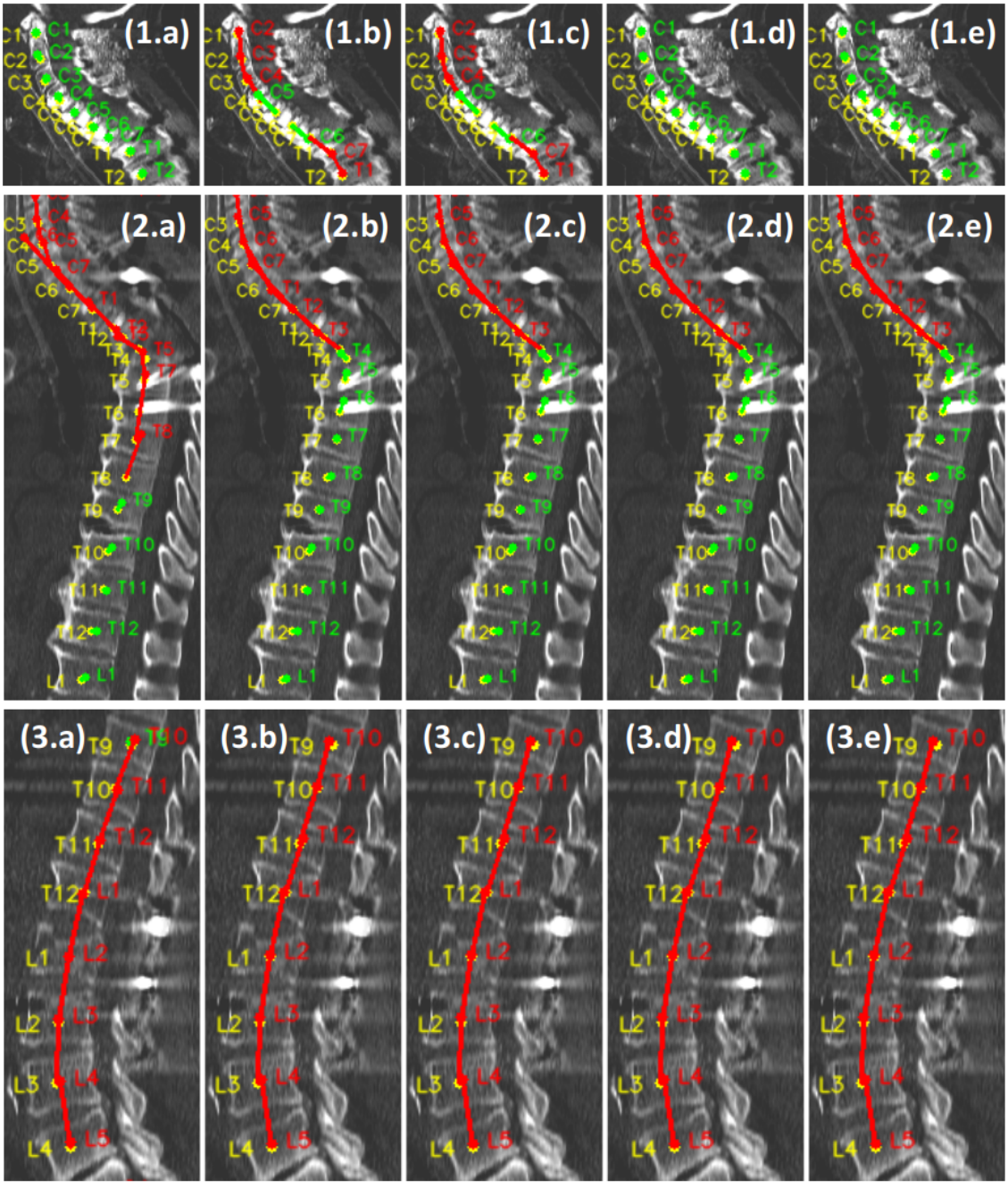

Figure 4.5: Examples of failure cases. The visualization scheme is the same as in Figure 4.3.

By analyzing the failure cases, we observe that severe pathologies and extreme imaging conditions may still negatively impact the model's performance on robustness. Therefore, future research efforts should be conducted to further investigate feasible methodologies to improve the robustness against these corner cases of severe vertebral compression fractures, very low imaging contrasts, strong imaging noises such as metal imaging artifacts, and lack of visually distinct anchor vertebrae.

Chapter 5

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

5.1 Introduction

Liver cancer is the 6th most prevalent cancer by incidence, but the 3rd by mortality in 2020 [87]. Liver cancer has many causes, and it has seen increasing trends due to reasons such as obesity. Liver cancer trends... statistics,... bad influence .... Causes.... The drawback/challenges for current liver cancer pandemic situation: 0) people's awareness, good life styles, etc less alcohol, more healthy food, more exercise, better body shape. 1) liver tumor has few symptoms until mid-late stage, thus missing early diagnosis 2) high cost of CT services/interpretation, causing less body checkups ¡– general reasons for traditional liver related pandemic, diagnosis, result, situation, not mentioning technical stuff.

Deep learning facilitated tumor detection has seen many applications cite papers, brain, lung, colon, liver, ...

Deep learning models can help segment liver tumors automatically for improved

diagnosis and treatment planing. It can also help detect liver cancer early with non-contrast CT images in the opportunistic setting without incurring additional cost. Combined deep learning models for lesion detection in other organs, liver tumor detection models could lower the cost and boost the screening effect.

This work proposes a method to reliably discover liver tumors from CT scanning. Our method is built from 3D UNet segmentation model, with lesion-sensitivity adjustment and context-aware lesion augmentation. We use semi-supervised learning and public liver segmentation datasets, models to boost multi-organ mask labeling of a large private dataset. We design a customized intelligent 3D medical image labeling software CT Labeler to improve the working flow of different roles during mask labeling, thus enabling to label a high-quality, multi-organ, multi-lesion liver dataset. We conduct detailed lesion analysis of the model result, providing the error analysis, case study. We design the context-aware lesion augmentation scheme and lesion model to effectively reduce False Positive detections. We use the multi-sensitivity to reduce the False Negative lesions, whose drawbacks could be managed by the context-aware augmentation and lesion model. Our proposed working flow is fully automatic, it can output precise individual lesion segmentation result. The experiment is based on highly challenging liver dataset, and our model works well on both enhanced setting and the non-contrast setting. In summary, our contributions are three-fold: 1) intelligent way to curate large scale multiple label segmentation 3D medical datasets. 2) context-aware lesion augmentation and lesion model 3) multi-sensitivity model to exhaust lesion discovery

why segmentation over detection? liver is a large organ with mostly uniform textures. The liver overall position and shape do not vary much, and the organ boundary and

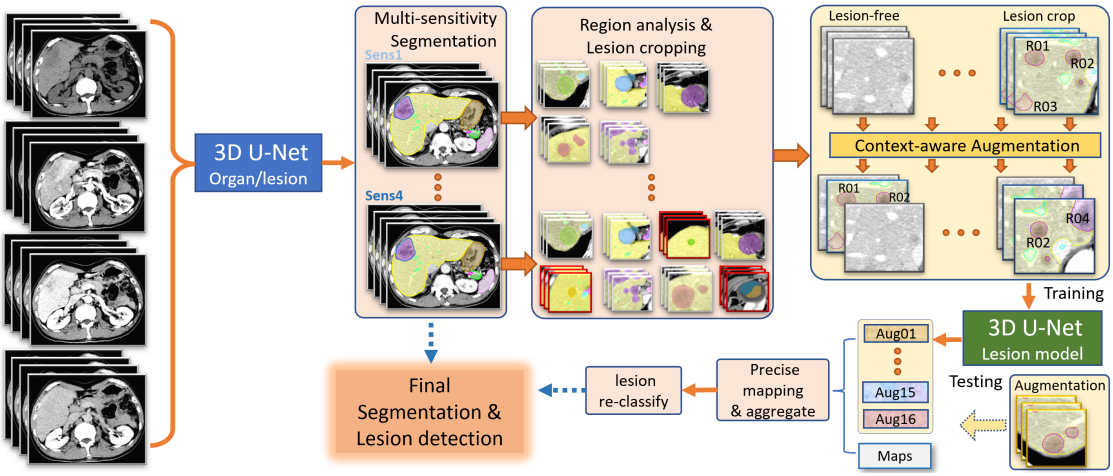

Figure 5.1: The proposed workflow. The CT images (four-phase or single-phase) go through the Organ/lesion U-Net to get the label-specific 3D probability maps, and we adjust the lesion sensitivity to generate Multi-sensitivity segmentation for a balance between False Positives and False Negatives. Medium and small lesions are cropped into patches, randomly combined together with lesion-free patches to form context-aware augmentation. The lesion segmentation model trains and inferences on augmented patches. The aggregated lesion reclassification results are used to update the final segmentation and detection outputs.

inner vessels usually remain clear to delineate. The lesions could appear anywhere in the liver, without a fixed shape, size, or texture. A lesion by nature is a cluster of abnormal tissue that has some distinct morphology characteristics to the surrounding liver. Therefore voxel-wise segmentation could adequately address the challenges and goals for lesion localization and identification. In this paper we employ Fully-Convolutional Neural Network to segment the liver and lesions, then design algorithms and procedures to detect and refine the results.

