Quantifying Uncertainty in Landmark Detection For Medical Imaging

James McCouat

St Hughs's College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Michaelmas 2023

Abstract

Machine learning, specifically deep learning, has the potential to revolutionise the way medical images are analysed by providing reliable, fast and accurate diagnostic readings automatically. This work focuses on a type of machine learning called landmark detection where a model predicts the location of keypoints (landmarks) of clinical interest in medical images. It demonstrates how a deep Convolution Neural Network (CNN) can be used to detect landmarks from which clinically useful angles and diagnoses are calculated. In particular, the method presented is used to diagnose Femoracetabular Impingement (FAI) from X-rays of the pelvis and Developmental Dysplasia of the Hip (DDH) from ultrasound infant hip images.

Moreover, this thesis proposes a new approach for assessing the uncertainty of landmark predictions. Uncertainty, or how 'sure' the model is in its predictions, has not been explored thoroughly in past landmark detection literature. However, it is a crucial aspect for the successful translation of this technology into clinical practice. This thesis presents a novel statistic called Expected Radial Error to quantify the uncertainty of landmark predictions. This statistic flags to the clinician when the model is likely to be inaccurate, thus enabling the clinician to efficiently review and modify the automatic landmark predictions to ensure their accuracy.

This thesis shows that the CNN model chosen achieved state-of-the-art landmark detection results on two public datasets. It also achieved a 91.96% and 87.50% diagnostic accuracy on the two types of FAI, and 82.98% accuracy at diagnosing DDH. In addition, the landmarks flagged by the ERE-based alerting system were, on average, at least 50% less accurate than those which were not flagged on each dataset experimented with. This means that the alerting system correctly flagged landmarks which had been predicted incorrectly, and would improve clinical confidence in the system if it were translated into the clinic.

Instead of following the status quo in the medical imaging AI field by proposing yet another network architecture to achieve slightly improved results on publicly available data, this thesis presents how CNN models can be used to diagnose orthopaedic conditions automatically and how an uncertainty measurement can be produced to build clinical trust.

Quantifying Uncertainty in Landmark Detection For Medical Imaging



James McCouat

St Hughs's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2023

Acknowledgements

Institutional

First and foremost I would like to thank my supervisors Dr Irina Voiculsecu and Professor Sion Glyn-Jones for their supervision. Irina offered an exceptional level of mentorship during my DPhil and I am sincerely grateful for the consistent meetings we held to discuss the direction of my research. I am additionally thankful for her guidance and support on the papers we published together. Sion Glyn-Jones offered invaluable medical data and expertise during my DPhil which provided me with a tangible understanding of how the methods I developed can make a meaningful impact in clinical settings. This is a rare experience for a Computer Scientist like myself. I would also like to thank my other clinical collaborators, Dr Abinav Singh and Professor Dan Perry, for providing data and guidance in ultrasound imaging. I enjoyed the fruitful collaboration we had in the final year of my DPhil.

Personal

I am immensely grateful for the friendships I've made during my DPhil journey. Firstly I would like to thank Andrius Vaicenavicius and Natasha Jeppu for being awesome office mates in my first year. I would like to thank Madeline White for all the virtual and non-virtual pub quizzes we have done together and Anna Pathak who became my climbing buddy and Go-playing friend.

I would like to thank Sandy MacAulay, Matthew Newton and Romain Guyon for being fantastic flatmates. You always provided a fun and supportive house, with plenty of jokes and competitiveness. Additionally, the Oxford University Squash Club has been a major part of my life during my DPhil and I would in particular like to acknowledge my teammates: Tom Paine, Tim Delport and Sam Greenrod.

I would like to thank my parents, Roger and Jane McCouat, for always believing in me so that I could achieve this. Finally, I would like to thank my girlfriend Benedetta Mussati for the great times we have had together and for being incredibly supportive through all the highs and lows of my DPhil journey.

Abstract

Machine learning, specifically deep learning, has the potential to revolutionise the way medical images are analysed by providing reliable, fast and accurate diagnostic readings automatically. This work focuses on a type of machine learning called landmark detection where a model predicts the location of keypoints (landmarks) of clinical interest in medical images. It demonstrates how a deep Convolution Neural Network (CNN) can be used to detect landmarks from which clinically useful angles and diagnoses are calculated. In particular, the method presented is used to diagnose Femoracetabular Impingement (FAI) from X-rays of the pelvis and Developmental Dysplasia of the Hip (DDH) from ultrasound infant hip images.

Moreover, this thesis proposes a new approach for assessing the uncertainty of landmark predictions. Uncertainty, or how 'sure' the model is in its predictions, has not been explored thoroughly in past landmark detection literature. However, it is a crucial aspect for the successful translation of this technology into clinical practice. This thesis presents a novel statistic called Expected Radial Error to quantify the uncertainty of landmark predictions. This statistic flags to the clinician when the model is likely to be inaccurate, thus enabling the clinician to efficiently review and modify the automatic landmark predictions to ensure their accuracy.

This thesis shows that the CNN model chosen achieved state-of-the-art landmark detection results on two public datasets. It also achieved a 91.96% and 87.50% diagnostic accuracy on the two types of FAI, and 82.98% accuracy at diagnosing DDH. In addition, the landmarks flagged by the ERE-based alerting system were, on average, at least 50% less accurate than those which were not flagged on each dataset experimented with. This means that the alerting system correctly flagged landmarks which had been predicted incorrectly, and would improve clinical confidence in the system if it were translated into the clinic.

Instead of following the status quo in the medical imaging AI field by proposing yet another network architecture to achieve slightly improved results on publicly available data, this thesis presents how CNN models can be used to diagnose orthopaedic conditions automatically and how an uncertainty measurement can be produced to build clinical trust.

Contents

List of Abbreviations **xi**

1 Introduction **1**
1.1 Thesis Inception . . . 2
1.2 Initial Methodology . . . 3
1.3 Generalisation . . . 4
1.3.1 Definitions . . . 4
1.4 Research Outcomes . . . 5

2 Background **7**
2.1 Automation in Medical Imaging . . . 8
2.2 Deep Learning . . . 10
2.2.1 Computer Vision Tasks . . . 12
2.2.2 Challenges and Competitions . . . 14
2.3 Landmark Detection . . . 15
2.4 Landmark Detection in Medical Imaging . . . 17
2.4.1 The Random Forest Method . . . 17
2.4.2 Emergence of Deep Learning . . . 17
2.4.3 Recent Architectures . . . 19
2.4.4 Self and Semi-Supervised Approaches . . . 20
2.5 Medical Landmark Detection Observations . . . 22
2.5.1 Comparison with Segmentation . . . 22
2.5.2 Comparison to Non-Medical Landmark Detection . . . 23
2.5.3 Gaps in the Current Literature . . . 25
2.6 Uncertainty . . . 29
2.6.1 Reliability Diagrams . . . 30
2.6.2 Temperature Scaling . . . 32
2.6.3 Uncertainty in Landmark Detection . . . 34
2.6.4 Gaps in Current Literature . . . 37

3 Datasets **39**
3.1 Introduction . . . 40
3.1.1 Dataset Types . . . 40
3.1.2 Ethical Considerations . . . 43
3.2 Cephalometric Dataset . . . 44
3.3 Hand Dataset . . . 49
3.4 Pelvis Dataset . . . 53
3.4.1 Femoroacetabular Impingement . . . 53
3.4.2 LCE-angle and α-angle . . . 54
3.5 Infant Hip Dataset . . . 59
3.5.1 Developmental Dysplasia of the Hip . . . 59
3.5.2 α-angle and β-angle . . . 60
3.6 Dataset Characteristics . . . 64
3.6.1 Landmark Characteristics . . . 64

4 Kolmogorov-Smirnov Test **67**
4.1 Disadvantages Of Random Set Selection . . . 69
4.2 Disadvantages Of K-Fold Cross Validation . . . 70
4.3 Kolmogorov-Smirnov Test Method . . . 71
4.3.1 Choosing The Number Of KS Tests To Perform . . . 71
4.4 Experiments Performed . . . 72
4.5 KS Test Results . . . 76
4.5.1 Distribution Of KS Test Scores . . . 76
4.6 Generalisability Of The KS Method . . . 79
4.7 Conclusion . . . 81

5 Landmark Detection **83**
5.1 Experimental Setup . . . 87
5.1.1 Preprocessing . . . 87
5.1.2 Post-processing . . . 91
5.1.3 Validation Metrics . . . 91
5.2 Hyperparameter Tuning . . . 94
5.2.1 Experiments Performed . . . 94
5.2.2 Data Augmentation . . . 95
5.2.3 Heatmap Generation . . . 99
5.2.4 Architectures . . . 102
5.2.5 Loss Function . . . 104
5.2.6 Hyperparameter Summary . . . 105
5.3 Publicly Available Dataset Results . . . 106
5.3.1 Cephalometric Dataset Results . . . 106

5.3.2 Hand Dataset Results . . . 109
5.3.3 Public Datasets Results Conclusion . . . 111
5.4 Pelvis Dataset Results . . . 112
5.4.1 Experiments Performed . . . 112
5.4.2 Landmark Detection Results . . . 112
5.4.3 LCE and Alpha Angle Results . . . 116
5.4.4 FAI Diagnosis Results . . . 118
5.5 Infant Hip Dataset Results . . . 120
5.5.1 Experiments Performed . . . 120
5.5.2 Landmark Detection Results . . . 120
5.5.3 Alpha and Beta Angle Results . . . 123
5.5.4 DDH Diagnosis Results . . . 124
5.6 Discussion and Conclusion . . . 125
5.6.1 Advantages of the Technology for Clinics . . . 125
5.6.2 Landmark Detection and Angle Calculation . . . 126
5.6.3 Diagnostic Performance . . . 127

6 Utilizing Output Heatmaps 129
6.1 Visual Characteristics of Output Heatmaps . . . 135
6.1.1 Observations . . . 135
6.1.2 Ideas . . . 135
6.2 Reliability Diagrams For Landmark Detection . . . 138
6.2.1 Method . . . 138
6.2.2 Experiments Performed . . . 138
6.2.3 Results and Conclusion . . . 141
6.3 Temperature Scaling . . . 144
6.3.1 Experiments Performed . . . 144
6.3.2 Temperature Scaling Results and Conclusions . . . 147
6.4 Expected Radial Error (ERE) . . . 149
6.4.1 Definition . . . 149
6.4.2 Examples . . . 156
6.4.3 Correlation . . . 156
6.4.4 Comparison To Highest Heatmap Value . . . 160
6.4.5 ERE Throughout Training . . . 161
6.4.6 ERE Conclusions . . . 163
6.5 Using ERE To Flag Potentially Unacceptable Predictions . . . 164
6.5.1 Method . . . 164
6.5.2 Experiments Performed . . . 165
6.5.3 ROC Results . . . 167

6.5.4 Quantitative and Qualitative ERE Results 167
6.5.5 Conclusion 173
6.6 Ensembling 174
6.6.1 Formal Definitions 174
6.6.2 Experiments Performed 175
6.6.3 Ensembling Results 178
6.6.4 Uncertainty in an Ensemble 179
6.6.5 Ensembling Conclusions 184
6.7 Hip Studio 185

7 Conclusion **191**
7.1 Summary 191
7.2 Filling the Gaps in the Literature 194
7.3 Future Work 197
7.3.1 Obtaining Data 197
7.3.2 Experimental Setup for Diagnostic Tasks 198
7.3.3 Applying Uncertainty to Angle Predictions 198
7.3.4 Using Uncertainty to Detect Incorrect Labels 199
7.3.5 Improving Hip Studio 199
7.3.6 Translating Software into Clinic 200

Appendices

A Dataset Images **203**

B Landmark Detection Examples **209**

C Pelvis Angle Predictions **221**

D Infant Hip Angle Predictions **225**

E Experimental Code **229**

F Failed Approaches **233**

References **237**

List of Abbreviations

AI . . . . . - Artifical Intelligence

AFPF . . . . . - Attentive Feature Pyramid Fusion

AP . . . . . - Anteroposterior

BCNN . . . . . - Bayesian Convolutional Neural Network

CLM . . . . . - Constrained Local Model

CNN . . . . . - Convolutional Neural Network

CPU . . . . . - Central Processing Unit

CT . . . . . - Computed Tomography

DDH . . . . . - Developmental Dysplasia of the Hip

ECE . . . . . - Expected Calibration Error

ERE . . . . . - Expected Radial Error, a measure proposed in this thesis

FAI . . . . . - Femoroacetabular Impingement

GAN . . . . . - Generative Adversarial Network

GPU . . . . . - Graphics Processing Unit

Grad-CAM . . - Gradient-weighted Class Activation Mapping

HOG . . . . . - Histogram of Oriented Gradients

ICC . . . . . - Intraclass Correlation Coefficient

KS . . . . . - Kolomorogov-Smirnov Statistic

LCE . . . . . - Lateral Center-Edge

MRE . . . . . - Mean Radial Error

MRI . . . . . - Magnetic Resonance Imaging

NLP . . . . . - Natural Language Processing

NHS . . . . . - National Health Service

NDORMS . . - Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences

OA . . . . . - Osteoarthritis

PACS . . . . . - Picture Archiving and Communication System

PCA . . . . . - Principal Component Analysis

ROC . . . . . - Receiver Operating Characteristic

RoI . . . . . - Region of Interest

SIFT . . . . . - Scale-Invariant Feature Transform

SOTA . . . . . - State of the art

SDR . . . . . - Successful Detection Rate

SVM . . . . . - Support Vector Machine

TPU . . . . . - Tensor Processing Unit

VGG . . . . . - Very deep CNN proposed in Simonyan & Zisserman [119]

ViT . . . . . - Vision Transformer

YOLO . . . . . - You Only Look Once (architecture proposed in Redmon *et al.* [109])

1 Introduction

Contents

1.1 Thesis Inception **2**
1.2 Initial Methodology **3**
1.3 Generalisation **4**
1.3.1 Definitions 4
1.4 Research Outcomes **5**

1.1 Thesis Inception

My DPhil work in the area of machine learning applied to medical images began by exploring how a condition called Femoroacetabular Impingement (FAI) could be diagnosed automatically from X-ray images of the pelvis. This was the first project of my DPhil because my two co-supervisors, Irina Voiculescu and Siôn Glyn-Jones, had an existing collaboration which involved the sharing of medical data and expertise between the Computer Science Department and the Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences (NDORMS). Prof Glyn-Jones provided a dataset of pelvis X-ray images acquired by a previous clinical trial (FAI Trial [102]) that had already undergone experimentation.

Not only did Prof Glyn-Jones provide the medical imaging data but also expert medical knowledge of the FAI condition, and highlighted the benefits an automatic diagnostic solution for FAI would bring to the clinic. FAI is a condition which causes damage to soft tissue and cartilage in the hip socket that may then progress to the development of Osteoarthritis (OA). OA is a degenerative joint disease which restricts a person's movement ability, can cause severe pain, and limits their ability to perform normal daily tasks [37]. With an ageing population, a large proportion of older adults will develop OA conditions that need to be treated. In 2018 the NHS spent £10.2 billion on OA and rheumatoid arthritis.[1] The majority of this cost is due to the limited and late-stage interventions on offer for OA with 109,000 people undergoing hip replacement surgery in England and Wales in 2019 [139]. If an automatic diagnosis solution to FAI could be developed then this would increase the amount of patients whose FAI is detected and treated using physiotherapy or hip arthoscopy surgery, which would lead to fewer patients developing more severe OA.

Any technology with this functionality has the potential to be a huge cost-saving to the NHS by reducing the amount of time clinicians need to spend interpreting images, minimizing the need for additional scans, and expediting surgery dates, thereby removing the need for long-term physiotherapy. It would streamline clinical workflow, reduce clinician fatigue and avoid patients living in pain for as long.

[1] https://www.england.nhs.uk/blog/tackling-the-elephant-in-the-room/

1.2 Initial Methodology

To automatically diagnose FAI from X-rays a machine learning model must be able to automatically calculate two angles on each hip present in the image. These angles are called the Lateral Center Edge (LCE) angle and the α-angle. Currently clinicians manually measure these angles for each X-ray using software in the clinic to diagnose FAI.

Although the benefits of an automatic diagnostic solution for FAI were clear, it was not initially clear what methodology could automatically calculate the necessary angles. It was evident from the literature that there were two possible avenues for automatically calculating the angles required for FAI diagnosis. Both methods involve training a machine learning model. However, the model could either automatically detect the shape of the hip socket and femur bone in a process called segmentation or, alternatively, it could detect certain landmarks (keypoints) which could be connected to calculate angles. The latter approach was chosen as a promising area of research. Apart from the technical reasons, there was also the compelling reason of generating ground truth. Any machine learning model needs to be trained on ground-truth annotations, and it is much faster to obtain landmark dots than segmentation masks or contours.

For this reason the work presented in this thesis focuses on automatically detecting clinically useful landmarks, in a process called landmark detection, from hip X-rays in the first instance, and also from medical images in general. This thesis shows that these landmarks can be connected to measure angles and perform diagnosis on orthopaedic medical images. Machine learning makes this process automatic, fast and reliable. The vision for this technology is to build it into the Picture Archiving and Communication System (PACS) software installed on NHS computers across the UK. Current software does not use AI to increase the speed at which clinicians can diagnose and process patients.

1.3 Generalisation

As well as testing the landmark detection method proposed in this thesis on pelvic X-ray images it was also tested on a dataset containing ultrasound images of infant hips, obtained from another clinical collaborator, and two publicly available datasets of cephalometric (head) and hand X-rays. This ensured the robustness of the proposed methods by testing them on datasets of different sizes, modalities, and which capture different anatomy.

Additionally, a method that outputs an uncertainty measurement with each landmark prediction is presented in this work. This provides a measure of how 'sure' the model is in its predictions. Uncertainty has not been explored in existing medical imaging landmark detection literature but will be very useful when the technology is translated into the clinic. This work shows that the uncertainty measurement proposed can flag and alert the clinician when the model has potentially produced an incorrect reading which would build clinical trust in the model.