Known to date, we curate the largest (1631 cases) multi-phase pathology-proven CT datasets with high-quality multi-organ lesion-specific masks [88] [89] [90].

Liver lesions can be born at any parts of the liver which does not has fixed shape or volume. Liver lesions have varied morphology, such as sizes, shapes, compositions,

densities, contexts. There may be no definite distinctions between normal tissues and deteriorated areas from the lesion inception. Therefore radiology based automation targets developed lesions of significant volumes and visual patterns. The voxel-level 3D segmentation is required and sufficient to discover and identify the varied lesions, with all the morphology possibilities.

Our contributions are the following folds. First we curate a large-scale high-quality multi-organ liver lesion datasets with semi-supervised learning and intelligent mask editing. We also develop lesion-level tools for precise analysis and manipulation. Second, develop the context-aware lesion augmentation scheme to effective reduce False Positive detections. Third, we explore the multi-sensitivity segmentation results with quantitative analysis and exploration to obtain high sensitivity and specificity in both the four-phase and NC-phase settings.

5.2 Related work

There are common consensus principles for the diagnosis of liver tumors via CT scanning. Although ultrasound can be used for liver cancer screening, CT and MRI improve detection of Hepatocellular Carcinoma (HCC), compared With ultrasound alone in patients with cirrhosis [91]. As means for noninvasive diagnosis of hepatocellular carcinoma (HCC) in patients with cirrhosis, both multiphasic CT and MRI are capable and have unique diagnostic benefits. Consensus guidelines for radiological diagnosis of HCC have been drafted by several large international working groups [92]. Dynamic and multiphase contrast-enhanced computed tomography and magnetic resonance imaging

still form the cornerstone in the diagnosis of HCC [93]. The EASL Clinical Practice Guidelines [94] gives advice for the clinical management of patients with HCC, as well as an in-depth review of the relevant supporting data. The Liver Imaging Reporting and Data System (LI-RADS), supported by the American College of Radiology (ACR), provides standardization for hepatocellular carcinoma (HCC) imaging in the contexts of screening and surveillance, diagnosis, and treatment response assessment. LI-RADS was developed by a multinational consortium of radiologists and other specialists with expertise in liver cancer imaging, and it was integrated into the most recent HCC clinical practice guidance by the American Association for the Study of Liver Diseases (AASLD) [95].

The 2-stage paradigm (coarse to fine) has been applied in many CT-based lesion detection tasks [96] [97] [98]. The chest or abdominal CT images contain different body regions, with different visual properties. As organs in the body pack together tightly, organ boundaries are not clearly determined by machine due to the flexible morphology and texture similarity. Lesions can vary a lot by the organ origin, stages, sizes, and other characteristics. The disproportionate volumes among organs and lesions, combining with the lesion uncertainty in the data renders segmentation or detection of the disease challenging. A typical 2-stage pipeline [99] would localize the target organ in the first stage, where the result can be used to focus or crop. In the second stage, a dedicated model is trained on the specific organ with lesion labels for the finer level results. This coarse-to-fine approach helps reduce lesion false positives and make it easier to mitigate the imbalance among lesion classes.

Although there have been several public CT datasets that can be used for liver tumor detection in recent years [100] [90], many shortcomings and challenges exist. The

Medical Segmentation Decathlon (MSD) [100] contains CT datasets for multiple organs (pancreas, lung, prostate, liver, colon) from different medical centers. The AbdomenCT-1K [90] is an abdominal CT organ segmentation dataset, with more than 1000 (1K) CT scans from 12 medical centers, including multi-phase, multi-vendor, and multi-disease cases. But AbdomenCT-1K borrows data from existing ones such as MSD and LiTS.

There are several ways to deal with lesion detection errors. There is a contradicting balance between the model performance of False Negative (when the segmentation or detection model miss lesions) and False Positive (when the model miss lesions). The HPVD model [101] accepts any combination of contrast-phase inputs with adjustable sensitivity depending on the clinical purpose.

Semi-supervised learning can be used to mitigate the data heterogeneity and labeling scarcity problems [102] [103]. Typically, we can build a lesion detection model on available public datasets, and use the model to automatically produce masks on unlabeled data. Human inspection and correction are needed with varied intensity as follow-ups.

The segmentation-based deep learning workflow is widely adopted for disease diagnosis and lesion detection in CT scans. The 3D UNet [104] extends the previous u-net architecture [105] by replacing all 2D operations with their 3D counterparts, which can performs on-the-fly elastic deformations for efficient data augmentation during training. The nnUNet [106], achieving high performance on CT segmentation tasks, automatically configures the deep learning workflow, including preprocessing, network architecture, training and post-processing for biomedical segmentation task.

5.3 Methodology

5.3.1 The Baseline model

The liver tumor segmentation/detection task via CT images can be well addressed by the 3D Unet architecture. We adopt the nnUNet framework as the baseline, with 5-fold cross-validation training scheme. Lesions and organs are treated the same during segmentation generation from the output probability map. Individual lesion detections are extracted through the region computation algorithm.

5.3.2 Fundamental algorithms for lesion computation

The proposed model relies on lesion sensitivity adjustment and precise lesion manipulation which requires fundamental algorithms such as region computation and lesion matching. The region computation (mask voxel clustering) is the basis for all analysis and procedures. The lesion matching results are not only used for statistics but also for lesion cropping and augmentation.

5.3.2.1 Region Computation

For multiple organs and lesions segmentation, we want to generate the connected voxel clusters (regions). Firstly the lesion regions transfer the segmentation result into precise detections. Secondly, it enables the lesion filtering/cropping and detailed analysis. The GenRegion 2 algorithm takes in a 3D mask M and produces regions each of which features a unique *region id (rid)*, the *clustering mask*, the *volume*, and the *density*. A

unified 3D *region mask* combines the *clustering masks* and marks each *region* with the corresponding *rid*.

5.3.2.2 Lesion Matching

Lesion-level analysis in a single mask requires localizing and isolating the individual lesions and the *GenRegion* algorithm could handle. However, it becomes more complex for comparing the predicted mask (*Pred*) with the ground truth mask (*Gt*) in several special scenarios. Some lesions may only appear in the *Gt* (FN) or the *Pred* (FP). Sometimes one large lesion in the mask may correspond to several smaller lesions in the other mask, commonly seen in metastasis cases. It is normal that the lesion mask partially overlaps in the *Gt* and the *Pred*. The lesion matching algorithm would influence the lesion cropping and performance statistics. Therefore we standardize the steps and operations of the *Mask Matching* 3 algorithm in all experiments.

$\mathbf{TP}$, $\mathbf{FL}$, $\mathbf{FP}$, $\mathbf{FN}$ represent *True Positives*, *False Labels*, *False Positives*, *False Negatives* respectively in the lesion-level matching process. The *Union* function unifies one or multiple masks by merging all lesion labels (*Lesion_Ls*) as one label. During the process the neighboring fragments of lesion predictions are re-joined together. The *Restore* function re-assign an appropriate label to each unified lesion by the label ratio sum of the original fragment volumes. The *VolumeFilter* function looks for noisy lesions ($< 0.1\ cm^3$) and converts them to the liver label. The *GenInfor* function examines each unified regions and computes the sub-mask constitutions (label & ratio pairs). Therefore each *region* in the $\mathbf{U\_infor}$ contains the label-volume pair lists for the *Pred* and the *Gt*

separately. The *Aggregate* function sets the primary label by volume constitution for the region, which enables the lesion-lesion matching.

5.3.3 Multi-sensitivity segmentation

The baseline model treats all organs and lesions with the same importance in the segmentation computation from the $\mathbf{Prob} \in \mathbb{R}^{(L+1)*Z*Y*X}$, which brings in noticeable *FN* detections due to the relatively weak lesion signals. Multi-sensitivity models consciously adjust the lesion sensitivity to reduce *FNs* while maintaining an acceptable level of *FPs*. To extract the liver lesions as complete as possible, we increase the lesion weight by multiplying the corresponding probability maps with scaling factor $f \in \mathbb{R}$ in Equation 5.1, where *Lesion_Ls* is the lesion label set. In Equation 5.2, $\mathbf{Seg}_f$ is the segmentation applying scaling factor *f*, and we use $Sens_f$ to represent the corresponding model adopting $\mathbf{Seg}_f$.

$$\mathbf{Prob}[i,:,:,:] \leftarrow \mathbf{Prob}[i,:,:,:] \times f\,, \quad i \in \textit{Lesion\_Ls}. \tag{5.1}$$

$$\mathbf{Seg}_f \leftarrow \textit{argmax}(\mathbf{Prob}). \tag{5.2}$$

Draw the relationship between scaling factor and lesion_Volume, lesion_number, FN, FP. Quantitatively speaking. for l in lesionLabels: Prob[l,:,:,:] = Prob[l,:,:,:] * Factor, the sensitivity Factor $\in [1, F_{max}], F_A, F_B, F_C, F_D$. The larger sensitivity factor, the larger lesion volumes and detections, then the more False Positives and the less False Negatives. For the ease of description, we measure medium and small lesion changes applying different sensitivity factors in this section. The Factor=1, the FP number and FN number of small and medium lesions are $Base_{FP}$ and $Base_{FN}$ respectively.

$Num_{FP}(factor) = \alpha(factor) * Base_{FP}, Num_{FN}(factor) = \beta(factor) * Base_{FN}.$

In the train-val set (cross validation inference), we records the lesion detection relationship with the sensitivity factors. The $\alpha(factor)$ and

$\beta(factor)$ has the changing trends described in Figure .... which can also be described by equation

$\alpha(factor) = \alpha * 1/3 + 2/3$, $Num_{FN}(factor) = Base_{FN} * 1/factor$ $\beta(factor) = 1/factor$. Similar trend are observed in the testing set. When the sensitivity factor = 1 (F_A), we have a small amount of FPs while a large amount of FNs.

To achieve the benefits of both low FP and low FN, we combine the results of the $Sens_1$ and the $Sens_4$. To mitigate the large FPs especially in the $Sens_4$, we introduce the lesion reclassification scheme.

5.3.4 Context-aware Lesion Augmentation

As the largest organ in the upper body, the liver texture has many variations due to the complexity of related diseases. The combination of local tissue deterioration, vessel anomaly, duct embolization could lead to ambiguity appearances for lesion detection. The holistic texture changes would appear in many liver sections. If we could tell the texture deterioration from lesions, we could effectively reduce the *FP* detections. So we propose a dedicated lesion segmentation model which applies the context-aware lesion augmentation. The lesion model relies on the segmentation and cropping results from the organ/lesion model, and it focuses on the context-aware lesion patches. During inference, the segmenting results of lesion patches are precisely mapped and collected for lesion

re-classification.

In the AugLesion (Algorithm 4), $\mathbf{U\_infor}$ is the lesion matching result (or the region computation result of the predicted mask during inference) containing lesion regions and the related information. $\mathbf{Img}$ is either the NC-phase CT image or the four-phase CT images. $\mathbf{Gt}$ is the ground truth mask only for training. For each lesion region, we randomly crop an area $\in \mathbb{R}^{ZC*YC*XC}$ in any directions while ensuring the lesion centroid is included. The same cropping profile is applied to both the CT image and mask. Negative croppings are obtained by selecting liver areas without lesion masks. The lesion cropping and the negative cropping are randomly concatenated to boost the augmentation effects. The augmented lesion patches and the corresponding mask patches are stored in the $\mathbf{CombImgs}$ and the $\mathbf{CombMasks}$ respectively. *AugRepeat* specifies the strength of the augmentation.

The high sensitivity segmentation result from the organ/lesion model contains many *FPs*, due to the liver texture variations and dubious liver tissues, such as fatty liver, cirrhosis, duct inflammation etc. The lesion patch model re-classify the context-aware augmentations with a emphasis on differentiating lesions from liver tissues. By randomly combining lesion with lesion-free liver context in the augmented patches, the model learns to ignore dubious liver texture deterioration. Unlike the first organ/lesion model which take in the whole CT scanning, the second model only work on liver and liver lesions to reduce mis-classification between liver and lesion.

5.3.5 Joint prediction of Multi-sensitivity models with Context-aware augmentation

Low sensitivity model ($Sens_1$ w. *Patch*) can output reliable lesions but has many *FN*s, while high sensitivity model ($Sens_4$ w. *Patch*) can effectively reduce *FN* lesions but has more *FP*s. In the proposed model ($Sens_1$ w. *Patch* + $Sens_4$ w. *Patch*), we harness the both benefits by dividing the predictions into the consensus group and the uncertain group. This paradigm works out for several reasons. Firstly these two models agrees on the patient-level predictions for most cases, only differing on the limited *FP* or *FN* cases. Secondly, when the disagreement appears, the ambiguity is highly likely due to some real or dubious lesion, which suggests further examination. Thirdly, the separation of different sensitive results imitate the real world diagnosis variations among radiologists, suitable for varying purposes.

$$\textit{MainLesion} \leftarrow \textit{CalcPatientLevel}(\mathbf{Regions}, priority) \tag{5.3}$$

In Equation 5.3, the patient-level calculating function takes in the predicted lesion regions and lesion priorities to determine the main lesion for a patient. Considering the malignancy and clinical needs, we set the lesion priority order as $HCC > Cholangio > Metastasis > Hemangioma > Other > Cyst > Normal$. The majority of patient-level predictions are the same from the $Sens_1$ w. *Patch* model and the $Sens_4$ w. *Patch* model (Consensus ratios are 97% and 93% in the four-phase and NC-phase settings respectively).

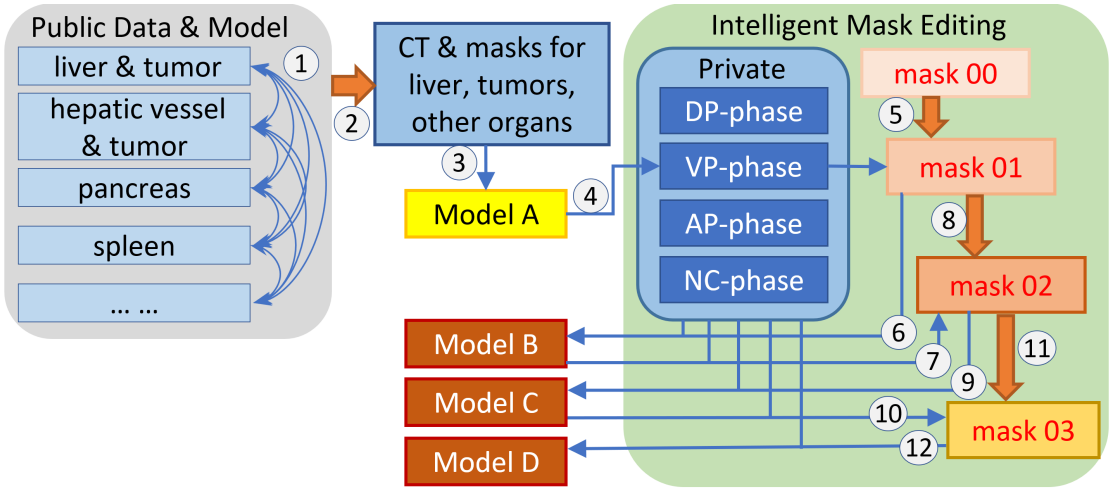

Figure 5.2: The data curation process: datasets, labels, models, relationship. The mask inspection and editing are efficiently done with the *CT Labeler*. The 12 steps are elaborated in the paper.