1.3.1 Definitions

This thesis describes work completed in the field of machine learning applied to medical images. Machine learning is an area of Computer Science concerned with building methods that learn from data and then make independent decisions, without being explicitly programmed, to perform a task. Machine learning is a type of Artificial Intelligence (AI) which is an umbrella term describing any algorithm which performs a task perceived as requiring intelligence close to human level [66]. Medical imaging is the process of viewing an interior part of the human body for medical intervention. It can be performed via different modalities including, X-ray, magnetic resonance imaging (MRI) and ultrasound. Different modalities are often used to diagnose conditions in different body parts depending on what anatomical structures the clinician would like to view to make a clinical decision. In this thesis machine learning is applied to medical images to automatically extract information from the image or scan such as automatic measures of the patient or a diagnosis.

1.4 Research Outcomes

The research presented in this thesis has, to date, been published or presented in the following proceedings and conferences:

- *Automatically Diagnosing Hip Conditions from X-rays using Landmark Detection* - Published in the proceedings of the 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) [89]. Presents the Kolmogorov-Smirnov test presented in Chapter 4 and the landmark detection results for the Pelvis dataset shown in Section 5.4.

- *Automatic Identification of Clinical Landmarks and Calculation of Graf Angles in Newborn Hip Screening: A Pilot Study* - Presented at the British Orthopaedic Association 2022 [120] (Third author). Presented the landmark detection results for the Infant Hip dataset as shown in Section 5.5.

- *Contour-Hugging Heatmaps for Landmark Detection* - Published in the proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2022 [88]. Presents the 'Expected Radial Error' statistic described and shown in Chapter 6.

- *Fast and Accurate Automatic Measurements of the Hip from X-rays Using AI* - Poster presentation at the Orthopaedic Research UK conference 2022 [90]. Presented the Hip Studio software shown in Section 6.7.

2 Background

Contents

2.1 Automation in Medical Imaging 8
2.2 Deep Learning 10
2.2.1 Computer Vision Tasks 12
2.2.2 Challenges and Competitions 14
2.3 Landmark Detection 15
2.4 Landmark Detection in Medical Imaging 17
2.4.1 The Random Forest Method 17
2.4.2 Emergence of Deep Learning 17
2.4.3 Recent Architectures 19
2.4.4 Self and Semi-Supervised Approaches 20
2.5 Medical Landmark Detection Observations 22
2.5.1 Comparison with Segmentation 22
2.5.2 Comparison to Non-Medical Landmark Detection 23
2.5.3 Gaps in the Current Literature 25
2.6 Uncertainty 29
2.6.1 Reliability Diagrams 30
2.6.2 Temperature Scaling 32
2.6.3 Uncertainty in Landmark Detection 34
2.6.4 Gaps in Current Literature 37

2.1 Automation in Medical Imaging

Research into using computers to increase the speed and accuracy at which patients could be diagnosed from medical images started not long after the invention of the computer, with early work dating back to 1960 [85]. It was thought at this time that computers had the potential to diagnose conditions faster and more accurately than humans because of the speed and reliability with which computers could perform calculations [33]. However, due to the constraints on medical image acquisition, the computational power of computers, and the statistical tools available at the time, early work was limited. An example of this is Meyers *et al.* [92] who, in 1964, used a computer to scan a medical image stored on magnetic tape, record the readings and then pass those readings through an algorithm that contained many conditional statements to produce a diagnosis.

Progress in this area was slow until the 1990s when medical images became digitized and statistical methods improved. In the Rueckert *et al.* [112] 1999 paper they use affine transformations and non-rigid deformations to register breast images to make it easier to spot cancer. Shape and image-based approaches such as active contours [15, 25] were popular in this period. These approaches had the advantage that prior knowledge could be incorporated however they lacked generalisability and the ability to cope with unusual cases.

By 2012 machine learning methods such as Principle Component Analysis (PCA), Support Vector Machines (SVMs), and linear regression had been experimented with; however, with increasingly powerful computer processors, new methods such as Random Forest began to emerge. For example, the 2014 paper Menze *et al.* [91] showed that the best-performing approaches for brain tumour segmentation used the Random Forest approach. This began a trend towards more powerful but also more uninterpretable methods.

In 2015 a huge shift towards deep learning took place with the advent of the Convolutional Neural Network (CNN). The CNN adapted the neural network, which had existed for decades, so that predictions could be made on images. A seminal paper is Ronneburger *et al.* [111] who proposed the U-Net CNN architecture,

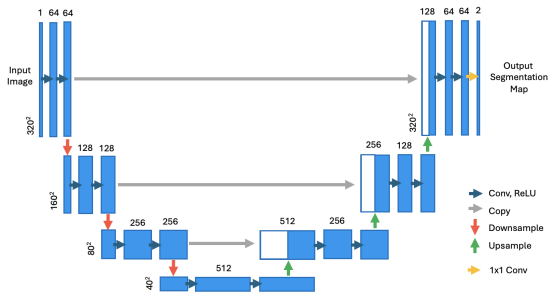


Figure 2.1: The U-Net architecture by Ronneburger *et al.* [111] proposed the idea of having a contracting and an expanding path, on the left and right side of the U respectively, with residual connections between them. The contracting path downsamples the images and the expanding path upsamples them so the U-Net learns features at different scales. This means that the U-Net learns features that represent the entire image and also small visual patterns, like corners and edges.

as shown in Figure 2.1, which achieved groundbreaking results at segmenting microscopy images as well as being easy to reproduce for other researchers. Until today, deep learning models continue to dominate the research landscape in this area, and the rate of publications has grown enormously [126]. This is hugely helped by faster processors, cloud computing, the accessibility of code on GitHub, the **`python`** programming language, and libraries, such as **`PyTorch`** and **`TensorFlow`**, which make it easy to set up and modify deep learning methods.

It is worth noting, however, that despite the growth of medical image analysis as a research field, and the acquisition of vast volumes of data in hospitals, this progress has not yet been fully translated into clinics, despite several approaches claiming better results than a doctor [7, 65, 95]. This lack of translation could be down to non-technological barriers such as patient data privacy, regulation, IT infrastructure, financial barriers, or the ability to commercialize work from universities. It could also be because of the un-interpretability and un-explainability of deep learning models, or that even though models are often accurate in a test setting they do not generalize well to new data. Both of these disadvantages can, rightly or wrongly, also lead to a lack of trust in artificial intelligence from the public, which is not exclusive to medical imaging.

2.2 Deep Learning

Deep learning is a subset of machine learning which, in turn, is a subset of artificial intelligence. Machine learning is defined as algorithms that 'train' on a set of data and then make predictions on new unseen data. These predictions are seen to require some degree of intelligence to make. Although not a deep learning technique, the machine learning method called Random Forests, which is an ensemble of decision trees, achieved good results on several medical imaging tasks during the early 2010s [82, 91]. However, it was soon surpassed by Krizhevsky *et al.* [70] in 2012, who proposed a deep CNN to classify images and achieved far higher classification performance than previous methods, helped by the recent advances in powerful GPUs.

This paper started a revolution where networks began to further leverage computation power, often by designing deeper networks, such as the 'Very Deep' CNN named VGG [119] and GoogLeNet [125]. In addition, the ResNet [55] (Residual Network) showed residual connections, where information is passed deeper into the network without being transformed, improved accuracy and reduced the computation overhead of networks. The aforementioned networks perform classification but networks trained to perform other tasks improved using a similar technique. For object detection; R-CNN [45], Fast R-CNN [44], Faster R-CNN [110], and YOLO [109] (You Only Look Once) each improves upon the previous, generally, by converting more of the detection pipeline into a CNN which is end-to-end trainable. For segmentation, Long *et al.* [86] proposed the fully convolutional network, a network not containing any fully connected layers, and then Ronneburger *et al.* [111] invented the U-Net architecture which extends this idea by adding an expanding path to produce more accurate segmentations. The Mask R-CNN [54] combines the Faster R-CNN and a U-Net-like architecture to achieve both object detection and segmentation. The contracting and expanding path, connected by residual/skip connections presented in the U-Net paper became very influential in the computer vision area and many papers leveraged this idea further, such as the Stacked Hourglass Network [96] and the Deep High-Resolution Network [124], both

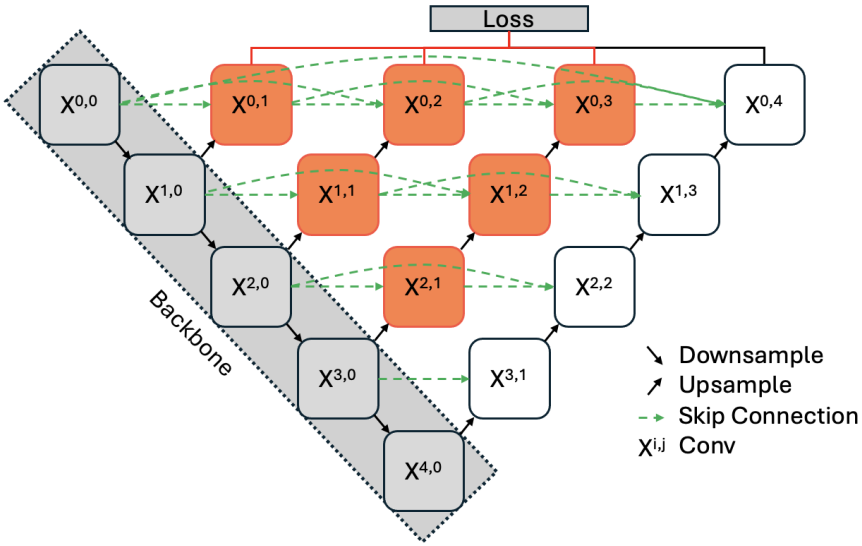


Figure 2.2: The UNet++ architecture, as proposed by Zhou *et al.* [149], builds upon the U-Net by nesting U-Nets inside each other and adding intermediate losses.

of which originally performed keypoint detection but, in more recent literature, have also been applied to segmentation and object detection [60, 136].

After the U-Net paper was published in 2015 the computer vision field grew quickly, with a rapid increase in papers published, often presenting new architectures, achieving SOTA performance, on a dataset. However, new difficulties began to appear in the field. The 'no-new-U-Net' paper [59], published in 2018, found that the original U-Net architecture could achieve better results than many newer, more complicated, architectures if it was tuned optimally. This paper concludes that many papers do not optimize the previous approaches which they then claim to perform better than. However, some new networks did achieve reproducibly better results over a variety of tasks such as the UNet++ [149], as shown in Figure 2.2.

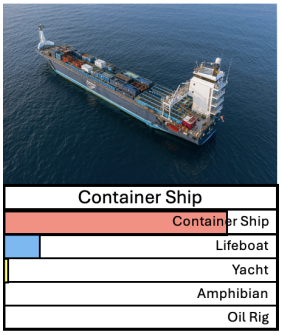
A new architecture called Transformer has recently come to the fore, significantly outperforming the previous SOTA on many imaging challenges. Originally invented for natural language processing (NLP) [132], it was translated to computer vision with the Vision Transformer (ViT) [35] which outperformed past classification networks. This concept was then extended to segmentation in the form of the Swin (Shifted-Windows) Transformer [84]. The Transformer architecture uses a concept called Attention which allows the model to extract information from the entire

image more effectively than a CNN. The Attention mechanism has time complexity $\mathcal{O}(n^2)$ in its native form and, as such, these networks leverage compute power even further than CNNs. This has partially been facilitated by the invention of the Tensor Processing Unit (TPU) which has even more processing power than a GPU. This processing power is financially expensive to acquire and run which means that an ever-increasing proportion of high-impact research in computer vision is conducted by large companies such as Google [35] and OpenAI [108] who have access to thousands of GPUs or TPUs. This is to the detriment of conventional solo researchers and academic groups.

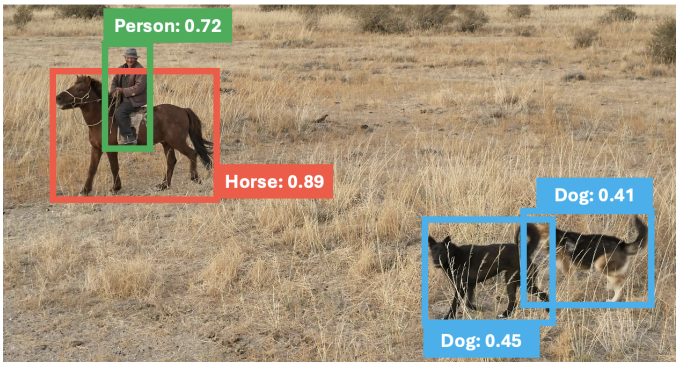
2.2.1 Computer Vision Tasks

Deep learning networks perform one, or more, predefined tasks. A task is defined by what is output or predicted by a network. The most common tasks are listed below.

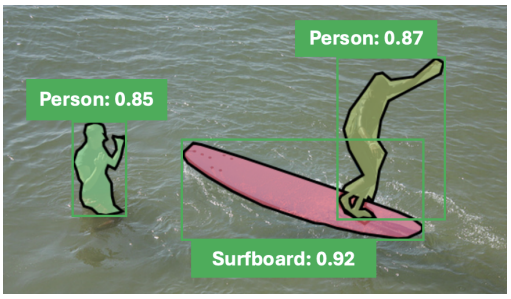
- **Classification** is a method which predicts a certain label for an image or a region of an image (Figure 2.3a). Classification can be either binary, where either a yes or no is predicted, or multi-class, where a label is predicted from a predefined set.
- **Object Detection** is a method which predicts bounding boxes around specific objects in an image (Figure 2.3b), and predicts the class of object contained within each bounding box. The bounding boxes are often called regions of interest (RoI) and the task of predicting their coordinates can be called object localisation.
- **Semantic Segmentation** is a method which predicts a class for each pixel in an image, where each class has a human interpretable meaning (Figure 2.3c). It is used to generate precise contours or masks around features of interest.
- **Landmark Detection**, otherwise known as Keypoint Detection, is a method which outputs 2D coordinates of significance in an image (Figure 2.3d).



(a) Image classification with probabilities shown for the most likely labels. This figure shows an image being predicted as a 'container ship'.



(b) Object detection puts bounding boxes around entities captured in an image. In this case, it is combined with classification to label each bounding box. This figure shows an image where four objects are detected, two of which are predicted as dogs, one as a person, and one as a horse.



(c) Semantic segmentation assigns certain pixels in an image to a class. This figure shows entities segmented alongside object detection on an image from the COCO dataset [80]. In this image, two surfboards and a person have been segmented as well as detected.



(d) Landmark detection localises points in an image, in this case, points are placed on human joints and connected with line segments. This figure shows landmark detection on a set of skiers.

Figure 2.3: Different machine learning tasks.

2.2.2 Challenges and Competitions

Since 2015, challenges based on publicly available datasets have become very popular in computer vision, with researchers aiming to outperform the current best method and become the new SOTA . Popular datasets include ImageNet [31], COCO (Common Objects in Context) [80], PASCAL VOC[1] (Visual Object Classes),[2] and ADE20K (20K Images).[3] The two best-known challenges in landmark detection are HumanPose, where models predict the joint positions of humans in images (Figure 2.3d and Figure 2.5), and facial landmark detection, where models predict key landmarks on a face (Figure 2.4). The latter can be used for downstream tasks such as facial or emotion recognition.

Labels for different machine learning tasks exist for some of these datasets, for example, object detection, landmark detection, and segmentation labels are all available for the COCO dataset. In addition to these general-purpose vision datasets, medical imaging challenges are also popular such as the BraTS (Brain Tumour Segmentation),[4] and the CheXpert (Chest X-ray) challenge.[5] Websites such as Kraggle,[6] which host challenges, and papers with code,[7] which keep track of previous methods' results, make it easier to keep up-to-date with the SOTA and participate in challenges. Challenges have accelerated research into methods for certain problems, however, the research area could be focusing too heavily on achieving the SOTA for certain challenges instead of making progress in translating these models to real-world applications.[8]

[1] http://host.robots.ox.ac.uk/pascal/VOC/
[2] https://groups.csail.mit.edu/vision/datasets/ADE20K/
[3] https://groups.csail.mit.edu/vision/datasets/ADE20K/
[4] https://paperswithcode.com/dataset/brats-2018-1
[5] https://paperswithcode.com/dataset/chexpert
[6] https://www.kaggle.com/competitions
[7] https://paperswithcode.com/sota
[8] https://www.technologyreview.com/2020/08/18/1007196/ai-research-machine-learning-applications-problems-opinion/

2.3 Landmark Detection

Landmark detection is the central topic of this thesis. It is more commonly known as keypoint detection in non-medical applications but in this thesis, for consistency, I will use 'landmark detection' throughout. The landmarks predicted by a landmark detection method are predefined and the method will make a location prediction for each of these landmarks for every image. An assumption this task makes is that each image contains all the predefined landmarks. Finding which landmarks are present in an image, as opposed to their location, is a separate and distinct task.

Landmark detection literature has followed a similar path to other machine learning tasks in moving from statistical models to deep learning approaches. Originally methods used filters such as Scale-Invariant Feature Transform (SIFT) [87] or Histogram of Oriented Gradients (HOG) [30] filters to extract features which are then input into deformable models such as active shape models [26] and constrained local models [28]. Other methods in this period include active contour models, otherwise known as 'snakes' [62], generalized Hough transform [10] and Support Vector Machines (SVMs) [13]. The Random Forest Regression-Voting method for feature extraction, which became prominent, involves decision trees which take in many quickly-computed features and vote for the placement of each landmark [81].

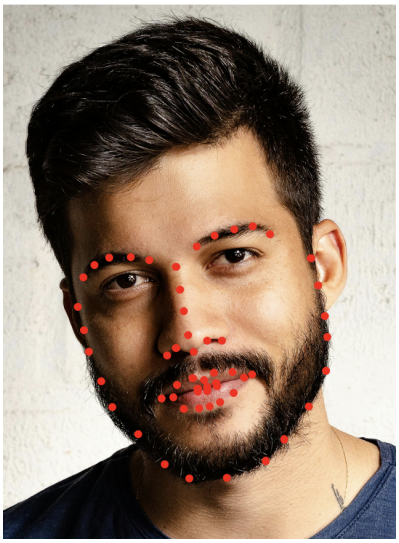


Figure 2.4: An example of facial landmark detection. 68 landmarks are detected and from their positions facial expressions and emotions can be predicted.



Figure 2.5: The image on the left shows the final human pose prediction, and the images on the right show training Gaussian heatmaps for landmarks situated on the ankle, knee, hip, shoulder and head.

After the publication of AlexNet [70], the landmark detection area increasingly began using CNNs. These outperformed Random Forest approaches because convolutions can capture more complex features than simple statistics such as averaging and max pooling used in Random Forest approaches. Landmark detection was performed by CNNs predicting (regressing) landmark coordinates [130] until Tompson *et al.* [129] proposed training a network with Gaussian 'heatmaps' with a certain standard deviation centred at the ground truth points. This latter approach achieved good results and using heatmaps centred on the ground truth became the most popular method in literature to the present day. An example of these heatmaps is shown in Figure 2.5. It is worth highlighting that in the context of this thesis, the word heatmaps means either the two-dimensional value maps which are used for training models or 'output heatmaps' which are output from a model. **Heatmaps in this thesis (discussed in detail in Chapter 6) are distinct from heatmaps used in the context of saliency maps [118] and GRAD-CAM [116].**

The publication of the U-Net architecture [111] also affected the landmark detection area with new networks such as hourglass [96] and cascaded pyramid network (CPN) [22] leveraging contracting and expanding paths to become the SOTA. The deep high-resolution network further leveraged this idea by effectively stacking hourglass networks together horizontally and vertically [124]. To date, Transformers have not made a big impact on landmark detection however it is likely this is due to how recently vision Transformers have been published, and the lack of sufficiently large amounts of training data, rather than them not being capable to do so.

2.4 Landmark Detection in Medical Imaging

2.4.1 The Random Forest Method

At the International Symposium on Biomedical Imaging (ISBI) in 2014 a dataset of landmarks on head X-rays was released to the public in the form of a challenge for the conference. The top four achieving methods at the time all used a Random Forest approach [17, 23, 58, 93] using different hand-crafted features. The best performing work by Ibragimov *et al.* [58] achieved its results by applying a superior shape model however even for this method only 70.21% of landmarks were detected within the clinically accepted threshold of 2mm. In addition, on the hardware available at the time, methods took between 20-90 seconds to produce predictions for an image.

In 2016, the approach of Ibragimov *et al.* was superseded by Lindner *et al.* who used a constrained local model (CLM) on top of a Random Forest. They had already developed this method and published work on facial landmark detection and hand X-rays [81]. This was the SOTA at the time but still only achieved 73.68% detections under the 2mm threshold.

2.4.2 Emergence of Deep Learning

During 2016 Random Forests and CNNs achieved similar results [123]. Urschler *et al.* [131] proposed a method of Random Forests for local and global contexts. That work rejects the CNN method because it *"depends on the availability of large training datasets and is limited by the need to restrict resolution of training images due to computational complexity and memory requirement"*. Whilst this was true given the computational resources available at the time, GPUs quickly became more powerful.

From 2017 onward all approaches achieving SOTA results have used CNNs. In 2017 Arik *et al.* [8] used a CNN on patches to output the probability of a landmark being at the centre of each patch. This was because computation resources were still limited and a whole image could not be passed into a GPU. A similar approach was taken by Lee *et al.* [75]. The disadvantage of these methods is that a separate CNN is required for each landmark. Subsequently, Payer *et al.* [103] published a

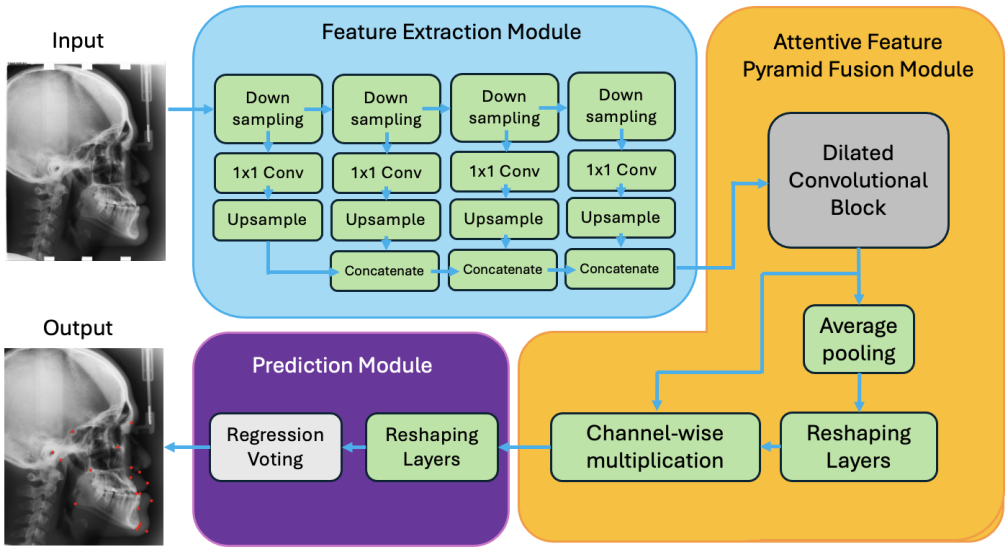


Figure 2.6: An illustration of the Attentive Feature Pyramid Fusion (AFPF) architecture proposed in Chen *et al.* [19] and also used to achieve SOTA results with a region attentive loss in [20]. It used a VGG network [119] as the feature extraction module then a dilated convolutional block [144] performed attention on the output before being passed to a final prediction module.

work which used a single U-Net architecture and training heatmaps proposed by Tompson *et al.* [129], as shown in Figure 2.5, to achieve results on an X-ray hand dataset. Lui *et al.* [83] and Wu *et al.* [138] also achieve SOTA results on X-ray hip datasets using a single CNN network and Gaussian heatmaps.

Some methods use a CNN to perform segmentation of relevant structures, from which landmarks are then computed via deterministic algorithms. Golan *et al.* [46] use a Generative Adversarial Network (GAN) to obtain realistic segmentations and then use least squares optimization to produce lines between landmarks. Hareendranathan *et al.* [52] and El-Hariri *et al.* [53] define the highest or lowest pixels in segmentation maps as their landmarks.

2.4.3 Recent Architectures

Approaches after 2019 often try to overcome the problem that, due to memory constraints, images have to be input into a deep learning model at a downsampled resolution. Medical landmark detection methods should extract both long-range and high-resolution information an image. Long-range information is required to locate landmarks that are in the middle of structures where an understanding of the overall anatomy is required. On the other hand, high-resolution information is needed for landmarks on boundaries where the edge of a bone needs to be captured in high resolution. Chen *et al.* [19] proposed an Attentive Feature Pyramid Fusion (AFPF) architecture that inputs features at different levels of scaling into a dilated convolutional block [144] to capture maximum contextual information at different scales, as shown in Figure 2.6. An alternative solution is proposed by Zhong *et al.* [148] and Zeng *et al.* [145] who use a two-stage U-Net approach, where the first U-Net detects an RoI, where a landmark is likely to be, and then the second U-Net more accurately pinpoints its position. However, this approach has the disadvantage that it requires a CNN per landmark. CephaNet by Quan *et al.* [106] used an R-CNN [45] method to detect landmarks and a shape model to suppress false positive detections. Other methods which use CNNs but with a different focus include: Oh *et al.* [99] who proposed a loss function that has a penalty term based on a shape model, Dai *et al.* [29] who proposed a GAN [47] to learn displacement maps for each landmark and Hu *et al.* [57] who train a model to perform segmentation and landmark detection, with the theory that anatomical features will be shared between the two tasks.

There is evidence that the phenomenon described in the 'No New U-Net' paper [59] is present in medical imaging landmark detection. Yao *et al.* [141] showed that a U-Net can achieve results within 2% of the SOTA for landmark detection. This could be because they trained the U-Net on higher resolution images, enabled by recent GPU advances, or because previous methods never fully optimized the U-Net. Determining which conclusion is correct is challenging due to the complexity of reproducing SOTA architectures, especially when the code is not publicly available.

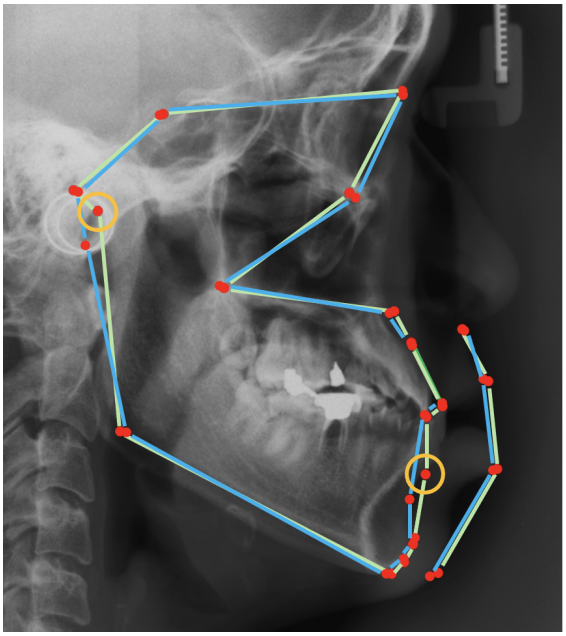


Figure 2.7: This figure illustrates the method proposed by Chen *et al.* [20]. The original landmark predictions are connected by green lines and the landmarks after shape adjustment are connected by blue lines. The circled landmarks are the most inconsistent with the shape model and need the most adjustment. The landmarks connected by blue lines are then fed back into the model for retraining. This process of generating pseudo-labels, adjusting using the shape model and then retraining is repeated until the position of the landmark predictions converges.

2.4.4 Self and Semi-Supervised Approaches

Some publications from 2020-2022 perform medical landmark detection but focus on maximising performance when given unlabelled or limited training data. Zhu *et al.* [150] train a model on different landmark datasets in the hope that the features learned by one dataset will help the other. The model shares a local network between the datasets but a global network is retrained. It achieves good results although it does not compare to all previous methods.

Some works aim to produce accurate landmark detections with limited training data. Yao *et al.* [142] aims to learn landmark predictions by training on only a single example, known as one-shot learning. This paper uses a technique called self-supervised learning where a model is trained to perform a task that requires no annotated ground truth, in this case predicting where a patch in an image was cropped from. After this model has been trained one image with manually-labelled

ground truth annotations is selected. Patches from this image, centred at the landmarks, are then taken. These patches are then passed into the model with new images so the model must predict where the patch is from in the image even though it was not cropped from that image. The location it predicts for each patch is a prediction of a landmark. These landmarks are called pseudo ground truth and once they have been generated, a U-Net architecture is trained on them to produce final landmark predictions.

The effectiveness of Yao *et al*'s method was found to be extremely dependent on which single image was chosen for manual annotation. Quan *et al.* [107] built on this work by researching how to choose the best single training image. Qian *et al.* devised a method of self-training a model using contrastive learning and then running this model over the dataset to obtain features. From these features a similarity score is calculated between each pair of images and the image similar to most other images is chosen as the training image. This is shown to improve the effectiveness of the method.

Finally, Chen *et al.* [20] is the only semi-supervised method of landmark detection on medical images. Semi-supervised methods utilise unannotated data alongside data with annotations. Chen *et al.* first train an AFTF model, proposed in their 2019 paper [19], on a small set of annotated images. This model is then run over the unannotated examples in the dataset to produce pseudo-labels for these images, similar to Dong *et al*'s mean teacher framework [34]. A PCA shape model, similar to those seen in early landmark detection work with Random Forests [58, 82] then refines these pseudo-labels. Figure 2.7 shows an example of this refinement process (described in the caption). The model is then retrained on the refined pseudo-labels and the process is repeated until there is a negligible difference between the pseudo-labels generated by consecutive iterations. This model is then tested over a test set containing separate images and the results reported.

2.5 Medical Landmark Detection Observations

2.5.1 Comparison with Segmentation

More publications in AI for medical imaging focus on segmentation rather than landmark detection. Many structures of clinical importance can be segmented in medicine such as tumours, organs, and blood vessels. The medical segmentation decathlon is a set of ten publicly shared datasets in segmentation which receives wide participation from researchers.[9] However, no orthopaedic dataset is included in the decathlon because orthopaedic conditions are often diagnosed by measuring angles or distances between landmarks. Some methods do semantically segment bony structures and then calculate landmarks from the segmentation masks on orthopaedic images [46, 52, 53, 147] however they have the following disadvantages:

- **Obtaining ground truth**. Ground truth landmarks are significantly quicker to draw than segmentation ground truths (contours or masks) because the clinician only has to click on the image to place landmarks as opposed to drawing a segmentation mask. This is especially significant in medicine where doctors are under significant time pressure. Moreover, a faster method of annotation leads to clinicians being less fatigued when drawing them which means they will likely be more accurate.

- **Algorithm complexity**. Performing segmentation to obtain landmarks means a function from segmentation to landmarks needs to be created. This algorithm is non-trivial and adds complexity to the pipeline.

- **Fragile pipeline**. It is possible that even small inaccuracies in the predicted segmentation mask could cause the landmarks to be placed incorrectly and thus performing segmentation brings an additional process into the pipeline which could malfunction.

[9] `http://medicaldecathlon.com/`

- **Difficult Validation**. Landmark detection has the advantage that the accuracy of landmark placements can be easily validated, using distance error, in an understandable and explainable way. There are numerous evaluation metrics for segmentations including Dice, intersection over union or Hausdorff distance, however, it is hard to interpret these readings in isolation and what they mean for any possible diagnosis [143].

For these reasons, and because we perform experiments on orthopaedic datasets obtained from clinical collaborations, **landmark detection is the focus of this thesis**.

2.5.2 Comparison to Non-Medical Landmark Detection

Although literature for medical imaging landmark detection has broadly followed general landmark detection by progressing from statistical methods to deep learning there are some key differences caused by discrepancies between medical and non-medical image datasets including:

Dataset Size

Medical imaging datasets are much smaller than general imaging datasets. Publicly available medical landmark datasets contain less than 1000 images, whereas general computer vision datasets such as COCO [80] or MPII [6] (used for human pose detection) have 328,000 and 25,000 images. Medical imaging datasets are harder to obtain because a collaboration with a hospital and regulatory approval is required in the UK. In addition, a qualified clinician has to do the annotating which takes up time in their busy schedule, unlike images of objects where labelling can be done by the general public.

Feature Similarity

Images within medical imaging datasets are more alike than images in general-purpose object detection datasets like ImageNet [31]. Datasets of photographs contain different objects, colours, patterns, textures, and lighting. As a result,

general-purpose deep network architectures need more layers and more weights than models trained on medical imaging datasets to cope with the large variations in images. In addition, these large models can be trained without overfitting because of the abundance and variety of training data. The most recent Transformer architectures, seen in general imaging, require substantial amounts of data to learn which is often not available in medical imaging [117]. Medical imaging datasets are greyscale and normally contain the same anatomy, with the same textures from the same view. In addition, medical datasets obtained for research purposes will often be more aligned than routine medical images taken in routine clinical practice.

Ambiguity

The ground truth in medicine can be ambiguous, clinicians often disagree on exact measurements [4]. In addition, a clinician can disagree with themselves when they reassess a patient's image for a second time. This ambiguity happens less often in general imaging datasets. The popular ImageNet dataset, for example, features images with a clearly visible object which can be clearly defined. Due to the ambiguity in the ground truth of medical imaging datasets, the maximum accuracy a model can achieve is more limited. For example, if the placement of a landmark is ambiguous and the inter-clinician variability of the landmark placement is 2mm then the highest accuracy the model can obtain is 2mm when compared to another clinician's ground truth placement. If this were achieved, the model could be said to be achieving the same accuracy as a clinician.

Due to this limit on the accuracy of a method it will be difficult to further improve the SOTA accuracy on publicly available medical imaging landmark detection datasets by any significant margin. More impactful research can be conducted by filling the gaps in the existing literature identified in Section 2.5.3.

2.5.3 Gaps in the Current Literature

Reproducibility

Unlike general imaging, where the major conferences require code to be shared, medical imaging does not require this. None of the recent papers achieving SOTA results share their code [19, 20, 145, 148]. In addition, these papers proposed complicated architectures, like the one illustrated in Figure 2.6, which are not easy to replicate, and often not all the hyperparameters to recreate the model are shared. Papers, especially those experimenting on challenge datasets, focus solely on the accuracy of their proposed architecture without considering how easy it would be to reproduce. There is also little fundamental theory presented as to why certain architectures work slightly better (2-5% improvement) than other architectures on a given dataset. Although there are exceptions, such as Payer *et al.* [104] and Chen *et al.* [19, 20], most publications only test on a single dataset, which, coupled with the difficulty of reproducing experiments, means that it is hard to establish a clear SOTA in the area. This problem is exacerbated by the fact that many publications use private data. For this reason, future work should move away from focusing on challenges and instead focus on proposing reproducible methods, tested on multiple datasets.

A Lack Of Images Taken During Routine Practice

The majority of previous machine learning medical imaging literature experiment with data acquired in a strict way to ensure that the images in the dataset are as uniform as possible. This is not representative of images that AI systems would see if deployed during routine practice in the clinic. It is not explored in medical imaging landmark literature what features are learned by these models and how the domain shift of data from a different scanner or scanner settings would affect the output. More collaboration between computer scientists and hospitals is required to source images from different sites and scanners and to provide guidance into what images can be taken in hospitals. Deployed AI systems should be able to

give assurance that they will work on images from different scanners and patients with different presentations.