5.3.6 Intelligent Multi-phase CT Dataset Curation

5.3.6.1 Intelligent mask editing tool: the *CT Labeler*

As the liver tumor segmentation task involves multiple phases of CT images, rounds of mask updating, varied tumor characteristics, existing tools such as ITK-snap could not handle these challenges well. We design the *CT Labeler* tool which has features solving all the labeling challenges and improves the productivity greatly. The software manages the patient cases and files smoothly and could automatically load and save images (CTs, masks) upon switching cases, masks. It could show four CT phases concurrently, load multiple masks concurrently with independent showing styles (mask transparency, contour thickness, contour transparency), adjust CT contrast flexibly, navigate easily with either mouse or keyboard (zoom, shift), draw and erase masks efficiently, automatically load and show patient reports, compute and list regions (voxel clusters in masks), conduct region

editing (delete, re-label, copy across masks), enable lesion-specific comments. Being used by all personnel of our project, the *CT Labeler* has shown great efficacy in case inspection, mask editing, doctor-engineer collaboration.

5.3.6.2 Data curation process

The data curation process in Figure 5.2 has 12 steps. The 1st step is to collect public datasets related to liver tumor detection and to conduct cross inference. We select the following organs/lesions for the public data: pancreas, spleen, liver, portal vein, hepatic vein, gallbladder, stomach, liver tumors. The 2nd step is to organize the CT images and inferenced masks, making a data split. The 3rd step is to train the 3D UNet *Model A* based on the curated public dataset. The 4th step is to inference the VP-phase CT of the private data with the *Model A*, to get *mask 01*. The 5th step is to update the *mask 01* with *mask 00* which are labeled by radiologists with pathology reports. The tumor lesion is divided into 6 specific lesions, namely HCC, Cholangio, Metastasis, Hemangioma, Other (including FNH, TACE equip, dead lesion, other benign tumor), Cyst. Obvious mask errors are also corrected. The 6th step is to train a 3D UNet model with four-phase CTs and improved *mask 01* to get *Model B*. The 7th step is to inference the four-phase CTs with *Model B* to get *mask 02*. The 8th step is to update the *mask 02* by comparing the lesion regions with ones generated from *mask 00*, where the differences are sent to liver experts to check against radiology and pathology reports. Organ mask errors and obvious lesion errors (cyst, TACE equipment) are corrected. The 9th step is to train a 3D UNet model with four-phase CTs and the updated *mask 02* to get *Model C*. The 10th, 11th, 12th

steps repeat the process of the 7th, 8th, 9th steps to get *Model D*. More details are in the experiment section.

5.4 Experiments

5.4.1 Data Curation

5.4.1.1 CT image collection

The private data comes from Chang Gung Research Database [52] (CGMH liver center between 2008 and 2018), Chang Gung Memorial Hospital, Taiwan. We follow the Helsinki declaration with ethical permission number IRB-201800187B0 (Liver tumor detection through CT images). There are 4000 pathology-verified liver cancer cases in total with four-phase CT images, radiology reports, pathology reports, counting in both the cancerous (about 2/3) and the non-cancerous (about 1/3) samples. The malignant (HCC, Cholangio, partial Metastasis) lesions have been marked with bounding boxes by experts checking the radiology and pathology reports. The non-cancerous cases do not contain malignant tumors, but more than half do have some liver lesions. After filtering out unsuitable cases (exam after operation, CTAP, scanning failure due to metal implants, uncertain lesions) at the quality-assurance stage, 1631 cases remain.

The CT images are in NIFTI format where patient information has been removed to protect privacy. The CT machine is ???. The CT xxx is xxx, ... The voxel spacing is between ... and ... . . The four-phase CT images are aligned by xxx algorithm, with small alignment error in most cases.

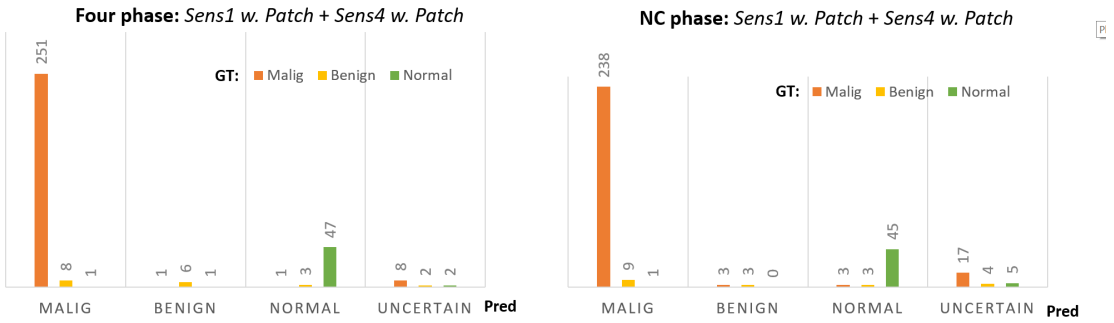

Figure 5.3: The patient-level classification on test set (331 cases) of the proposed model. A small portion is labeled as *uncertain*, while the rest has high performances for malignancy prediction. The majority of the *uncertain* cases are actually abnormal.

5.4.1.2 Label assignment

There are several stages to curate high-quality liver data set for the experiments, such as CT phase alignment, lesion labeling, organ labeling, mask updating. During the multi-step curation process, we pay especial attention to several detailed aspects such as lesion labeling scopes, relevant organ contexts, handling mask updates. The four-phase CT alignment algorithm ensures tiny alignment errors in most patient cases. We determine the lesion types for the data set through the discussion with collaborating liver experts on relevant publications and the data characteristics. We follow the principles of 1) incorporating the malignant tumors, 2) including benign tumor-like lesions, 3) grouping less-frequent lesion types together. The final labeling types (HCC, Cholangio, Metastasis, Hemangioma, Cyst, Other) consider both the lesion appearance and distribution characteristics to enable the smooth training of deep learning models. To maximize the segmentation model learning capability, we incorporate the surrounding organs as training goals as well which could reduce error predictions on organs and bring in more stability. The selected organs (hepatic vein, portal vein, gallbladder, spleen, pancreas, stomach) are

either directly connected or closely bordering to the liver. The labeling selection helps localize and identify both the liver and tumors more accurately.

5.4.1.3 The labeling process

To accomplish the multi-target labeling on a large scale multi-phase CT image dataset, we design a customized CT masking tool, the *CT Labeler* with the adequate functionality and usability. The *CT Labeler* features case file management, user configuration auto-loading, concurrent four-phase CT display, style-specific multi-mask showing, 3D connected region computation, convenient mask editing (drawing, erasing, copying), lesion-specific commenting, and medical reports auto-loading. As shown in Figure 5.2, the *CT Labeler* empowers the intelligent multi-step mask curation process with automatic updating and efficient lesion-specific inspection. By automatic mask computation and comparison between existing and new predictions, the software could localize the differences effectively and involve precise human interaction efficiently. By following the determined labeling guideline and editing protocol, the machine learning engineers and liver experts could improve the mask quality in iteration. We divide the data into testing set and 5-fold train-validation sets. The models (B,C,D) are trained using 5-fold cross-validation paradigm, and the train-val sets are inferenced correspondingly. The test samples are inferenced by model ensemble. In the end we get 1631 well-labeled cases (1300 in the train-val set and 331 in the test set).

5.4.1.4 Lesion statistics

Our data is from a regional liver medical center whose population is limited to East Asian Han Chinese and the majority records have liver-related diseases. Based on lesion distribution and appearance characteristics we assign 6 labels: HCC, Cholangio, Metastasis, Hemangioma, Cyst, Other (FNH, TACE equipment, deceased tumor, other benign tumors). To better serve the liver disease diagnosis and screening purposes, we assign the priorities from high to low as HCC, Cholangio, Metastasis, Hemangioma, Other, Cyst, Normal at both the lesion level and the patient level. If multiple lesion types present, the patient-level main lesion is defined by the highest priority (most severe) one. When multiple malignant lesions occur, it is defined as largest volume type. The lesion has varied sizes/volumes, and the majority non-HCC lesions are smaller than 16 cm^3. The 331 testing cases comprises of 216, 12, 33, 6, 11, 2, 51 cases for HCC, Cholangio, Metastasis, Hemangioma, Other, Cyst, Normal respectively (patient-level) in Table 5.2

5.4.2 Performance Metrics

As our work flow derives lesion detections from 3D segmentation in the Figure 5.1, the performance metrics consider both lesion-level and patient-level matching results. The CT mask contains all the voxel labels, while a lesion or an organ is a voxel cluster of the same label. 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 in *Pred* or *Gt*. We also gather the lesion-level detection statistics, including 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 lesion-grouping manners, which are lesion-specific, All-lesion (*All*), 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 on lesion overlapping, ignoring lesion types.

At the patient-level, the metrics are based on main lesion matching results. Each prediction or ground truth mask has exactly one main lesion tag under 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, Cholangio, Metastasis$\}$, Level2 = $\{$Hemangioma, Other$\}$, Level1 = $\{$Cyst$\}$, Level0 = $\{$Normal$\}$) to reflect disease severity, and the Level Accuracy (*L.Accu*) measures the level-specific correctness. Alternatively we can treat malignant types (HCC, Cholangio, 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) generate final predictions from the joint consensus of two context-aware segmentation at different lesion-sensitivity levels. The high sensitivity model discover as many potential lesions as possible leading to high *FP* and low *FN*, while the low sensitivity model selects high-confidence lesions

resulting in low *FP* and high *FN*. In their joint prediction at the patient-level, the majority common outcomes are more reliable while the differences (Uncertain) usually result from risky cases indicating further examination. 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 have malignant lesions in the result. In Table 5.1, the proposed (Sens1 w. Patch + Sens4 w. Patch) achieve 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 3D UNet

We adopt the nnUNet framework to generate 3D segmentation in all experiments. The output $Prob \in \mathbb{R}^{14*Z*Y*X}$ contains individual probability maps for all 13 labels and the background, and $argmax$ operation produces the segmentation mask. We use the result for the Baseline, and run the region computation and lesion matching algorithms 23 to get the precise lesions and to calculate the patient-level metrics in Table 5.1. In the four-phase setting, it reaches 85.2% lesion-specific accuracy and 92.4% level-specific accuracy. Compared with previous published papers ....

compared to existing works: better quality labeling; larger amount of CTs; multi-organ segmentation; precise-lesion computation and post-processing.

5.4.4.2 Context-aware lesion augmentation model

The Baseline result contains some False Label (*FL*) and a few *FP* lesions, which can be mitigated by the context-aware lesion re-classification module in Figure 5.1. The Lesion Matching algorithm produces the prediction statistics and mask matching information, and the AugLesion algorithm generates lesion varied patches with adequate augmentation. The random combination between lesion croppings and lesion-free liver context make the model more robust to noises and non-tumor texture anomaly. During the process, we can not only adjust the augmentation strength, but also give more attention to *FP* and *FN* lesions. The context-aware augmentation process exposes 'hard textures' or error-prune lesions thoroughly and frequently in order to improve the re-classification purposely. During lesion model training, both the augmented CT image croppings and the corresponding *Gt* croppings are used. 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 with the correspondingly augmented prediction 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 it is re-classified as liver. In this way the context-aware lesion model significantly reduce the *FP* predictions with little increase of *FN*.

The organ model firstly generate a base detection result, then the lesion model refine the results by excluding major *FP* predictions. The mask computation, lesion cropping, lesion retrieval, re-classification, metric computation are all conducted in an automatic manner. The resulting context-aware lesion augmentation model (Sens1 w. Patch) has

1.8% and 0.3% detection accuracy increases compared to the Baseline model (Sens1 w/o Patch) in the four-phase and NC-phase settings respectively in Figure 5.1. While the lesion detection accuracy boost is not obvious, the main advantage of the lesion patch re-classification lies in the reduction of False Positives, where the specificity improves 5.7% and 12.8% in the four-phase and NC-phase settings respectively.

5.4.4.3 High lesion sensitivity model

Building on the baseline model, the context-aware lesion augmentation process ($Sens_1\ w.\ patch$) reduces *FP*s at the cost of slightly increasing *FN*s. To minimize the missing detections, we increase the scaling factor for lesion labels in Equation 5.1. Higher sensitivity effectively brings down the *FN*s, at the cost of more *FP*s. With the help of context-aware lesion re-classification module, $Sens_4\ w.\ patch$ (lesion sensitivity factor=4) model keeps a low number of *FN*s while maintaining an acceptable *FP*s. In Table 5.1, $Sens_4\ w.\ patch$ 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 advantage becomes even larger when it comes to level-specific accuracy. These comparisons clearly prove the combined power of lesion sensitivity scaling and context-aware reclassification modules.

5.4.4.4 Performance comparisons

We investigate the effect of context-aware lesion augmentation module on patient-level malignancy prediction by comparing $Sens_1\ w.\ patch$ against $Sens_1\ w/o\ patch$ in

Table 5.1. Although the sensitivity slightly decreases, the precision and specificity get substantially enhanced, especially in the NC phase setting (3.4% and 12.8% increase). 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 avoid this drawback. 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 the baseline 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 four-phase and NC-phase 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. 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 the four-phase and NC-phase settings.

$Sens_4\ w.\ patch$ discovers more lesions but increases the *FN*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 jointly make the prediction based on consensus, emulating the varied judgement strictness from different radiologists. The $Sens_4\ w.\ patch$ differs from $Sens_1\ w.\ patch$ by segmenting more lesions with slightly enlarged volumes, usually not changing the main lesion type. When the main lesion type differs from these two, the patient case is marked out as uncertain. By referring to the common ground of multi-sensitivity patient-level results, we achieve both low *FP*s and low *FN*s. The consensus portion takes the majority (97% and 92% in four-phase and NC-phase respectively), and the disagreed cases usually result from highly suspicious lesions requiring further examination.

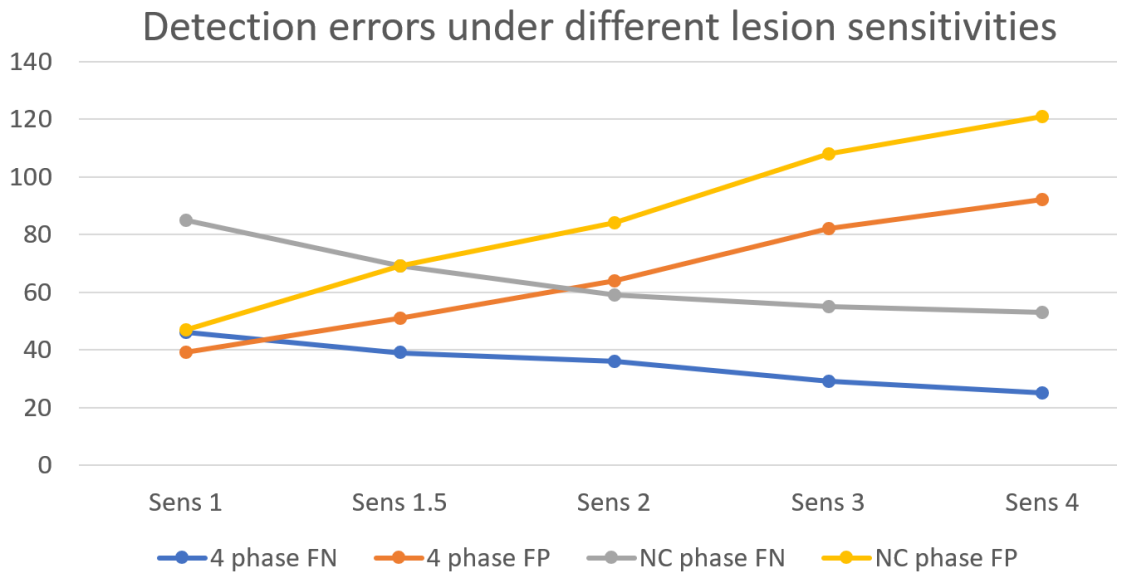

Figure 5.4: The effect of lesion sensitivity scaling factor f. Compare lesion *FN*s and *FP*s using $Sens_f\ w/o\ patch$ model with $f = 1, 1.5, 2, 3, 4$ respectively in both the four-phase and NC-phase settings.

5.5 Ablation study

5.5.1 The adjustment of sensitivity scaling factor

Increasing the lesion factor f in Equation 5.1 reduces the *FN*s while increasing the *FP*s. We compare the scaling f effect on the $Sens_f\ w/o\ patch$ model in Figure 5.4. The *FN* reduction (primary goal) slows down beyond $f = 3$, while the *FP*s keeps increasing (side effect) all the time. Therefore the appropriate sensitivity scaling factor f would be around 3 or 4.

5.5.2 Context-aware augmentation

The dedicated context-aware lesion patch module is able to reduce *FP* detections from the organ/lesion segmentation output because of the adequate random cropping and combination of lesion-free context and lesions. There are many nuances such as lesion cropping range, negative patch cropping, combination manner, augmentation strength. Particularly interested in the effect the negative context in the lesion re-classification process, we compare the context-ignorant ($Sens_1\ w.\ patch\ (no\ context\ aug)$) with context-aware $Sens_1\ w.\ patch$ augmentation in Table 5.35.4. For the training and testing preparation, the context-ignorant model randomly crops the lesion, without explicit combination of non-lesion regions. It can control the *FP*s with similar strength, but at the cost of increasing *FN*s. On the other hand, the context-aware patch augmentation achieves low *FP*s while retaining the *FN*s, which proves to be more robust.