Another reason why there is little experimentation on routine clinical datasets is that it is difficult for researchers to meet both the reproducibility criteria required for publication and the clinical criteria required for translation into the clinic. For experiments to be reproducible datasets need to be publicly available but this is often not possible with images taken in routine practice. A shift in how papers are evaluated for publication, or in how anonymised data is made available for research, could encourage more work to be performed on routine data.

Experimental Robustness

Roughly half of recent papers performing landmark detection on private medical imaging datasets only split their data into training and test sets with no mention of a validation set [46, 79, 83, 140]. One of these papers states that the test set was used to tune the network [46], however, most state that hyperparameters were tuned, but without giving details. Although it is possible these papers could be using part of their training set to tune parameters, it is also possible that the test set is being used. This is called data leakage and constitutes bad practice. The papers which do mention a validation set either state they split the total dataset randomly [48] or do not state how they split the dataset [14, 21, 68]. Splitting randomly may not be a robust method because medical imaging datasets tend to be small, often less than 200 images [14, 21, 68]. When using a small test set it is common that only a few difficult examples can fool the model and produce erroneous results that skew statistics like the mean localization error.

Robustness is also an issue for public datasets. Many works perform experiments on a Cephalometric (head) dataset. When it was released by Wang *et al.* [134] it had three pre-defined sets: 'Training' and 'Test Set 1' and 'Test Set 2'. Some works explicitly state that Test Set 1 is used as a validation set [19, 99]. Some explicitly state that the number of epochs the model trains for was decided by a validation set taken as a subset of the training set [76, 104, 145]. However most do not mention

how hyperparameter tuning was completed and just say the dataset was split into training and test sets [8, 20, 106, 141, 148, 150]. There is a problem with consistency and past approaches using a subset of the training data to tune are at a disadvantage.

Another publicly available dataset of hand landmarks has three training/test splits [103] defined by the original authors. Works perform validation on the first split and then present on an average of the three splits [20, 123]. Whilst this is a consistent protocol it is a form of data leakage between training and test sets.

In addition, no papers found during the literature review in medical landmark detection performed repeat experiments on the same test set. Training a deep learning network is a random process which results in different models even when the training data and hyperparameters are kept the same. Results should be reported as an average over several runs to obtain a fair representation of how effective a method is, especially when the accuracy of results for publications are within as little as 2% of each other.

Uncertainty

All AI works referenced are known to make mistakes. For example, even the SOTA for publicly available datasets [20] detects 2.46% of landmarks over a 4mm distance from the clinician's ground truth. It is important for patient safety and trust in the AI system, that the AI system can provide some levels of certainty with its predictions. Uncertainty is a large topic in general deep learning literature [2] however it has not been fully explored in landmark detection. Uncertainty encoding and assessment forms a major part of this thesis, and a comprehensive literature review is given in Section 2.6.

Requires Training Data

The vast majority of existing approaches train on more than 100 annotated images. Despite landmark detection methods requiring less annotation time than for a segmentation task, as discussed in Section 2.5.1, it is still important to reduce the amount of labelled data required for training because clinicians often have very limited time to spend annotating.

Few Ensembling Approaches

There are very few landmark detection ensembling approaches using CNNs in the literature. The key motivation behind ensembling models is to improve accuracy by obtaining a model that outperforms each of the individual models in the ensemble [101]. Random Forests are an ensemble method however they are inferior to the more recent CNNs for reasons discussed in Section 2.3. Dropout is another form of ensembling although it has been shown to not improve results over batch normalization for CNNs on several tasks [42].

Any method to ensemble CNNs needs a way of aggregating the predictions from the individual models, and the common methods of aggregation would not generalise well or would be sub-par for landmark detection. For example, the majority voting aggregation strategy is where the most common prediction made by an ensemble is chosen as the final prediction. This method is often used for classification tasks [5, 146], yet it does not generalise to landmark detection because the likelihood of two of the base estimators choosing the same pixel as their predicted point is very low relative to the total number of pixels in the image.

Some publications mention ensembling in a different context, such as Zou et al [151] use it to mean a group of landmarks in a hierarchical structure, rather than to mean a group of separate models. The only publication that has ensembled deep landmark detection models in Schobs *et al.* [115] who ensemble models in two ways: 1) By mean averaging the landmark positions predicted by the individual base learners and 2) By multiplying the output heatmaps of the individual base learners together and taking the hottest point. This area could be further explored. For example, a method of ensembling individual base learners based on their uncertainty is yet to be proposed.

2.6 Uncertainty

As an increasing amount of AI gets deployed into real-life situations, it becomes all the more important to make sure these systems are reliable [61]. Uncertainty exists in two main forms in machine learning. Firstly aleatoric, which is the uncertainty of data being fed to the network, for example, noise in the sensor or annotation ambiguity. Secondly epistemic, which is the uncertainty the model has in its prediction caused by a lack of knowledge. The total uncertainty of a model is aleatoric plus epistemic.

Ideally, deep learning models should output both a prediction and a level of certainty in that prediction, a capacity that humans naturally have [27]. Literature that proposes methods for outputting an uncertainty measurement for deep learning tasks include Harakeh *et al.* [51] and He *et al.* [56] for object detection, and Wang *et al.* [135] and Badrinarayanan *et al.* [9] for segmentation.

These methods can be split into two categories: sampling-based and sampling-free. Sample-based methods sample multiple predictions from a network to obtain a distribution for model uncertainty, methods within this category are: Bayesian neural networks [78], Monte Carlo Dropout [40], and bootstrap ensembles. Because multiple predictions are required for these methods, they are computationally expensive and have had to be adapted to work in the imaging domain [94]. However, these adapted methods are still slow to perform multiple predictions, especially for the deeper networks emerging in imaging [74].

The sample-free method, on the other hand, is not as slow because it outputs an uncertainty value directly alongside its prediction. Such networks are trained using a Gaussian log-likelihood loss function, a form of log-likelihood where the uncertainty output is taken into account [73, 74]. Sample-free approaches are considered the SOTA for uncertainty in imaging.

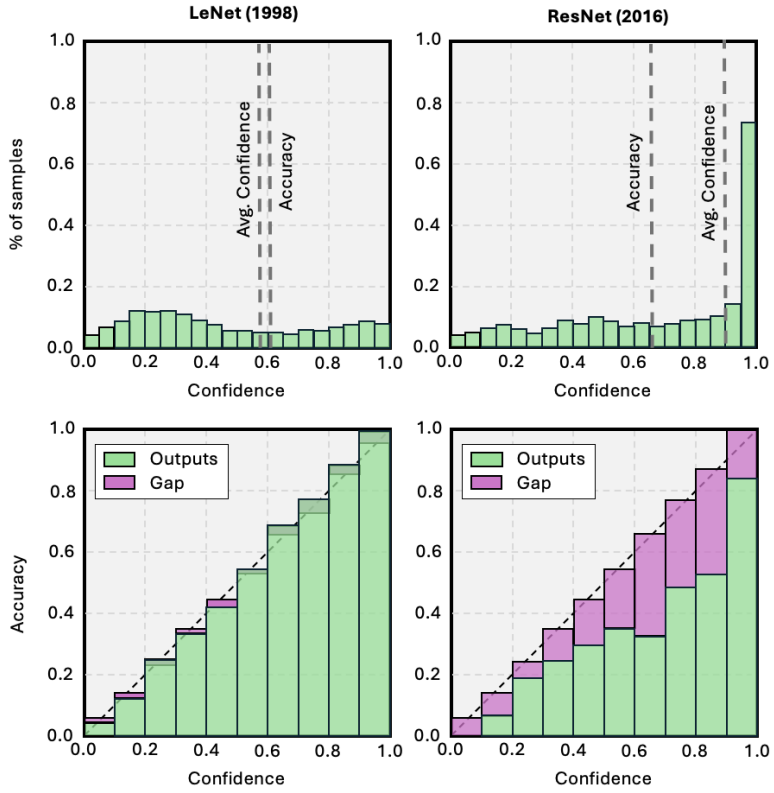


Figure 2.8: This figure shows the results presented in Guo *et al.* [50]. On the left is a five-layer LeNet from 1998 and on the right is a 110-layer ResNet [55] from 2016. The top graphs show that models have become overconfident over time. The average confidence a LeNet has in its predictions is similar to the accuracy of the predictions. On the other hand, a large proportion of the ResNet's predictions are overconfident. This skews the average confidence of predictions to a higher value. The bottom graphs are reliability diagrams [97] and work by putting each prediction into a bin based on its confidence score. Each bin is represented by a bar in the diagram where the green 'output' bar is the accuracy achieved by the predictions in a bin and the pink 'gap' bar is the difference between that accuracy and the average confidence in the bin. A well-calibrated network will have a small gap between the confidence of each bin and the accuracy of each bin, however, Guo found this does not happen in modern deep architectures.

2.6.1 Reliability Diagrams

When neural networks were first developed the values in their final layer (known as the prediction's confidence) could be used as an uncertainty measurement directly. For example, if the network output that an image was a 'dog' and the label 'dog' had a confidence value of 0.8 in the network's output layer then 80% of the time that image would truly contain a dog. This means that the network was 'calibrated'. However, in 2017 Guo *et al.* [50] showed that as neural networks have been getting

deeper their outputs have become less calibrated to accuracy and uncertainty. They showed this using reliability diagrams as seen in Figure 2.8. To compute reliability diagrams, each prediction made by a model is placed into one of M bins based on that prediction's confidence. Predictions those confidence value falls within the interval $(\frac{m-1}{M}, \frac{m}{M})$ where $m \in \{1...M\}$ are placed into bin B_m. Once predictions have been placed into these bins two values are computed for each bin. One of these values is the accuracy of the bin. This is how many of the predictions within the bin are correct (equal to the ground truth). The accuracy of B_m is defined by:

$$acc(B_m) = \frac{1}{|B_m|}\Sigma_{i \in B_m}\mathbf{1}(\hat{y}_i - y_i) \tag{2.1}$$

where $\hat{y}_i$ and y_i are the predicted and ground truth class for input i. The second value calculated for each bin is the average confidence. The average confidence of B_m is defined by:

$$conf(B_m) = \frac{1}{|B_m|}\Sigma_{i \in B_m}\hat{p}_i \tag{2.2}$$

where $\hat{p}_i$ is the confidence of the prediction i. Reliability diagrams plot two separate but superimposed histograms. One histogram plots the accuracy for each bin on the y-axis and the other plots the average confidence of each bin on the y-axis. For both histograms, the x-axis represents the confidence of predictions. Bin B_m contains predictions with confidences within the interval $(\frac{m-1}{M}, \frac{m}{M})$ and as such this bin is represented as a bar positioned between $\frac{m-1}{M}$ and $\frac{m}{M}$ on the x-axis on both histograms. A perfectly calibrated model should have $acc(B_m) = conf(B_m)$ for all $m \in 1, ..., M$. In other words, the height of the bars representing accuracy and confidence should be the same height for each bin.

Expected Calibration Error

Reliability diagrams give an intuitive graphical representation of model calibration. However, a quantitative measure is also useful so that models can be compared numerically. Guo *et al.* [50] propose the Expected Calibration Error (ECE) statistic. It is a weighted average of the difference between the accuracy and confidence of each bin, where the weight is how many predictions fall into each bin. This is formally defined as:

$$ECE = \Sigma_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)| \tag{2.3}$$

where n is the total number of predictions. The ECE statistic can be computed over all sets of landmarks for which there are predictions and ground truth labels.

2.6.2 Temperature Scaling

Guo *et al.* [50] analyses several methods for calibrating neural networks. In their work, they find that a method called Temperature Scaling improves model calibration the most and is relatively simple compared to other methods. Temperature Scaling works by dividing the outputs of a neural network by a parameter T before they are put into the softmax activation function. Mathematically this is defined as:

$$S(x) = \frac{e^{N(x)/T}}{\sum e^{N(x)/T}} \tag{2.4}$$

where $N(x)$ is the value of the output vector N of the network for input x. $T > 0$ is a scalar value. Figure 2.9 shows how Temperature Scaling improved the calibration of their neural network.

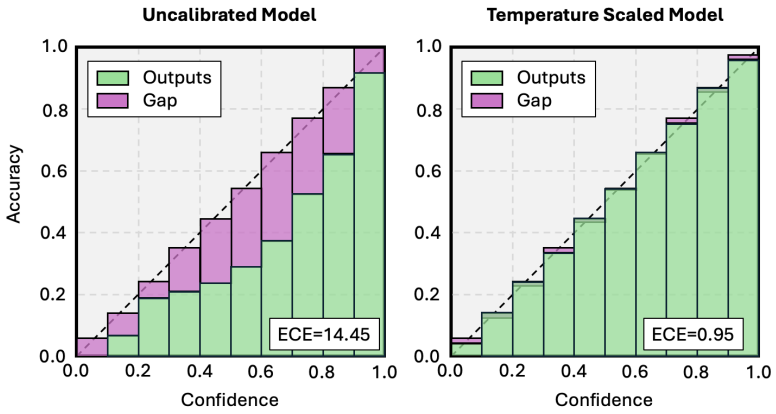
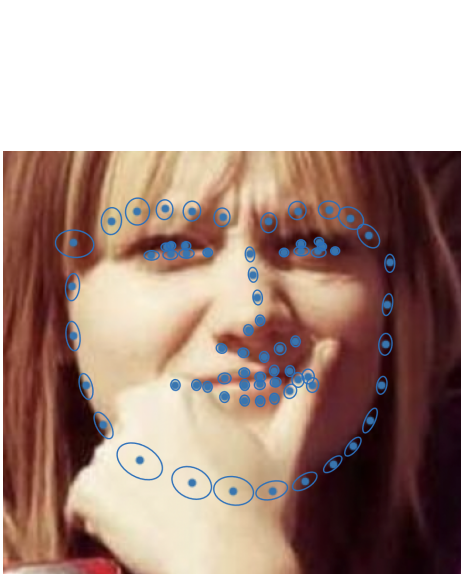
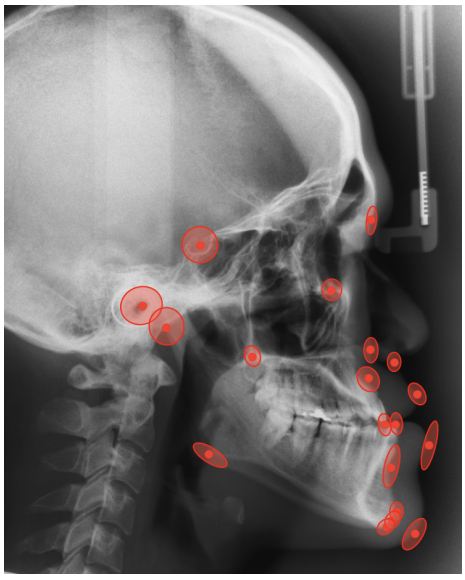


Figure 2.9: Guo *et al.* [50] presented how Temperature Scaling improved the calibration of their network. This figure illustrates this by showing that Temperature Scaling reduces the gap between the superimposed histograms in the reliability diagram and reduces the ECE score.



(a) This figure shows multi-variate gaussian directly regressed along with the predicted points (blue dots) similar to what is presented in Kumar *et al.* [71].



(b) This figure shows an example of multi-variate gaussians which could be derived from the sampling-based method presented in Lee *et al.* [77]

Figure 2.10: Figures showing existing works which either directly (a) or in-directly (b) derive Gaussian distributions for each landmark prediction.

2.6.3 Uncertainty in Landmark Detection

There have been no attempts in the literature to quantify the calibration of landmark detection networks using reliability diagrams or calibrate landmark detection networks using Temperature Scaling. It is therefore yet to be determined whether the confidence of landmark detection networks is equal to the certainty of the prediction. There has, however, been other work in landmark detection uncertainty since 2019. **It is worth noting that most of the works cited in this section have been published concurrently with the work I present in this thesis and were not known to me at the time of carrying out the research.**

In 2019 Gundavarapu *et al.* [49] published the first work on uncertainty in landmark detection, experimenting on the human pose dataset. They proposed a sample-free approach where values representing the standard deviation of a multivariate Gaussian are regressed alongside a landmark position and Gaussian log-likelihood is used as a loss function. In this paper, they show that landmarks with the highest uncertainty are likely to be occluded landmarks. Kumar *et al.* [71] built on this approach by integrating similar outputs into a stacked hourglass network [96]. They performed thorough uncertainty validation by creating their own annotation set with labelled occluded landmarks and then carried out correlation experiments between uncertainties and localization errors. This method was tested on facial landmark data, an example of the output, with multivariate Gaussians, is given in Figure 2.10a. The multivariate Gaussians are displayed in the figure using ellipses. These ellipses are centred on the landmark prediction and their size represents the uncertainty of landmark placement.