5.5.3 The consensus of Multi-sensitivity models

The proposed ($Sens_1\ w.\ patch$+$Sens_4\ w.\ patch$) work flow utilize the computational modules and context-aware re-classification modules in different lesion sensitivity levels, and divide the results into the *consensus* and the *uncertain* portions. We show the consensus portion performances in Table 5.55.6. 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 whole 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, Cholangio, Metastasis *FL*s seem high, the model usually classifies them as malignant lesions, which explains the high malignancy precision. At the patient-level, the proposed model achieve similar malignancy precision and specificity in the NC-phase setting compared to in the four-phase setting, which is phenomenon.

5.6 Discussion

5.6.1 Dataset limitations

Our data is from a regional medical center for liver diseases, where the patients have much larger chances for tumors or other lesions. In the testing set, only 15% cases have not any liver lesions, different from the regular clinical reality. In the opportunistic setting, the majority patients would not contain malignant liver tumors. Solving the distribution limitations requires data source selection from multiple cross-region medical centers and

larger amount of lesion annotation, which are difficult and currently unavailable. Acknowledging the limitations, the performances in this paper only apply to similar data distributions and may vary substantially on other datasets. However, general principles of lesion sensitivity adjustment, region computation, context-aware lesion augmentation would apply to liver tumor detection tasks in other datasets.

5.6.2 The prediction errors

The proposed mode eventually has controlled the *FP*s, *FN*s, *FL*s to a small portion. Many *FP*s are highly suspicious even to radiologists. The majority *FN* lesions are well blended with surrounding tissues which are hard to tell from. The majority *FL*s occur among malignant lesion types, especially the lesion lacks the typical characteristics. We examine some prediction errors in Figure.

5.6.3 The performance upper bound and evaluation applicability

Due to the complex diseases and various textures in the liver, there exist many ambiguous lesions which could not be definitely classified to 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 biology tissues which 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 Conclusion

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

Algorithm 2: Region Computation: GenRegion

Input: $\mathbf{M} \in \mathbb{R}^{Z,Y,X}$
from *imantics* **import** $Polygons, Mask$;
$\mathbf{active\_Rs} \leftarrow \emptyset$;
$\mathbf{finish\_Rs} \leftarrow \emptyset$;
for z = *0,1,...,Z-1* **do**
    $\mathbf{polygons}, \mathbf{labels} \leftarrow Polygons(\mathbf{M}_z)$;
    $\mathbf{masks} \leftarrow Mask(\mathbf{polygons})$;
    UpdateActive$(\mathbf{masks}, \mathbf{labels}, \mathbf{active\_Rs}, z)$;
    UpdateFinish$(\mathbf{active\_Rs}, \mathbf{finish\_Rs}, z)$;
end
Def *UpdateActive*$(\mathbf{masks}, \mathbf{labels}, \mathbf{active\_Rs}, z)$**:**
    for m, l *in* $\mathbf{masks}, \mathbf{labels}$ **do**
        $matches \leftarrow \emptyset$;
        for *region in* $\mathbf{active\_Rs}$ **do**
            if *region[label] != l* **then**
                continue
            end
            if *region[z-1][mask] & m* **then**
                update *region*[z][mask] with m ;
                $matches$ += *region* ;
            end
        end
        TryMergeRegions$(matches)$;
    end
Def *UpdateFinish*$(\mathbf{active\_Rs}, \mathbf{finish\_Rs}, z)$**:**
    for *region in* $\mathbf{active\_Rs}$ **do**
        if *region[z][mask]* == $\emptyset$ **then**
            $\mathbf{finish\_Rs}$ += *region* ;
            $\mathbf{active\_Rs}$ -= *region* ;
        end
    end
Result: regions in $\mathbf{finish\_Rs}$

Algorithm 3: Lesion Matching: MatchLesion

Input: $\mathbf{Pred} \in \mathbb{R}^{Z,Y,X}$, $\mathbf{Gt} \in \mathbb{R}^{Z,Y,X}$
$\mathbf{TP}, \mathbf{FL}, \mathbf{FP}, \mathbf{FN} \leftarrow \emptyset, \emptyset, \emptyset, \emptyset$;
$V\_thres \leftarrow 0.5\ cm^3$;
$\mathbf{Pred\_U} \leftarrow$ Union([$\mathbf{Pred}$], Lesion_Ls) ;
$\mathbf{Pred\_UR} \leftarrow$ Restore($\mathbf{Pred\_U}$) ;
$\mathbf{Pred\_URV} \leftarrow$ VolumeFilter($\mathbf{Pred\_UR}$) ;
$\mathbf{Pred\_Rs} \leftarrow$ GenRegion($\mathbf{Pred\_URV}$) ;
$\mathbf{Gt\_Rs} \leftarrow$ GenRegion($\mathbf{Gt}$) ;
$\mathbf{PG\_U} \leftarrow$ Union([$\mathbf{Pred\_URV}, \mathbf{Gt}$],Lesion_Ls);
$\mathbf{U\_infor} \leftarrow$ GenInfor($\mathbf{PG\_U}, \mathbf{Pred\_Rs}, \mathbf{Gt\_Rs}$) ;
for *region in* $\mathbf{U\_infor}$ **do**
    gt_l, *gt_V* = Aggregate(*region*[gt]) ;
    pred_l, *pred_V* = Aggregate(*region*[pred]) ;
    if *gt_V* $<$ *V_thres* **then**
        if *pred_V* $>$ *V_thres* $+$ *0.1* **then**
            $\mathbf{FP}$ += *region* ;
        end
        continue ;
    end
    if *pred_V* $<$ *V_thres* $-$ *0.1* **then**
        $\mathbf{FN}$ += *region* ;
    else
        // may use lesion-level criteria instead ;
        if *pred_l* $==$ *gt_l* **then**
            $\mathbf{TP}$ += *region* ;
        else
            $\mathbf{FL}$ += *region* ;
        end
    end
end
Result: $\mathbf{TP}, \mathbf{FL}, \mathbf{FP}, \mathbf{FN}, \mathbf{U\_infor}$

Algorithm 4: Context-aware: AugLesion

Input: $\mathbf{U\_infor}, \mathbf{Img}, \mathbf{Gt}$
$\mathbf{CombImgs}, \mathbf{CombMasks} \leftarrow \emptyset, \emptyset$;
AugRepeat $\leftarrow 10$;
$ZC, YC, XC \leftarrow 8, 128, 128$;
$V\_thres \leftarrow 64\ cm^3$;
for *iter in AugRepeat* **do**
  for *region in* $\mathbf{U\_infor}$ **do**
    if *region[Volume]* $>$ *V_thres* **then**
      continue ;
    end
    LesionImg, LesionMask $\leftarrow$
      RandomCrop($\mathbf{Img}, \mathbf{Gt}$, *region*) ;
    NegImg, NegMask $\leftarrow$
      RandomCrop($\mathbf{Img}, \mathbf{Gt}$);
    k $\leftarrow$ randInt(ZC) ;
    CombImg $\leftarrow$ Concate(NegImg[:k],
      LesionImg, NegImg[k:]) ;
    CombMask $\leftarrow$ Concate(NegMask[:k],
      LesionMask, NegMask[k:]) ;
    $\mathbf{CombImgs}$ += CombImg ;
    $\mathbf{CombMasks}$ += CombMask ;
  end
end
Result: $\mathbf{CombImg}, \mathbf{CombMask}$

Table 5.1: 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%

Table 5.2: The test set lesion distribution. Lesions smaller than 0.5 cm^3 are excluded during evaluation. Number of cases containing the lesion is in the last column.

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 5.3: Lesion-level performance comparisons of model variants in 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%

Table 5.5: In the proposed work flow, 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%

Chapter 6

Conclusions and Future Perspectives

We have covered both the technology basis and applications of medical imaging analysis in previous chapters. Now we will summarize the critical aspects and discuss some concerns. From these valuable experience and discussion, we can grow and create better systems. In the end, we will look into the future, discussing about the technological development, the 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 researches and high-performance computing. Convolutional neutral network can extract vision pattern 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.