Finally, in the non-medical space, Chen *et al.* [18] proposes the kernel density network to extract uncertainty estimates. This has the advantage over other methods in that it does not limit the output heatmaps to Gaussian distributions. This is achieved by changing the typical loss function by embedding a kernel function within it. The disadvantage of this method is that a kernel function has to be specified in advance of training, and it is unclear which kernel is optimal. In

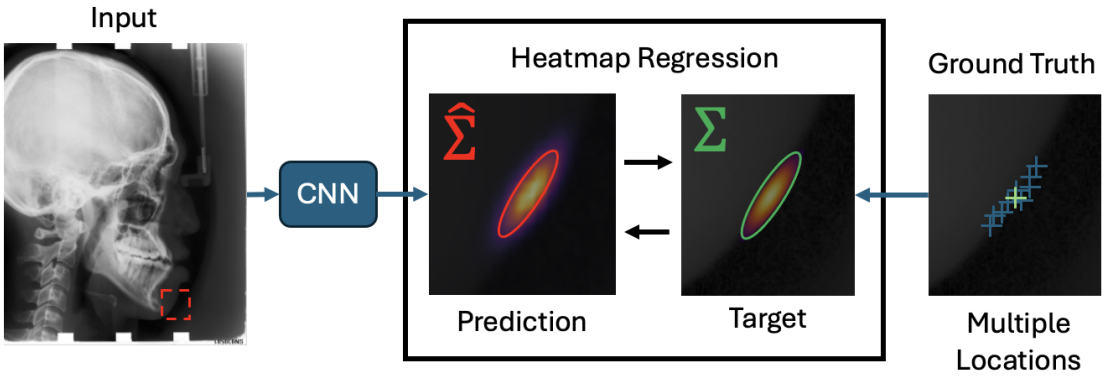


Figure 2.11: This figure illustrates the approach taken by Payer *et al.* [105]. The Gaussian training heatmaps used to train the model are represented by the values in the green Σ covariance matrix which are learned throughout training. Additionally, a multivariate Gaussian is approximated from the output heatmap represented by the red Σ covariance matrix and red ellipse. The results in this publication show that the values in the green Σ correlate with aleatoric uncertainty and the values in the red σ correlate with epistemic uncertainty for this task.

addition, although the uncertainties obtained can be used to deduce occluded landmarks, they are not validated further.

Although there are relatively few publications in the area of landmark detection uncertainty, of the published works, a high proportion are on medical images. These cite that understanding uncertainty in the medical space is of particular importance. The first work in this space with an application in the medical imaging domain was Lee *et al.* [76] who used a Bayesian CNN (BCNN) to localize landmarks. A BCNN outputs multivariate Gaussian distributions centred at the landmark prediction point, an example is shown in Figure 2.10b. The size of the ellipses in the figure corresponds to each landmark prediction uncertainty. A disadvantage of this approach is that BCNNs work by performing multiple passes for single images at test time. In addition to this, their model was patch-based and required that all pixels in a RoI be passed into the second stage BCNN for each pass. Passing an image through their network at test time took 192 seconds on a GPU, which is not scalable. To fix this problem in 2021 Payer *et al.* [105] built on their previous work [104] by making the training heatmaps multivariate Gaussians represented by learnable parameters. They also fitted multivariate Gaussians to the output

heatmaps. This is shown in Figure 2.11 where a red ellipse is overlayed on a prediction heatmap. The ellipse is centred at the landmark prediction, which is the expected value of the heatmap, and its size is represented by a covariance matrix. The expected value and covariance matrix are computed by interpreting the output heatmap as a discrete probability distribution across the domain of pixels and then calculating the expected value and variance in the x and y directions of the distribution. This work shows that there is a correlation between aleatoric uncertainty and the standard deviation parameters on the training heatmaps and that there is a correlation between the epistemic uncertainty and the standard deviation of the output heatmaps. In late 2021 Thaler *et al.* (Payer 2nd author) [128] future validated this concept by obtaining annotations from nine more clinicians on a subset of their dataset to show that the inter-rater variability between these clinicians correlates with the uncertainty of their model.

In a separate line of research Schobs *et al.* [114] analyse whether the highest value in an output heatmap correlates with uncertainty. They conduct experiments on a cardiac MRI dataset and find that putting all landmarks above a certain value into one bin and below into another bin, the former bin, on average, contains significantly more accurately located results. Schobs *et al.* [115] recently built on this work in 2022 by trying different confidence pseudo-measures including one where the distance between predictions made by three separate models was used as an uncertainty measurement. They find that the best pseudo-measure is taking the heatmaps from each model in the ensemble, multiplying them together, and taking the value of the hottest point of the product heatmap. This is validated with the same binning approach as their first work.

2.6.4 Gaps in Current Literature

Retraining Required

Most existing methods, including the three non-medical landmark detection uncertainty publications [18, 49, 71], proposed new networks be trained to obtain uncertainty measurements. Of the medical approaches: Lee *et al.* [77] uses a BCNN, Payer and Thaler [105, 128] train networks on multivariate Gaussian heatmaps where the parameters are learnt during training. Researchers and companies could already have their own trained models and do not want to re-code and train a new method to obtain uncertainty measurements. Ideally, it should be possible to add a module to existing, trained, networks to obtain an uncertainty measurement without the need for training a new model with a new scheme. The only method for which this is the case is Schobs *et al.* [114, 115].

Assuming Gaussian Distribution

The majority of approaches assume that the input and output heatmaps should be Gaussian distributed despite it being accepted in recent literature that training on Gaussian heatmaps is not representative of the true landmark position distribution [128]. Current approaches have not posed an alternative. In addition, current approaches either regress multivariate Gaussian distribution parameters to represent uncertainty as shown in Figure 2.10, or fit a Gaussian distribution to the output heatmap, as shown in Figure 2.11. Although ultimately the output heatmap must be represented by a small set of values for quantitative analysis, assuming the distribution is Gaussian is likely lose useful information contained in the heatmap. Schobs *et al.* [115] also does not assume a Gaussian distribution however their method only uses the highest value in the output heatmaps thus useful information is also lost from that process, for example, the size of the heatmap and whether it is multi-modal.

Limited Heatmap Analysis

No analysis has been conducted in any of the papers regarding how well calibrated the values in the outputs are to the true probability of a clinician placing the landmark there. A potential method for doing this is through reliability diagrams. Current methods take statistics such as the Gaussian covariance matrix and find a correlation between them and the accuracy of the landmarks [49, 71, 105, 128].

Lack of Practical Analysis

Little analysis has been completed into how individual patients could be affected by a system with/without uncertainty embedded. The most recent works by Thaler *et al.* [128] and Schobs *et al.* [115] do attempt work in this direction. However, both have their flaws. Thaler *et al.* show a correlation between their uncertainty statistic and individual measurements in a figure in the publication, however, no quantitative evaluation is performed. It is hard to see whether a threshold could be applied to their uncertainty statistic to warn the doctor that a likely inaccurate measurement has been taken and how effective that threshold would be. Schobs *et al.* show that a threshold could be placed over their uncertainty measurement but not what knock-on effect that would have on patients. Neither of these approaches explores whether a customisable threshold could be proposed to satisfy specific end-user requirements.

3 Datasets

Contents

3.1 Introduction 40
3.1.1 Dataset Types 40
3.1.2 Ethical Considerations 43
3.2 Cephalometric Dataset 44
3.3 Hand Dataset 49
3.4 Pelvis Dataset 53
3.4.1 Femoroacetabular Impingement 53
3.4.2 LCE-angle and α-angle 54
3.5 Infant Hip Dataset 59
3.5.1 Developmental Dysplasia of the Hip 59
3.5.2 α-angle and β-angle 60
3.6 Dataset Characteristics 64
3.6.1 Landmark Characteristics 64

3.1 Introduction

As identified in Section 2.5.3, a current gap in the literature is that published methods are not tested on multiple datasets. This thesis presents methods which generalise to a wide range of medical images and, to this end, four medical imaging datasets were used for experimentation. The datasets used for the experiments described in this thesis contain images of Cephalometric (head) X-rays, Hand X-rays, Pelvis X-rays and Infant Hip ultrasounds. The Cephalometric and Hand datasets are publicly available and the Pelvis and Infant Hip datasets were privately sourced via clinical collaborations with our research group. In the first section of this chapter, the motivation for using, and the broad differences between, publicly available datasets and private datasets is provided to give an overall context. Then an overview of each dataset with descriptions of the images, landmarks and any angles or diagnoses which can be computed from the landmarks is provided. These datasets are distinct in the anatomy captured and in the landmark placement characteristics which ensures appropriate benchmarking for testing the methods presented in this thesis.

3.1.1 Dataset Types

Publicly Available Datasets

Distinguishing datasets used to increase the likelihood of research publication in Computer Science from those essential for advancing the adoption of methods in clinical practice is important. Publicly available datasets are chosen when the goal is to publish at leading Computer Science conferences because the assessment criteria for these conferences is to show that the proposed method outperforms previous methods on a dataset. Although reproducing previous methods and then applying them to a new dataset is theoretically possible, in practice it is very challenging because many publications either do not their share code or omit the precise details of their proposed method. It is therefore easier and quicker to use existing datasets and compare against previous method's published results when the goal is to publish. The Cephalometric and Hand datasets described in this thesis were chosen for this

reason and the former was used for the experiments presented in my paper *Contour Hugging Heatmaps for Landmark Detection* presented at CVPR 2022 [88].

These datasets, and publicly available datasets in general, are often released to the public in the form of a challenge presented at a conference, as discussed in Section 2.2.2. The dataset for a challenge is made available via online download, such as through the popular Grand Challenge website.[1] Challenges incentivise researchers to work on the dataset because of the promise of publication for the highest-achieving methods. Publicly accessible datasets generally feature fewer images and include images with more conventional presentations compared to images captured during routine clinical practice.

Private Clinical Datasets

Although publicly available datasets provide a good benchmark for new methods to outperform, their effectiveness at pushing forward automatic AI approaches into clinics is limited. This is because achieving good automatic results on a dataset is only one of many stages involved in translating technology into the clinic. A suitable dataset must first be obtained with the help of clinical collaborators and any regulatory barriers stopping data obtained in hospitals from being used for research purposes must be passed. This process can take years because it requires in-depth knowledge of the protocols in place and good connections to the relevant clinicians, administrators and technicians in the hospitals where the images are being acquired.

An example where this has been achieved successfully is Murphy *et al.* [95] who conducted experiments on two datasets containing 429 and 2,364 routine images where images in the latter dataset contained hip fractures. This publication showed that an automated CNN approach achieved better diagnositic perform than any single clinician. An additional example is Beetz *et al.* [12] who experiment with a dataset of 1021 Magnetic Resonance images taken in a multi-center study to assess how accurately the shape of the heart can be reconstructed in 3D.

[1] `https://grand-challenge.org/challenges/`

In order for a dataset to push forward translation of AI technology into clinics it must be different from publicly available datasets in several ways. One such way is that more metadata should be included with each image such as age and ethnicity to ensure that the algorithm developed is not biased. In addition, more effort needs to be made to ensure that the dataset represents images taken during routine practice in different hospitals so that the model trained from the data can be confidently deployed in different sites. Finally, the experiment objective needs to be different. The objective of publicly available datasets is to provide a benchmark for future publications whereas private datasets should be used to validate how the technology would improve the toolkit clinicians currently have access to and what the effect it would have on both clinican time and patient care.

Data collected via private clinical collaboration often cannot be distributed to the public. This means Computer Science researchers often have to choose between two distinct paths at any one time; either to work towards a publication at a highly rated Computer Science conference which requires a publicly available dataset or perform translational work which might not get published at the top conferences in Computer Science. For this reason it is also likely that other research groups have been performing machine learning experiments on their private datasets but have not been published their results.

Both the Pelvis and Infant Hip datasets described in this thesis were obtained by clinical collaborations and shared with the research group I belong to. When the research described in this thesis was being performed only the a preliminary subset of the overall Infant Hip dataset was available for experimentation. A larger version of this dataset has now been obtained containing images taken during routine practice.

3.1.2 Ethical Considerations

When working with medical data it is important to consider a range of ethical considerations to ensure the protection of individuals and the responsible use of their data. All datasets in this thesis adhere to the following principals:

- Privacy and Confidentiality - All images in these datasets were anonymised by removing names and addresses from their metadata as well as any identifying information captured in the scan itself. This means that it is not possible for any researcher or collaborator to trace back the images to the individuals captured in them.

- Data Security - Data was stored on Oxford University's Advanced Research Computing (ARC) service when performing experiments and on external hard drives kept within the research group. Both of these locations are password protected and only members of our research would have access to them.

- Data Minimization - Our group only accessed the minimal amount of data required for each experiment. We did not obtain any extra images or metadata of patients that were not necessary for the experiments to be performed.

- Non-Discrimination - We ensured that the use of medical data does not perpetuate or exacerbate existing health disparities or lead to discrimination against any particular group. Whilst we did not have control over which patients and demographics were captured in the images of the datasets used we did check what demographics were represented in each. Additionally, in Chapter 4 a method is proposed to ensure future models are trained on representive data so they are not biased towards particular demographics.

3.2 Cephalometric Dataset

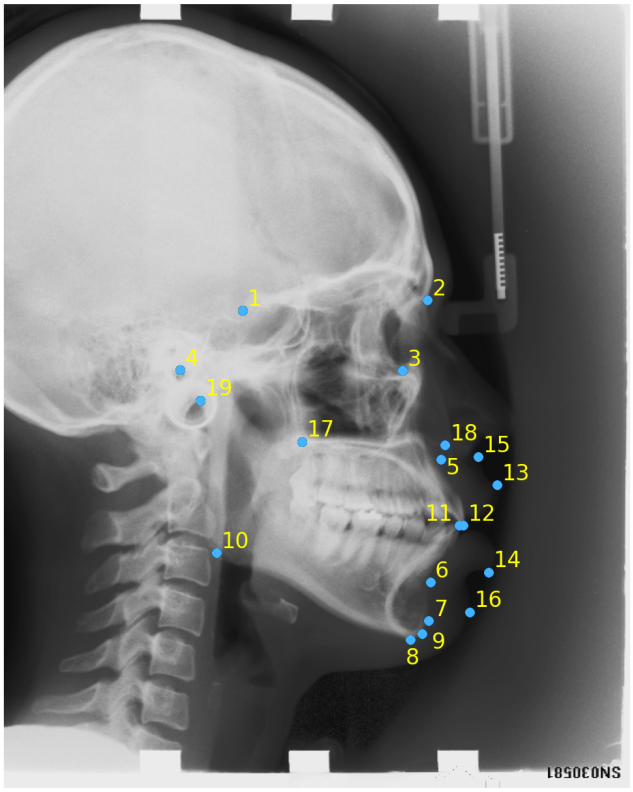


Figure 3.1: This figure shows the 19 Cephalometric landmarks annotated on each image.

Dataset Details

The Cephalometric dataset comprises X-ray images of the head, it is publicly available, and has been widely used by researchers to test landmark detection methods since it was released for public research at the IEEE ISBI conferences in 2015 [133] and 2016 [134]. It is also a challenge dataset at the MICCAI 2023 conference. The dataset was collected for research purposes related to orthodontic analysis which involves evaluating the angles between specific landmarks on the head. Consequently, the dataset was published with annotations for these landmarks on the images.

The dataset comprises 400 images, split into a training set of 150, plus Test Set 1 and Test Set 2 of 150 and 100 images respectively. The images were released as `Windoms BMP image` files with the file extension `.bmp`. The images have a greyscale bit-depth of eight bits, meaning there are $2^8 = 256$ different greyscale intensities,

	Intra-Observer Variability		Inter-Observer Variability
	Expert 1	**Expert 2**	**Expert 1 vs Expert 2**
MRE (mm)	1.73±1.35	0.90±0.89	1.38±1.55

Table 3.1: Intra and inter-rater variability mean and standard deviation for two experts, each annotating twice in millimetres (mm) for the original Cephalometric dataset [133]. The statistics for Experts 1 and 2 are calculated by finding the mean and standard deviation between the measurements they each took on the two separate occasions they assessed the images. The statistics for Expert 1 vs. Expert 2 are calculated by first averaging the two measurements each expert took and then calculating the mean and standard deviation of the difference between the experts' measurements.

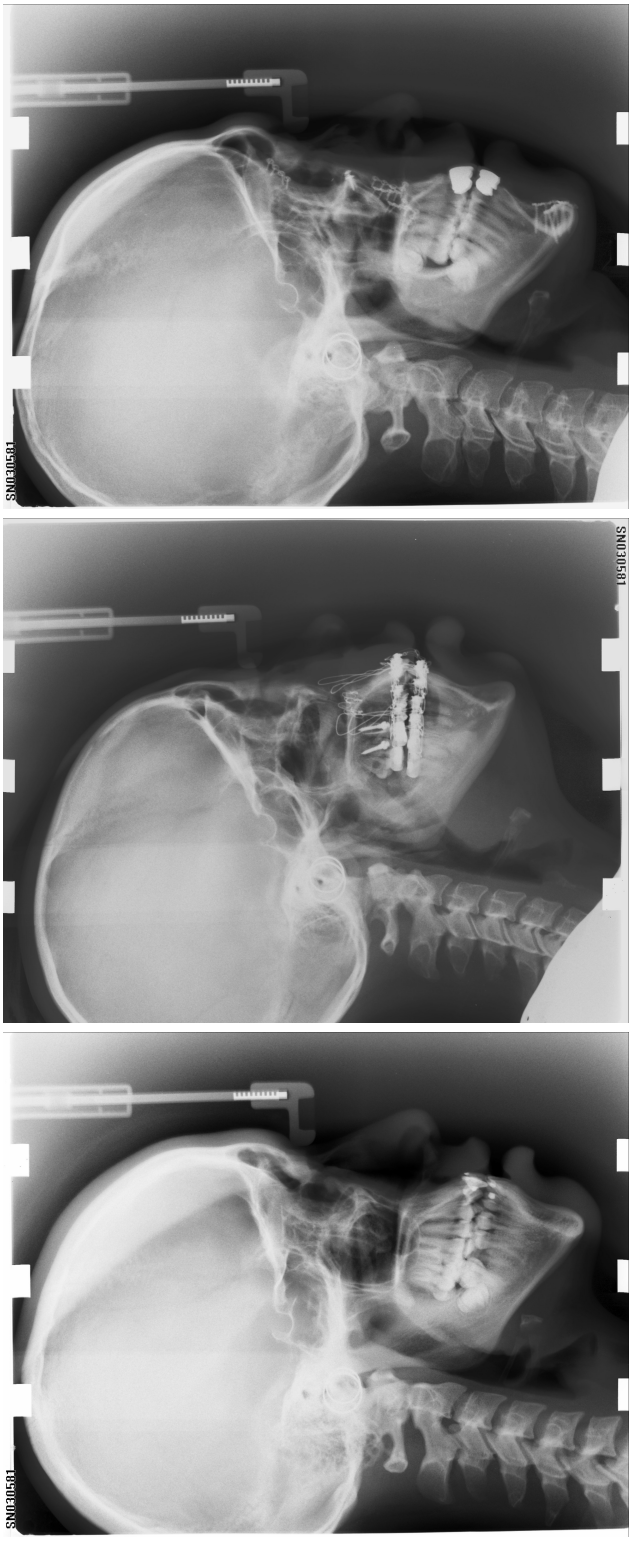
from pure black to pure white, available to each pixel. All images were taken using the same scanner with patients being between six and sixty years old. No curation, exclusion criteria or preprocessing steps are stated in the original paper which released the dataset [133]. The resolution of each image is 1935×2400 pixels. All past publications that experimented with the dataset have stated that they interpreted each pixel as a 0.1mm x 0.1mm squared area, including publications by Linder *et al.* [82] and Chen *et al.* [22]. However it is worth noting that each pixel must represent a larger area than 0.1mm x 0.1mm because 2400 pixel multiplied by 0.1mm is only 24cm which is significantly shorter than the average height of the head and neck of a person. Nevertheless this pixel spacing was also used for the experiments described in this thesis so that the methods presented could be directly compared to previous literature. This dataset is the only dataset experimented on where the images are all the same resolution and, in addition, the only dataset which has been annotated by multiple clinicians. Two orthodontists, with six and 15 years of experience respectively, labelled 19 landmarks on each image twice and the final landmark positions are the average of these four ground truths. The software used to annotate the images is not stated in the original publication. These landmarks are shown in Figure 3.1. The availability of this data means intra and inter-rater variability for the landmarks are known, as seen in Table 3.1. The inter-operator variability of 1.38mm (which is explained in the caption of Table 3.1) gives automated approaches an accuracy target to achieve for its measurements to be considered in the range of clinicians.

Dataset Discussion

Images in this dataset are well aligned, with each head being positioned using a guide visible in all images. All images contain the entire head vertically, and they slightly crop the back of the skull. More example images from the dataset are given in Figure A.1 in Appendix A. Landmarks are always located in the middle to the bottom right quadrant of the image, although how close the chin is to the bottom of the image varies.

There are considerable differences between the presentation of patients, with some having extensive implants around the face, as shown by Figure 3.2b and 3.2c. The placement of the landmarks on Cephalograms is also ambiguous as shown by the inter and intra-operator variability statistics in Table 3.1. Some landmarks are less controversial than others, for example, Thaler *et al.* [128] show that there is less variability in the human placement of landmarks eleven and twelve because the anatomical point where they lie is unambiguous.

The landmarks with most variability are located on an edge such as Landmark 16 (see Figure 3.1). On top of this, there has been a systematic shift in how this landmark has been placed by the annotators between the training set, Test Set 1, and Test Set 2 [82, 148]. In the former two subsets, the two annotators disagree with each other on the placement of the landmark by a consistent 5mm. However, in Test Set 2 they consistently agree on its placement [105] with them both placing this landmark in a lower position. As the final landmark position is the average of the annotators' placements this means there is a significant difference between Landmark 16's position in Test Set 2 compared to the training set and Test Set 1. **This discrepancy is the reason why all past methods achieve superior results on Test Set 1 compared to Test Set 2**. The original paper does state that there are two ways of positioning Landmark 16, as shown in Figure 3.3, so it is possible one annotator changed their annotating procedure from one method to the other before annotating Test Set 2.



(a) A standard example from the dataset containing no abnormalities. **(b)** An example where the patient has dental implants. **(c)** An example containing foreign objects in all areas of the face.

Figure 3.2: Three images from the Cephalometric dataset.

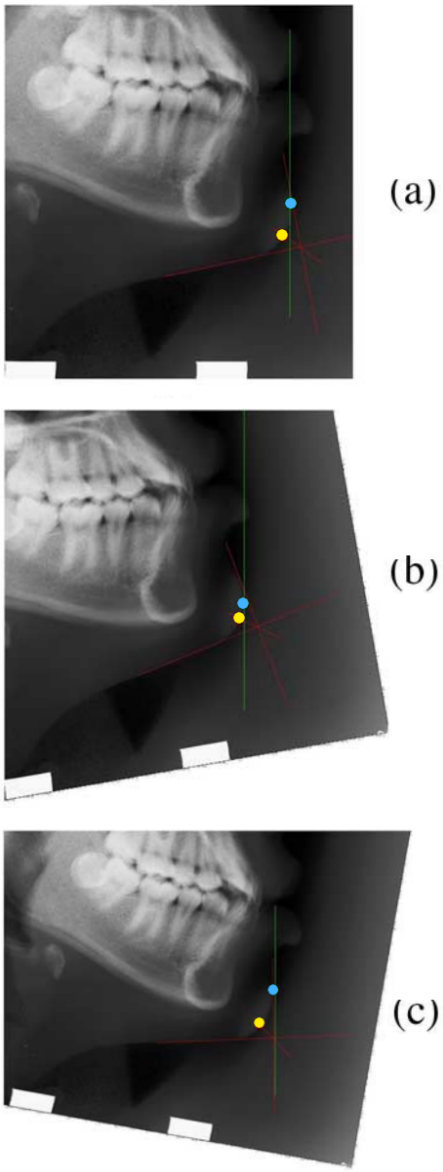


Figure 3.3: These figures taken from Wang *et al.* [133] show two methods for placing Landmark 16. The older method is represented by the blue dot and the newer method by the yellow dot in a, b and c. The blue dot is placed at the point where a vertical line intersects the chin. This causes the blue dot to be placed high up the chin and its placement is also dependent on the rotation of the head. On the other hand, the yellow dot is placed using multiple intersecting lines and is consistently lower than the blue dot as well as being less dependent on the rotation of the head. One annotator likely changed their method between annotating the Training and Test Set 1 and the Test Set 2.

3.3 Hand Dataset

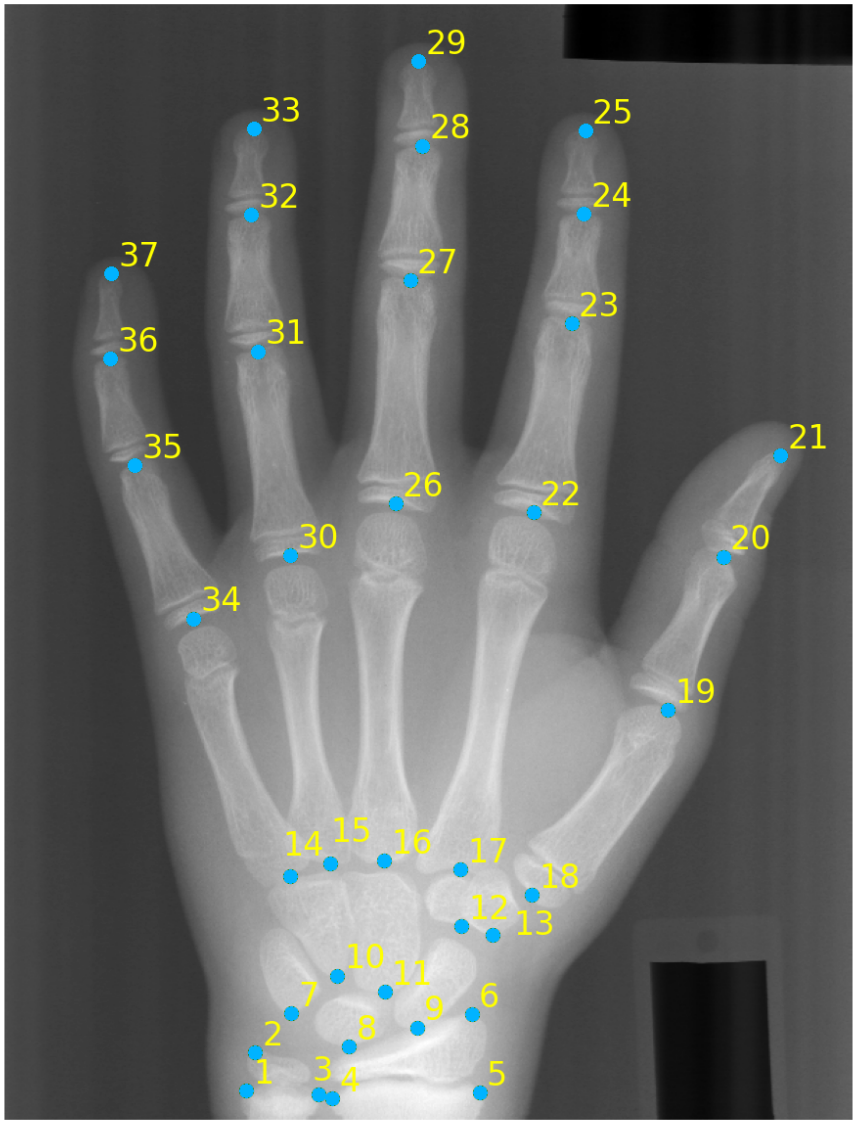
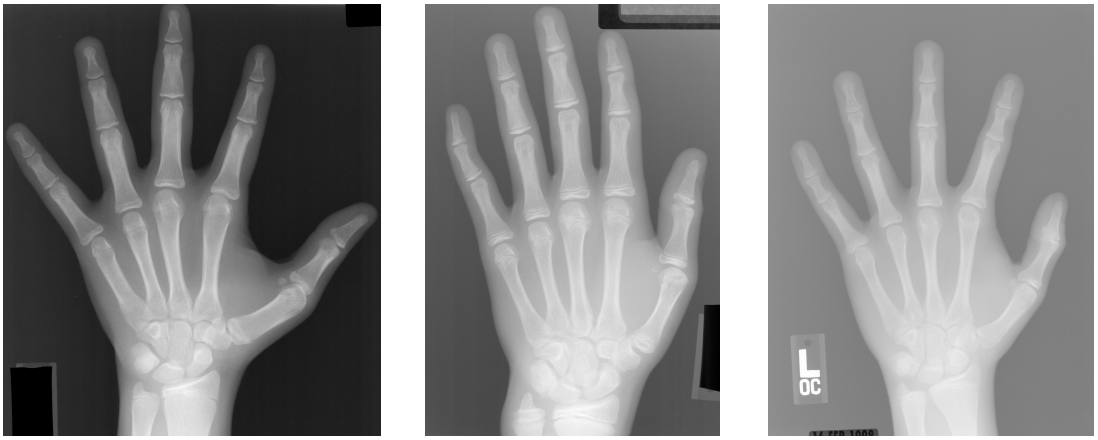


Figure 3.4: This figure shows the 37 Hand landmarks annotated for each image in the dataset.

Dataset Details

The Hand dataset was introduced by Payer *et al.* [103] in 2016 after they annotated 895 images from the publicly available Digital Hand Atlas [43]. The Digital Hand Atlas contains X-rays of children's left hands between the ages of one and eighteen. No exclusion criteria or preprocessing steps are stated in the original paper which



(a) Digits are splayed and background is dark.

(b) Digits are tilted to the left.

(c) Hand is small and image is bright.

Figure 3.5: Three images from the Hand dataset.

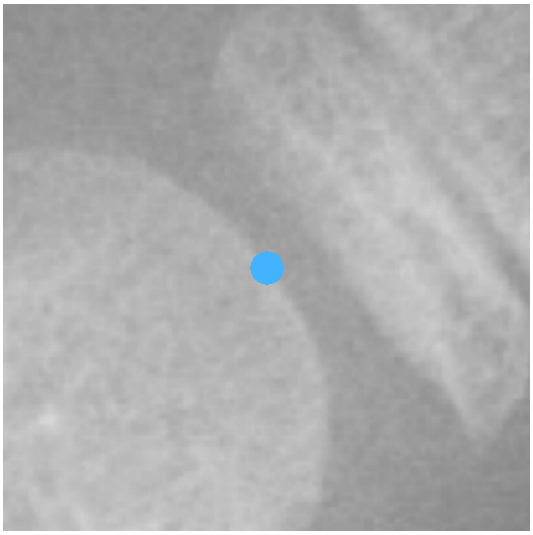
released the dataset [43]. It is stated that an equal number of images were obtained from Caucasian, African American, Hispanic and Asian ethnic groups and that the images were taken at different sites across the Los Angeles area. Originally the dataset was acquired to develop methods that could predict a child's age from the wrist bone but Payer repurposed it for landmark detection by annotating 37 landmarks on fingertips and bone joints, shown in Figure 3.4. The landmarks were manually annotated on each image once and minimal information is provided about the annotator in Payer *et al.* [103] except for describing them as an 'expert'. The software used to annotate the images is not stated in the original publication. The images are in `.jpg` format and have an average resolution of 1560×2167 pixels although these dimensions can vary by up to 200 pixels for some images. The images have a greyscale bit-depth of eight bits which is standard for the `.jpg` format. In line with previous publications, in the absence of metadata, we calculate the pixel size by assuming that the wrist diameter of each hand captured is 50mm. Three different training and test set combinations are given by the authors and, as in previous work [81, 104, 131], experiments in this thesis are performed on all three of these combinations and the average results reported.

Dataset Discussion

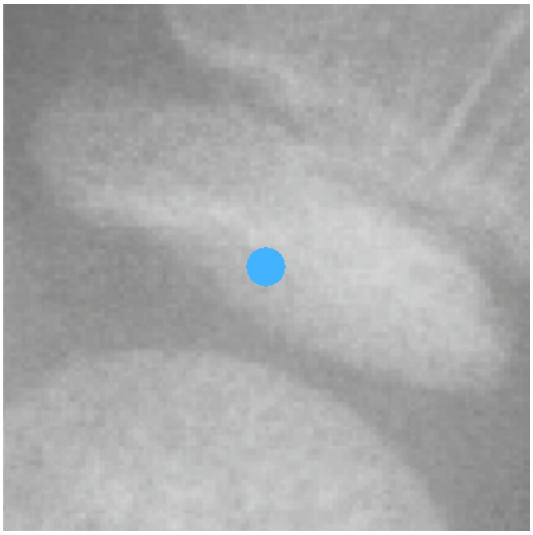
In all images, the palm of the hand appears roughly in the centre to the bottom of the image with all five digits captured. There is little rotation of the hand, the most tilt is shown in Figure 3.5b. Fingers can be close together like in Figure 3.5b or spread out like in Figure 3.5a but in all cases the fingers are fully extended. In most images, there are dark patches as seen in Figure 3.5b and 3.5a. Although not stated, these patches could be to obscure patient-identifying details in the image. Some images have a big letter L present like in Figure 3.5c. One difference between this dataset and Cephalometric is that the hands were taken with different scanners and settings, hence a difference in image brightness and contrast can be observed. Figure 3.5a shows a hand with a dark background and high contrast as opposed to Figure 3.5c which is a light background and low contrast. More example images from the dataset are given in Figure A.2 in Appendix A.

Another difference is the fundamental structure of the hand. Many past publications which experiment on the Cephalometric data use a patch-based approach [8, 145]; however none of these methods are validated on the Hand dataset. This is likely because it would be very difficult to know which part of a hand a patch comes from if only a single digit is captured in the patch.

This dataset only has one set of annotations so calculating inter and intra-observer statistics is not possible. However, Payer *et al.* observed that Landmark 19 is the hardest to localise because it can be placed at either the superior point on the lower bone, as in Figure 3.6a or at the inferior point on the upper bone, as in Figure 3.6b.



(a) Figure from Thaler *et al.* [128] shows the landmark placed at the superior point on the lower bone.



(b) Figure from Payer *et al.* [105] shows the landmark placed at the inferior point on the upper bone.

Figure 3.6: Two different ground truth placements of Landmark 19.

3.4 Pelvis Dataset

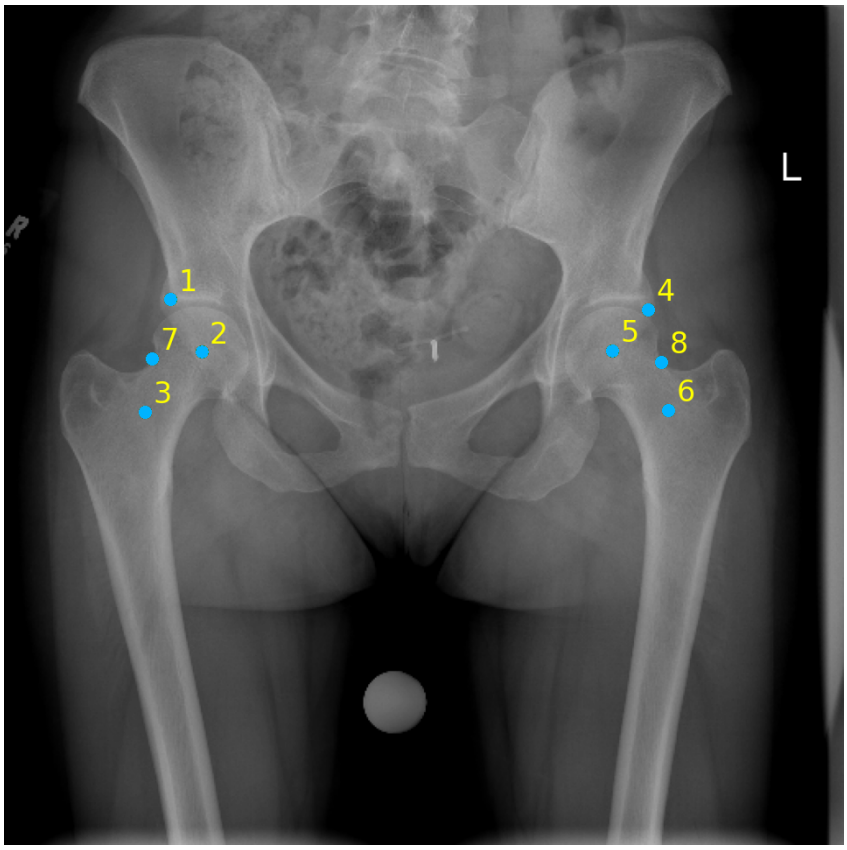
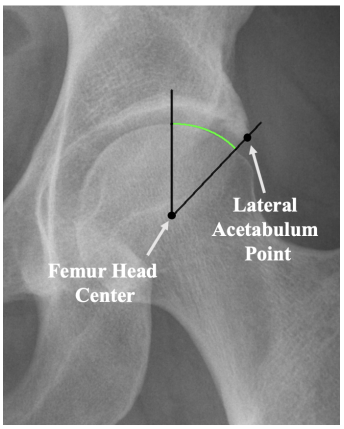


Figure 3.7: This figure shows the eight X-ray pelvis landmarks annotated for each image in the dataset.