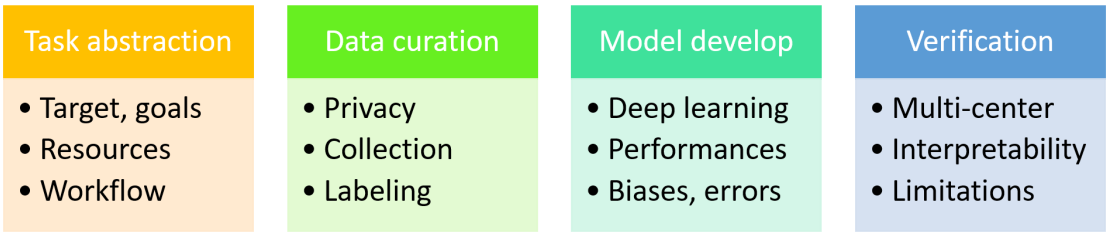

Figure 6.1: The formulation of medical image analysis tasks. There are multiple aspects to consider, including task abstraction, data curation, model development, verification.

As an important sub-branch of computer vision, medical imaging analysis benefits from deep learning technologies. Combining the anatomy constraints and medical knowledge with scanning images, medical imaging tasks can be solved with elegance and satisfaction. Deep learning models for medical image analysis have been applied in many clinical settings, diagnosing diseases such as cancer, Alzheimer's Disease, Covid-19. With careful design and verification, it is foreseeable to play more roles 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/redesign, training and testing. The differences are also obvious, such as privacy concerns, data sharing and labeling, clinical interpretability. A mindful formulation calls for close collaboration between researchers, engineers and doctors. Successful deployments depend on efforts from on multiple parties, who conduct multiple rounds of verification.

Broadly speaking, medical imaging tasks includes not only radiography and MRI, but also ultrasonography, histopathology and photoplethysmogram. Radiography and MRI images generate 2D or 3D projection of internal body parts, with high resolution and

accuracy, making them the first choice for disease screening and diagnosis. Depending 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 for patients infected with Covid-19.

There may be many challenges for different tasks, but the common one is data collection and sharing. Medical data collection is restricted by government and medical centers, and the patients must be acknowledged. Researchers must follow the Helsinki declaration and acquire ethical permission for the data collection and modeling. Before the patient data leave the hospital's database, thorough data cleaning to eliminate personal information must be conducted first, to protect privacy. The collected data should be protected and limited to only proved 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 distributions in the society for several reasons. Firstly, medical centers usually cover a regional population, bearing specific ethnicity characteristics. And many diseases are directly influenced by living environments and habits. Secondly, sample collection rate for patients may vary because of disease severity, economic status, willingness to share. Therefore the final samples tend to represent patients with severe diseases in many tasks. Thirdly, the training of convolutional neural networks encourage models to predict conservatively, biased toward classes with smaller sample numbers. 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 treats are necessary, promoting clinical interpretability. Before the real application, the model also need to go through verified in cross-center clinical 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-ray with DXA measurements. After studying the task characteristics and identifying the challenges, we propose the attentive multi-ROI workflow, which can yield 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, transformer-based feature fusion. Given chest X-ray image, we train models to predict BMD of four lumbar vertebra independently in four-fold cross-validation scheme. We package all the modules into a docker image, and develop a HTTP server for user interactions through web browsers.

Spine vertebra localization and identification via CT images. Vertebra detection in the CT scans has several challenges, namely narrow field of view, large spine curvature, metal implantation, vertebra fracture, scanning noise. Previous methods rely on techniques such as feature fusion, message passing, joint prediction, 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 by cancer mortality, and it has increasing trends in many countries. One way to mitigate the society of cancer is screening and detecting cancer at early stages. Low-dose CT scanning is being used in many body checking scenes, and it can be used to examine the 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 reach high sensitivity and specificity for malignancy detection at the patient-level.

6.2 Future perspective

6.2.1 Prospective directions of medical imaging analysis

Deep learning models have achieved super-human abilities in recognizing everyday objects. Given enough training samples for distinguishable patterns of predefined classes, 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 learning, adversarial learning have been successfully adopted in computer vision tasks. It is reasonable

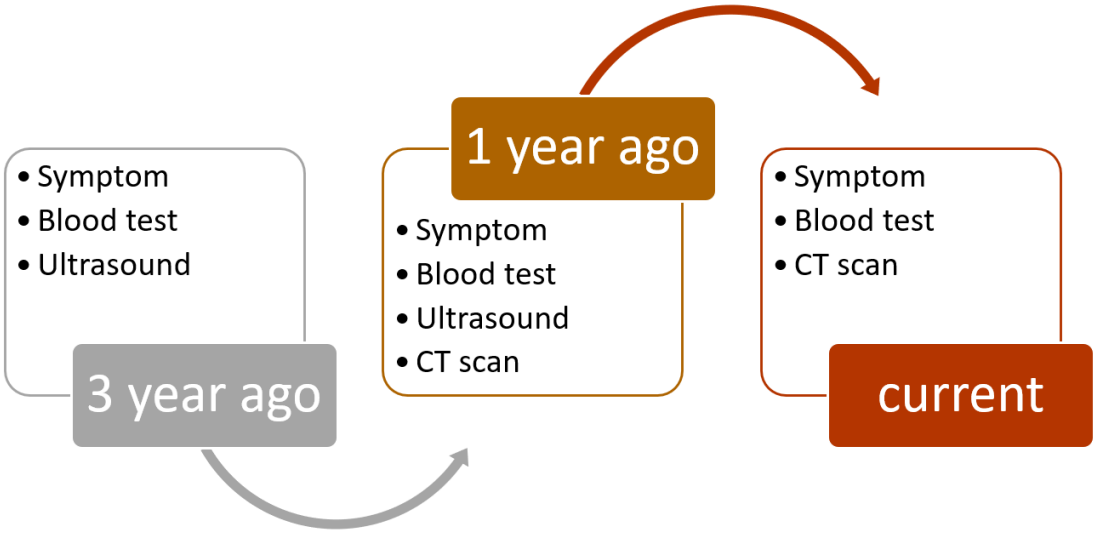

Figure 6.2: The longitudinal study involves health data from different periods.

to assume that deep learning models could achieve similar degree of success given enough qualified training samples, utilizing radiology images for defined disease diagnosis in any single organ.

The longitudinal characteristics of human disease should be emphasized with more importance in the medical image analysis. As a integrated process of disease research and prevention, the relevant practitioners should pay attention not only to current medical status of the patients, but also the disease patterns in historical manifestation. For example, the radiology scans of the same patient at different time would record the disease changing patterns. By monitoring and comparison, medical care provider could identify emerging anomalies for suggest follow-ups. Machine learning models can sense subtle or suspected changes, in a much better way. Another example is using historical medical scans to prediction disease progression. Every patient has unique genes, habits, conditions, which influence the lesion developments. Customized medical care requires doctors know patient health records and provide more precise treatment. Based on both general and personalized

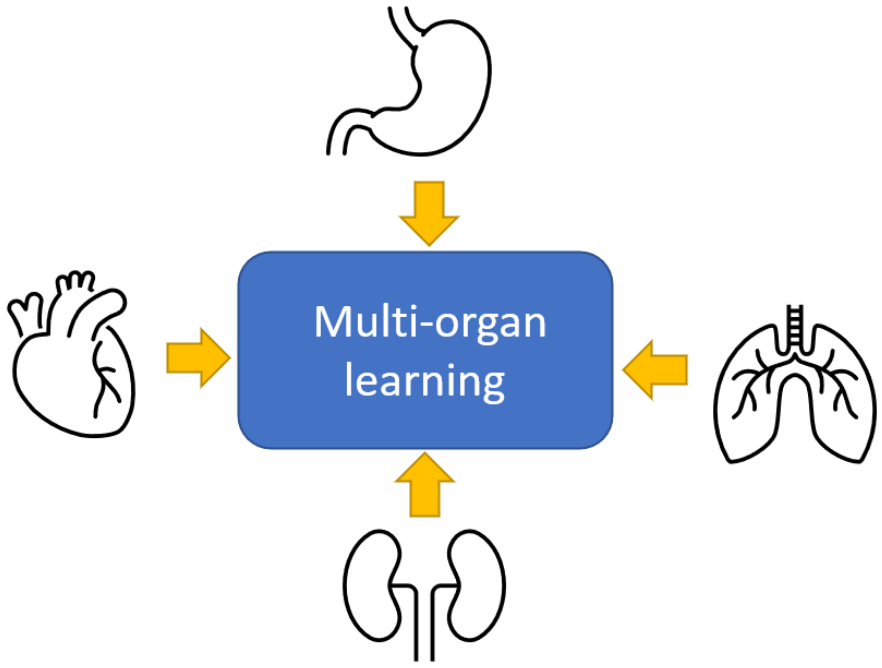

Figure 6.3: Many diseases show symptoms simultaneously at related organs. The collective information may benefit machine learning models.

disease evolving patterns, it is possible to make customized treating plans under expected conditions. Deep learning models would play an important role in pattern learning and prediction.