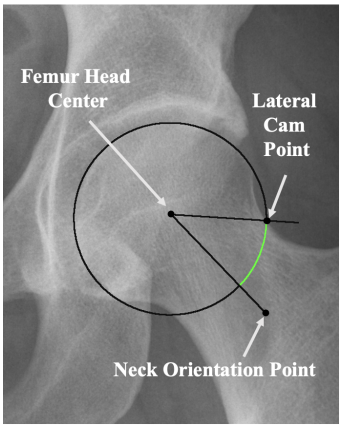
This dataset contains X-rays of the pelvis and was collected at hospitals around Oxford as part of the Femoroacetabular Impingement Trial (FAIT) [102]. It is not publicly available. This trial was designed to determine the best treatment for Femoroacetabular Impingement (FAI) and the images began being collected in 2014.

3.4.1 Femoroacetabular Impingement

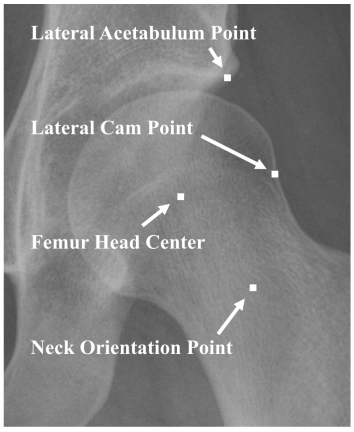
FAI is a condition where the ball of the femur bone and the socket of the pelvis, called the acetabulum, rub together during movement [41]. There are two types of FAI: cam impingement, where the femoral head has an abnormal 'bump' present, and pincer impingement where extra bone extends further than the normal rim of the acetabulum potentially damaging a piece of cartilage called the labrum inside



(a) This figure shows an LCE-angle and the detected landmarks used to calculate it.



(b) This figure shows an α-angle and the detected landmarks used to calculate it.



(c) This figure shows all landmarks detected for each hip.

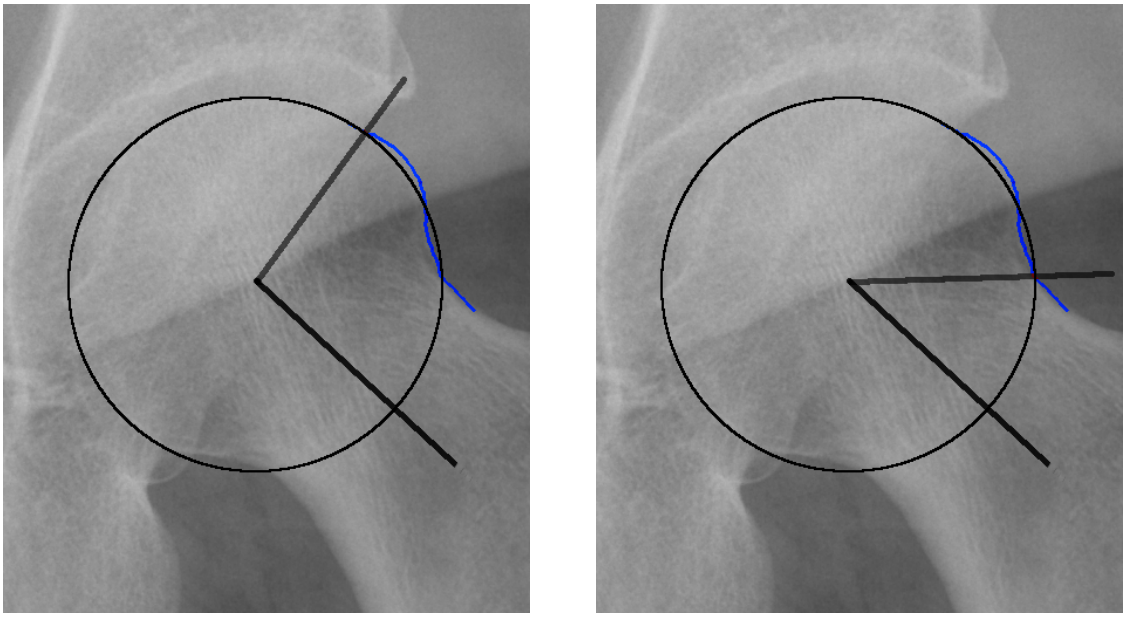
Figure 3.8: Figures taken from my work *Automatically Diagnosing Hip Conditions from X-rays using Landmark Detection* [89] show the LCE and the α-angle. Figure (c) shows what landmarks are detected for each hip. Figure (a) shows that the LCE-angle is the angle between the line segment drawn from the femur head centre to the neck orientation point and the line segment from the femur head centre to the lateral acetabulum point. Figure (b) shows that the α-angle is the angle between the line segment drawn from the femur head centre to the neck orientation point and the line segment from the femur head centre to the lateral cam point.

the hip socket. Both of these types are likely to cause groin pain and the inability to perform high hip flexion activities or sit for prolonged periods of time [11]. FAI can be treated either by physiotherapy or hip arthroscopy. Hip arthroscopy is a type of keyhole surgery where the clinician uses mechanical tools to repair the cartilage inside the hip socket of a patient.

3.4.2 LCE-angle and α-angle

Both cam and pincer impingement are diagnosed by finding angles on the X-ray of a patient. Pincer impingement is diagnosed by measuring the Lateral Center Edge (LCE) angle between the vertical axis and the lateral edge of the acetabulum [113] as shown in Figure 3.8a. An LCE-angle threshold of $\geq 40°$ is used to diagnose pincer impingement in this thesis in line with previous literature [72].

Cam impingement is diagnosed by measuring what is called the α-angle, the angle between the axis of the femoral neck and a point where the femoral head stops being spherical, as shown in Figure 3.8b and Figure 3.9a. α-angle thresholds of $\geq 70°$



(a) An α-angle placement that matches the definition given how this circle is placed.

(b) A more conservative non-geometrically accurate α-angle reading.

Figure 3.9: This figure shows two different interpretations and placements of the lateral cam point with the blue contour following the edge of the femur. This example illustrates how different α-angle readings can be taken from the same hip.

for men and $\geq 61°$ for women have been proposed to diagnose a cam impingement [39]. In this thesis, an average of these ($\geq 65°$) is used to diagnose cam impingement.

Agricola *et al.* [4] observe a 0.73 intraclass correlation coefficient (ICC) between clinicians taking the α-angle measurement (see Section 5.1.3 for a definition of ICC). This indicates that there is regular disagreement between clinicians when taking this measurement. Often clinicians do not place the angle according to the geometric definition and, instead, place the landmark where the contour leaves the circle for the second time near the femoral neck, as in Figure 3.9b. This may not be a mistake: where the circle is plotted is subjective because the femoral head is not perfectly spherical and also because bumps on the femoral head can be very small. In addition, clinicians use their experience and intuition when making diagnoses. They may choose an α-angle which matches the condition they believe (for other clinical reasons) the patient has as opposed to measuring the α-angle without prior knowledge.

Dataset Details

Every patient captured in this dataset is between the ages of eighteen and sixty and presented with evidence of the condition. The dataset contains 375 AP X-rays of the pelvis. Patients with previous hip surgeries, advanced osteoarthritis, hip dysplasia or who had undergone physiotherapy were filtered out of the trial before any experimentation was performed on the dataset [102]. This means that no image in the dataset contains implants. This dataset contains longitudinal data with up to five images per patient captured at baseline, six months, 12 months, 24 months and 36 months. The images are in `.dicom` format which contains the pixel-to-mm resolution. This differs from image to image but the average resolution is one pixel representing a 0.134mm $\times$ 0.134mm squared area. These images were exported from `.dicom` to `.png` format so that they could be opened in `python` script for machine learning. The `.png` images have an average resolution of 2961 $\times$ 2794 pixels and a greyscale bit-depth of eight bits.

Each image has four landmarks for each hip for a total of eight landmarks as shown in Figure 3.7. Landmarks were manually annotated on each image once. A senior consultant in Orthopaedics annotated 100 images and then subsequently a technically trained annotator annotated a further 275 images under the consultant's supervision. Annotations were produced using the GIMP image editing software.[2] Each image was opened and coloured dots were placed by the annotator for each landmark. These dots were then converted to image coordinates by a custom python script to produce the final ground truth landmark positions.

The four landmarks for each hip are called the 'femur head centre', 'neck orientation point', 'lateral acetabulum point' and 'lateral cam point' as shown in Figure 3.8c. These landmarks have the following definitions: The femoral head centre is placed at the perceived centre of the femoral head geometrically. As the femoral head is not perfectly spherical the centre does not have a precise locus. The neck orientation point is placed in the centre of the shortest line which can be drawn across the femoral neck. The lateral acetabulum point is placed at the

[2] `https://www.gimp.org/`

most lateral point of the acetabulum rim. The lateral cam point is placed at the point where the femoral head stops being circular moving from the socket towards the lateral side of the femoral head.

The task of automatically calculating the LCE-angle and α-angle from images has not been formulated as a landmark detection task in previous literature and as such these landmark definitions are original to this work. However, this does not mean they are without clinical basis. These landmark definitions are derived from the definition of LCE-angle and α-angle in Ruil *et al.* [113] and Fraitzl *et al.* [39] described in Section 3.4.2. The LCE-angle is the angle between the line segment drawn from the femur head centre to the neck orientation point and the line segment from the femur head centre to the lateral acetabulum point. The α-angle is the angle between the line segment drawn from the femur head centre to the neck orientation point and the line segment from the femur head centre to the lateral cam point (see Figure 3.8).

Dataset Discussion

The images in this dataset always contain both hips in full view. Figure 3.10a and Figure 3.10b shows two images from the dataset. Images were taken at different sites and a little difference in the brightness and contrast can be seen between images. In addition, Figure 3.10a shows a male pelvis which is thinner and has a longer pelvis and femurs whereas Figure 3.10b shows a female pelvis. Every image contains a small ball between the legs of the patient which is used for scale calibration and many images also contain letters such as 'L' or 'R' or 'FAI'. More example images from the dataset are given in Figure A.3 in Appendix A.



(a) A male pelvis with non-impinged hips.



(b) A female pelvis where both hips are arthritic and have high α-angle.

Figure 3.10: Images from the Pelvis dataset.

3.5 Infant Hip Dataset

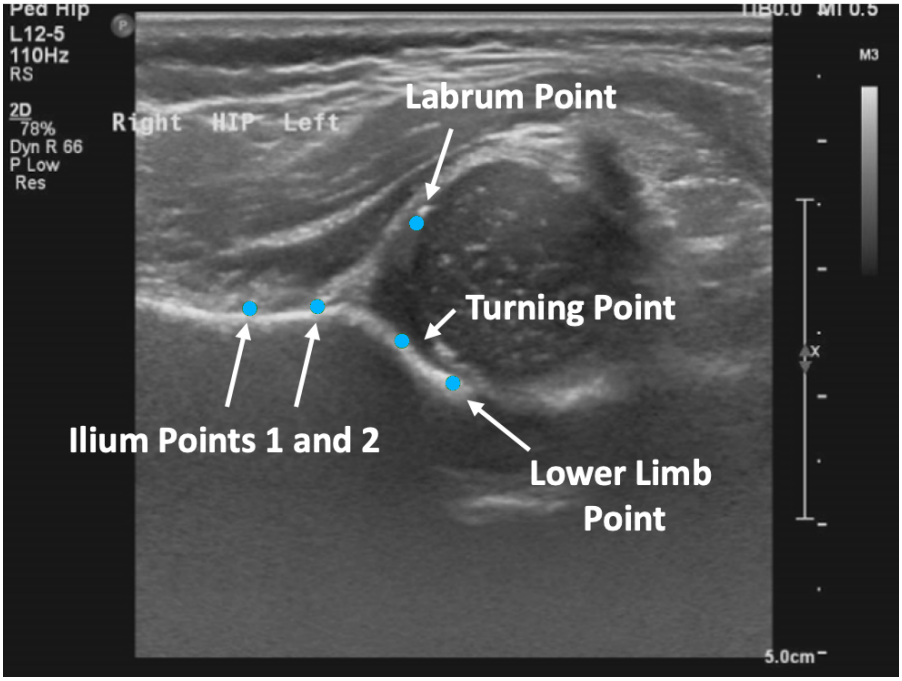


Figure 3.11: This figure shows the five ultrasound hip landmarks annotated for each image in the dataset. The names of the landmarks are given in this figure.

This dataset captures single images of ultrasound scans of babies' hips and was obtained via a collaboration with a clinician at the Alder Hey Children's Hospital in Liverpool. This dataset was collected to build an automatic method for diagnosing a condition called Developmental Dysplasia of the Hip (DDH). It was released for the first time to us so we could perform the research presented in this thesis. This is different from the Pelvis dataset which was more mature and had been used for other research by different groups.

3.5.1 Developmental Dysplasia of the Hip

DDH is where the ball of the hip joint has formed either partially or completely outside the socket. It is especially important to detect DDH early because the earlier it is detected the quicker and more painlessly it can be treated, requiring fewer surgeries and a reduced cost burden on the NHS [32, 127]. For this reason, the images in this dataset are ultrasound images which can be taken of children under three months old as opposed to X-rays.

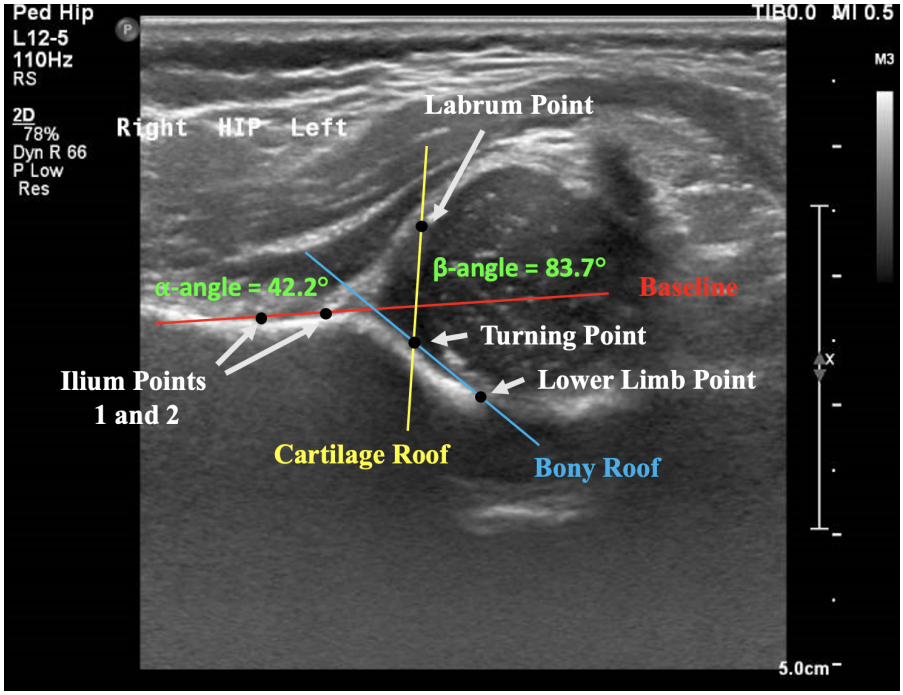


Figure 3.12: The α-angle is the angle between the baseline, drawn using the ilium points, and the bony roof, drawn using the turning and lower limb points. The β-angle is the angle between the baseline and the cartilage roof, drawn using the turning and labrum points.

3.5.2 α-angle and β-angle

Two angles called the α-angle and the β-angle are automatically computed from the Infant Hip landmarks as shown in Figure 3.12. The α-angle is the angle between the baseline and the bony roof. The baseline is a line along the ilium bone which is present at the left side of the image and should be close to horizontal. The bony roof is a line along the ilium bone when it curves down around the hip. The β-angle is the angle between the cartilage roof and the baseline. The cartilage roof is a line between the turning point of the ilium bone and the labrum which is a small piece of cartilage on the superior side of the socket.

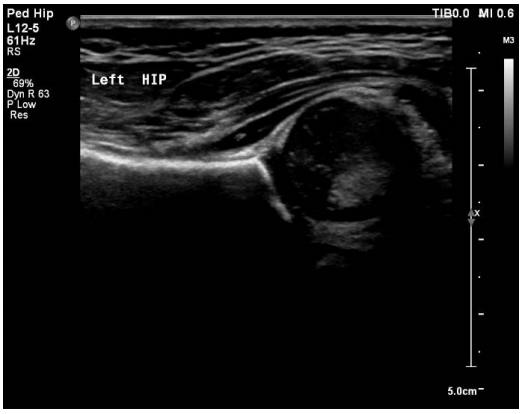
The α-angle and β-angle are used for, what is called, the Graf method of DDH classification. The α-angle measures the size of the bony socket and determines the types of DDH and the β-angle measures the size of the cartilage roof and determines the sub-type of DDH. The α-angle is recognised to be the easier angle for a clinician to measure because the ilium bone is visible. In contrast, the shape

of the labrum cartilage, upon which the β-angle is based, is hard to see because it is soft tissue. This is supported by the larger inter-rater variability for beta angles measurements [100]. For this reason, the British Society of Children's Orthopaedic Surgery has released consensus guidance that only the α-angle should be used to make a diagnosis [1]. This is therefore adopted in this section with hips with α-angles smaller than 50° being diagnosed as having DDH.