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 [107]. 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, pancreas, liver, and gallbladder. The malfunction of liver may lead to abnormal status of

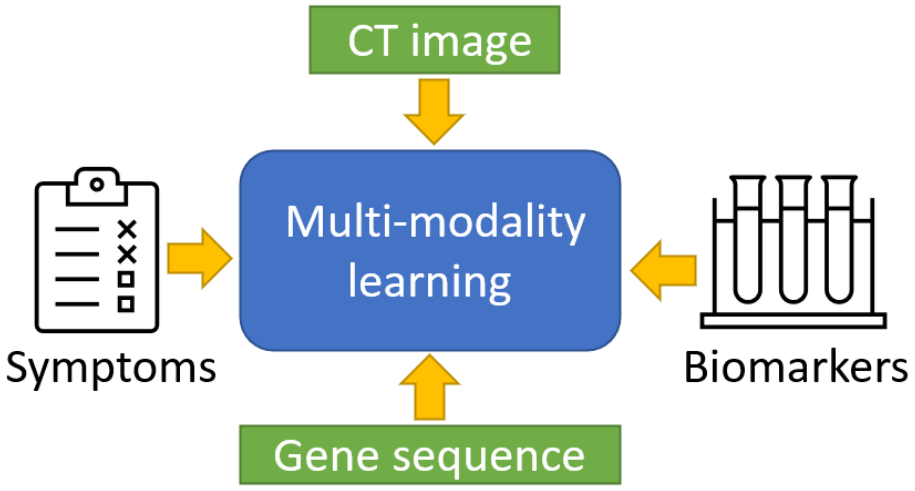

Figure 6.4: Some diseases not only have visual changes in the organ, but also bring in bio-marker anomaly. In clinical reality, patient would use multiple examinations to make judgement for the best possible accuracy. Therefore machine learning models need to learn multi-modality patterns for the disease classification.

gallbladder, subsequently increasing burdens of 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 are necessary. When the more comprehensive medical imaging data is available in the future, deep learning models for the multi-organ joint analysis are expected to improve our understanding of many diseases.

Multi-modality analysis would contribute better understanding of disease. Disease diagnosis and therapy 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 appearance, the genes. Radiology analysis on the organ alone sometimes would not capture the body condition accurately, so it needs multi-modality learning of related information sources to understand and monitor diseases in a more comprehensive way.

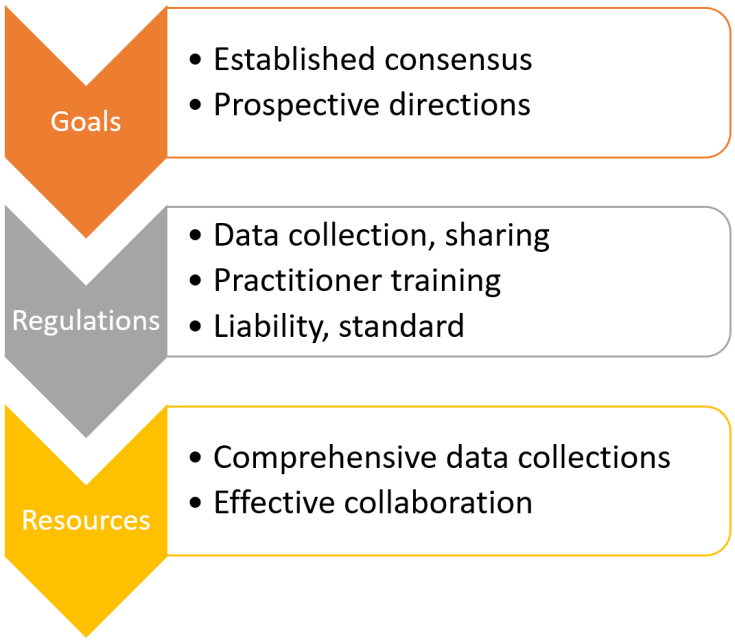

Figure 6.5: The future perspective for the medical imaging community. Although deep learning is promising and expected to revolutionize the medical care industry, there are challenges and risks ahead. As these changes are associated with many parties and resources in the society, it would be a long process.

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 the medical care is an essential part in any society, the medical standards, quality, affordability, 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, utilization to protect privacy and to promote the industry. Medical centers and practitioners should be trained to harness the computer-aided diagnosis and therapy facilities. Through continues integration and coordination, the quality and efficiency

are expected to improve. More elaboration in liabilities of computer-aided diagnosis is needed. Human 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 longitudinal study, multi-organ joint analysis, multi-modality analysis. In the clinical practice, the treatment for many diseases already involve comprehensive analysis, but medical image analysis community has not seen enough related studies due to several reasons. First and foremost, there is not enough public comprehensive data collections. Collecting such kind of data relies on both clear medical goals and clinical support. Secondly, comprehensive disease understanding is essential to design and implement machine learning models for multi-source tasks. Therefore it calls for close and effective collaboration between doctors and researchers, which is not the current reality for various reasons.

6.2.3 The limitation of deep learning applications

How much will Artificial Intelligence replace human labor in the 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 human in many medical imaging tasks, which are defined by established knowledge and practices. However, machine learning models have its limitations, resulting from data collection, labeling, training. It can learn and distinguish common lesions and symptoms in the radiology

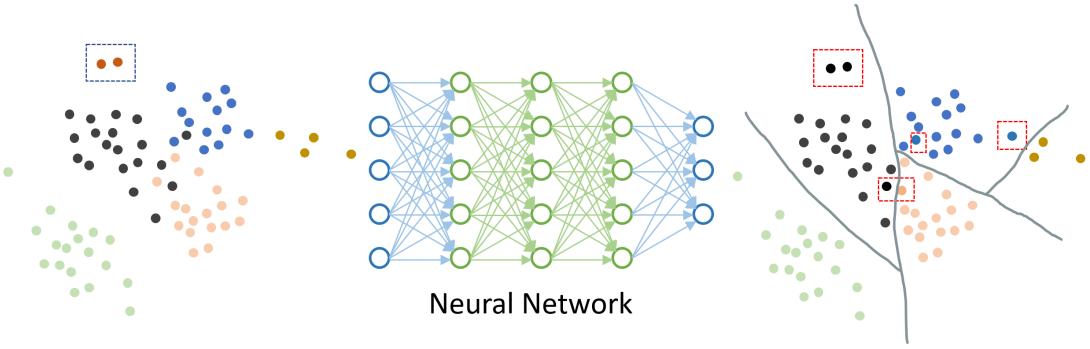

Figure 6.6: The neural network learns to classify data points in its representative space. The points on the left are the input data, where the colors represent the ground truth labels. The points on the right are the output, where the colors represent the predicted labels. The red cases are not learnt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

images quite well because of large quantity 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 mis-classify 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 the diagnosis scenarios, deep learning model predictions can serve as an efficient reference. However, the doctor usually need to conduct multiple examinations to make the final judgement. Only human has the extensive knowledge and logical reasoning to specify rare cases, to identify new discoveries, and to make sensible judgement in situation of conflicts.

[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] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

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

[24] 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.

[25] 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.

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

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

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

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

[44] 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.

[45] 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.

[46] 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.

[47] 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.

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

[49] 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.

[50] 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.

[51] 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.

[52] 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.

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

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

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] 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.

[76] 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.

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

[78] 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.

[79] 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.

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

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

[82] 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.

[83] 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.

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

[85] 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.

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

[87] 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.

[88] 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.

[89] et al. Patrick Bilic, Patrick Ferdinand Christ. The liver tumor segmentation benchmark (lits). *ArXiv*, abs/1901.04056, 2019.

[90] 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.

[91] 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.

[92] 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.

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

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

[95] 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.

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

[97] 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.

[98] 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.

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

[100] 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.

[101] 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.

[102] 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.

[103] 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.

[104] 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.

[105] 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.

[106] 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.

[107] 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.