Dataset Details

This is the only non-X-ray dataset used in this thesis. The whole dataset consists of over 10000 images however only 190 of those were available when the research presented in this thesis was being conducted. The individual images were selected from ultrasound sequences based on the clinician's judgement. If the clinician believed they had a sufficient view enabling them to make a screening decision, they saved the image. In this context, a 'sufficient view' means the clinician could clearly see the ilium bone aligned horizontally on the left of the image. Furthermore, they could also observe the labrum and ischium bone, the latter of which is visible below the lower limb point.

The 190 images were taken from 100 multi-ethnic babies, which were obtained by experienced radiographers during routine NHS practice. 71 of the hips captured are normal, 66 are dysplastic which means a hip socket that doesn't fully cover the femoral head, and 53 are dislocated. All children were under three months old when their ultrasound was taken. Images that were too dark or did not fully capture the hip were excluded from the dataset by a clinician before any experimentation was performed. The images were obtained as `.jpg` after being converted from `.dicom` by a clinical collaborator. This process is lossy and the average resolution of an image in the dataset is 689 pixels in width and 523 pixels in height which is likely to have been less than that of the original `.dicom`. The images have a greyscale bit-depth of eight bits which is standard for the `.jpg` format. Once obtained from the clinician the images were not resized or cropped. As there is no pixel-to-mm metadata, distance is reported in the pixels for this dataset in this thesis.



(a) A dark scan capturing a non-dysplastic hip.



(b) A light scan capturing a dysplastic hip. This can be seen by observing that at least half the hip is above the ilium bone.

Figure 3.13: Images from the Infant Hip dataset.

Each image has five landmarks labelled in Figure 3.11. Landmarks were manually annotated on each image once by a Speciality Registrar in Orthopaedics. Images were annotated using the GIMP image editing software in the same way as the Pelvis dataset. There are two landmarks along the ilium bone which are used to draw the baseline. The ilium turns downwards and the point where the bone has the steepest downward angle is labelled with the turning point landmark. The lowest visible point on the ilium is labelled as the lower limb point and the line between this point and the turning point is called the bony roof. Finally, there is a piece of cartilage called the labrum on the upper side of the hip socket. A landmark is placed in the centre of this cartilage and the line between it and the turning point is called the cartilage roof.

Dataset Discussion

The hip components can appear in different quadrants of the images, as shown by Figure 3.13. All images feature a border containing details on how the scan was taken on the left and a measuring rule on the right. A wide range of image brightnesses appears in the dataset from dark images like Figure 3.13a to lighter images like Figure 3.13b. Hips of different dysplasia levels are shown in Figure 3.13 with 3.13a being a normal hip and Figure 3.13b being the most abnormal which

can be seen by the amount the femur has risen above the baseline. More example images from the dataset are given in Figure A.2 in Appendix A.

Only a single clinician annotated these images so no intra or inter-operator variability statistics have been available. However, these landmarks are likely to have high variability because they are defined ambiguously. The ilium points can be placed anywhere on the ilium bone, the approximately horizontal bone in the images. The turning point and lower limb points are well defined however the labrum point was defined as the centre point of the labrum cartilage. However, the boundary of this bone is not clearly defined and neither is the concept of the 'centre' of bone of this shape. It is likely, therefore, that there would be variability in the placement on that point. Despite having a variable placement it is useful to know that the labrum point is present because that indicates the image captures the correct plane in the ultrasound.

3.6 Dataset Characteristics

In this section, the differences between the datasets are quantified to give further evidence that these datasets are sufficiently varied. Table 3.2 details the basic information of each dataset stating that the number of images in the datasets ranges from 190 to 895, with between five and 37 landmarks detected and that the datasets' have a wide range of image resolutions.

Dataset	No. of Images	No. of Landmarks	Mean Resolution
Cephalometric	400	19	(1935, 2400)
Hand	895	37	(1560, 2167)
Pelvis	375	8	(2961, 2794)
Infant Hip	190	5	(689, 523)

Table 3.2: A table showing the basic information for each dataset.

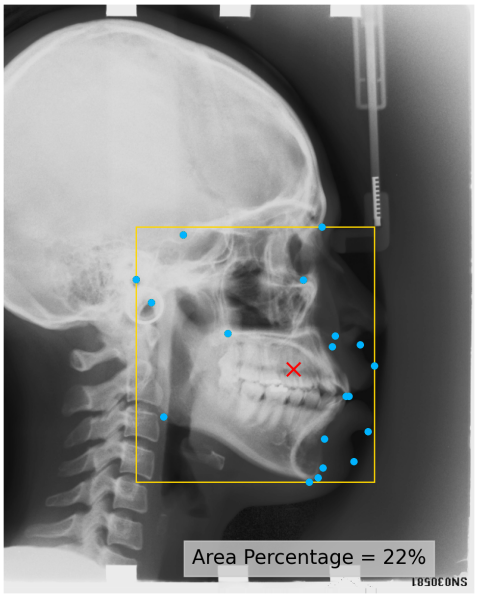
3.6.1 Landmark Characteristics

Dataset	Landmark Area Percentage	Mean Landmark Variance
Cephalometric	$22.43\% \pm 2.81\%$	$\leftrightarrow 2.7\%$ $\updownarrow 3.2\%$
Hand	$73.23\% \pm 6.47\%$	$\leftrightarrow 8.8\%$ $\updownarrow 4.1\%$
Pelvis	$9.64\% \pm 1.21\%$	$\leftrightarrow 2.3\%$ $\updownarrow 7.3\%$
Infant Hip	$5.24\% \pm 1.11\%$	$\leftrightarrow 4.9\%$ $\updownarrow 10.5\%$

Table 3.3: A table showing the novel statistics proposed for each dataset.

Two novel statistics, called Landmark Area Percentage and Mean Landmark Variance, are proposed in this thesis to quantify the differences in the characteristics of landmark placement for each dataset. Statistics for each dataset are shown in Table 3.3.

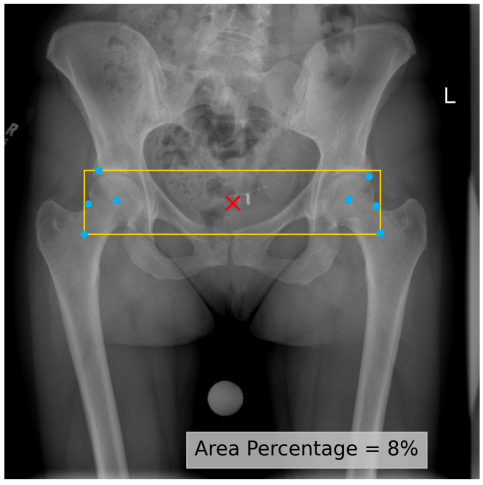
- **Landmark Area Percentage** - This statistic quantifies how spread out the landmarks are across the images in a dataset. The landmark area for a single image is the area of the minimal bounding box which encapsulates all landmarks divided by the total area of the image. Examples of the bounding boxes and the area percentages are given in Figure 3.14. This statistic for each



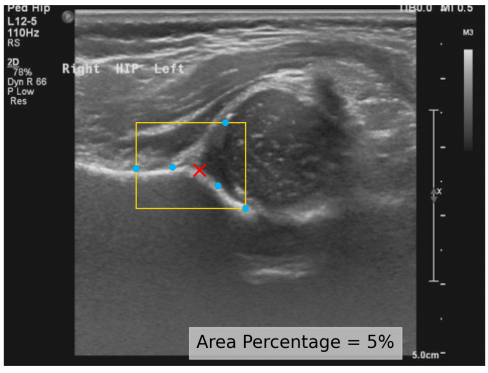
(a) Cephalometric example.



(b) Hand example.



(c) Pelvis example.



(d) Ultrasound hip example.

Figure 3.14: Each image shows the smallest bounding box around the landmarks with the percentage area that the bounding box takes up of the total image size displayed. The red cross in each image signifies the average landmark position which is used to measure how much shift is present within each dataset, as shown in Table 3.3.

dataset is given in the second column of Table 3.3 and it shows that the Hand dataset's landmarks cover more area in each image than the other datasets, with the ultrasound dataset containing landmarks closest to each other. It is important to note this statistic because different landmark detection methods could struggle with either large or small landmark areas. For example, if a CNN has a small receptive field (number of pixels the CNN sees when making a prediction) it might confuse landmarks that are far away from each other in an image, such as those in the Hand dataset.

- **Mean Landmark Variance** - This statistic quantifies how similar the position of the landmarks are in each dataset. This statistic is calculated by finding the centroid position of the set of landmarks in each image in the dataset, signified by the red cross in Figure 3.14. The standard deviation of this point, in each of the x and y directions, across each dataset, is calculated and normalized to the width and height of the images. As shown of Table 3.3 we find the Cephalometric dataset contains the most uniform images and the Infant Hip dataset contains the least, closely followed by the Hand dataset. In these datasets, the body part can appear in different positions in the image. This is an important statistic to analyze because, if a dataset is very uniform and the landmarks are in similar pixel positions in each image, then the method could overly rely on learning the absolute positions of the landmarks instead of identifying them using surrounding pixel intensities.

Table 3.3 shows that the Landmark Area Percentage varies from 5.24%, for the Infant Hip dataset, to 73.23%, for the Hand dataset. The table also shows that the Mean Landmark Variance varies from 2.3% to 8.8% horizontally and 3.2% to 10.5% vertically between the datasets experimented with. Therefore, these novel statistics further prove that the four datasets described and used in experiments are adequately varied such that robust validation of the methods presented is performed. The datasets are varied in all aspects, from the body part and modality to the number of images to the landmark spread and variation.

4

Kolmogorov-Smirnov Test

Contents

4.1 Disadvantages Of Random Set Selection 69
4.2 Disadvantages Of K-Fold Cross Validation 70
4.3 Kolmogorov-Smirnov Test Method 71
4.3.1 Choosing The Number Of KS Tests To Perform 71
4.4 Experiments Performed 72
4.5 KS Test Results 76
4.5.1 Distribution Of KS Test Scores 76
4.6 Generalisability Of The KS Method 79
4.7 Conclusion 81

An important consideration to ensure the robust validation of methods is how the images in datasets are placed into training, validation, and test sets. In machine learning the training set is a set of images used to train a model. Then, during training, model hyperparameters are adjusted to maximise the performance of the model over the validation set. Finally, once a model has been trained with the best-performing hyperparameters, it is then run over the test set where the final accuracy is reported. Which images are in each of the training, validation and test sets is important for model training and model validation. No images can be in more than one of these sets.

The datasets presented in this thesis fall into different categories. The Cephalometric and Hand datasets have existing training, validation and test sets proposed by previous publications. However, the Pelvis and Infant Hip datasets have been sourced privately and their images needed to be distributed carefully into training, validation and test sets.

It is important to note that this chapter describes a method used for placing images from datasets into training, validation and test sets. This is different from the task of determining how many images should be in each of these sets. For the method presented in this chapter it is assumed that the number of images in the training, validation and test sets is manually predefined. **This chapter refers to the images in a particular training, validation and test set as an 'allocation', whereas the number of images in each set is referred to as the 'split ratio'.**

4.1 Disadvantages Of Random Set Selection

The majority of the existing literature that uses private datasets does not mention a validation set [46, 79, 83, 140]. Of those that do, the only method given for placing images of a dataset into training, validation and test sets is random selection. Although this approach is simple and fast, it has the disadvantage that there is no guarantee that the training, validation, and test sets contain similar proportions of patients with certain conditions, ethnicities or unusual presentations (e.g. implants). These sets must be similarly distributed to ensure that method validation is robust. If the training set includes images with a similar presentation (e.g. medical conditions, scanner settings, demographics) to images in the test set, then it is likely a model trained on this set will perform well; however, if the test set contains images with different features to the training set, then it is likely the model will make mistakes on these images. This phenomenon is known as domain shift. An extreme case of domain shift is when a model is trained on one scanning modality and then used to make predictions on images acquired with a different modality.

Medical imaging datasets are generally smaller than datasets used in general computer vision and, in addition, there are often a small number of outlier examples in these datasets that can skew results. When only a small proportion of small datasets contain images that are difficult for the model (e.g. patients with unusual presentation) it is unlikely that randomly placing images into training, validation, and test sets will distribute these examples across these sets evenly.

4.2 Disadvantages Of K-Fold Cross Validation

An alternative to using a single random training, validation and test is to use cross-validation where each experiment is performed a pre-determined number of times (for example ten) and each with a different, randomly chosen, training, validation and test set allocation. Whilst this approach produces more reliable results by averaging over multiple random allocations it has the following disadvantages. Firstly, it is time-consuming to perform a large number of experiments when each involves training a model. A model takes in the order of hours to train so training a model ten times may not be viable for hyperparameter optimisation. Secondly, even using modern profiling tools, it is difficult to keep track of the different training, validation and test sets used which makes it hard to reproduce experiments. Thirdly, understanding the limitations of a model is harder when random sets are chosen. This is vital in medical imaging where it is useful to know which demographics and conditions are in each subset.

For these reasons, a novel method is presented in Section 4.3 for producing a single training, validation and test triplet in a way which distributes images evenly across these sets to perform more robust validation and help to understand which images the model is trained on.

4.3 Kolmogorov-Smirnov Test Method

Instead of generating just a single training, validation and test set allocation, this method generates a large number of random allocations (see Section 4.3.1 for details) and then uses the Kolmogorov-Smirnov (KS) test [67, 121] to determine which random training, validation and test set allocation has the most evenly distributed images.

The KS test is a general measure of how similar distributions are. In this context it is used to test which training, validation and test set allocation contains the most similar distributions of a particular statistic. The statistic used for the KS test can be any function that maps either the image, annotations or image metadata to a real number. For each random training, validation and test set allocation a predefined statistic is computed from each image in each set. These statistical values are then sorted for each set and their cumulative distributions are plotted such that the x-axis is the value of the statistic and the y-axis is the cumulative proportion of that set which has that value or less. For this task, the KS test score equals the greatest distance between any two of the cumulative distributions of the training, validation and test set (Concrete examples are given in Section 4.4).

The lower the KS test score the more alike the training, validation and test sets are with respect to the chosen statistic. The final training, validation and test set allocation is the one with the lowest KS test score out of all the random allocations generated and tested. This method ensures that the training, validation and test set each contains images with chosen characteristics. This method has not been used to evenly distribute images into training, validation and test sets in any existing machine learning literature and is novel to this thesis.

4.3.1 Choosing The Number Of KS Tests To Perform

How many random training, validation and test set allocations need to be generated and tested depends on several factors. With a larger dataset, more allocations need to be generated to find one with a close-to-optimal KS test score. However, a lot of computational power and time is required to perform the KS test over

large numbers of training, validation and test sets. Whilst the KS test can be calculated for a single allocation very quickly (0.3ms on an average laptop for the Pelvis dataset with 375 images) there is a huge number of possible ways images from a dataset can be allocated into training, validation and test sets. The number of allocations is given by the formula:

$$\frac{n!}{k_1! \times k_2! \times k_3!} \tag{4.1}$$

where n is the total number of images in a dataset and k_1, k_2 and k_3 are the numbers of images in the training, validation and test sets such that $k_1 + k_2 + k_3 = n$ and the split ratio is $k_1 : k_2 : k_3$. For the Infant Hip dataset, the smallest dataset presented in this thesis, $n = 190$, $k_1 = 95$, $k_2 = 48$ and $k_3 = 47$ (when placing images into the training, validation and test set with a split ratio of 50:25:25). Applying these numbers to the formula above results in 2.919×10^{83} possible training, validation and test set allocations which is an intractable number to perform the KS test over. Ultimately, the number of random allocations that need to be generated and tested is a practical balance between sufficiently many to ensure that an allocation with a KS test score much lower than an average random allocation is found and sufficiently few that the calculation remains computationally feasible.

4.4 Experiments Performed

In the experiment described in this section the α-angle was used as the statistic for both the Pelvis and Infant Hip datasets. It is worth noting that the α-angle has a different definition for the Pelvis and Infant Hip datasets as explained in Section 3.4.2 and Section 3.5.2. The reason these statistics were chosen was to make sure that models would be trained on a collection of patients with a full range of hip presentations. It was also because patients that have hips with higher α-angle have either FAI or DDH conditions and appear less frequently in the population [3]. It was therefore important to have patients with these conditions suitably represented in the training, validation and test sets.

	Number Of Images		
Dataset	Training Set	Validation Set	Test Set
Pelvis	263	56	56
Infant Hip	95	47	48

Table 4.1: A table showing the number of images in the training, validation and test set for the Pelvis and Infant Hip dataset. One hundred thousand (10^5) different random training, validation and test set allocations were generated with these numbers of images in each set and the KS test was performed on each of these. The number of images in the Cephalometric and Hand training, validation and test sets are omitted from this table because the KS method was not performed on these datasets as they already have defined training, validation and test sets given by the original authors of the dataset.

The images in the Pelvis dataset and the Infant Hip dataset were placed into training, validation and test sets with a split ratio of 70:15:15 and 50:25:25 respectively. The Infant Hip dataset contains only 190 images compared to 375 in the Pelvis dataset so different training, validation and test set split ratios were chosen such that a reasonable number of images were placed into the validation and test sets. The number of images in each set for each dataset is shown in Figure 4.1.

One hundred thousand (10^5) different random training, validation and test set allocations were generated, α-angles were calculated for each, and the allocation with the lowest KS test score was chosen as the final training, validation and test set. Although there is still an element of randomness in this process, there was a very high chance with this number of allocations generated that one considerably more representative than the average random allocation would be found. This is analysed in Section 4.5.1.

Calculating the best KS test score for 10^5 different random allocations computed in less than five minutes for these datasets on a CPU. The statistics for each image were pre-computed before the KS tests were performed to increase the efficiency of the algorithm.

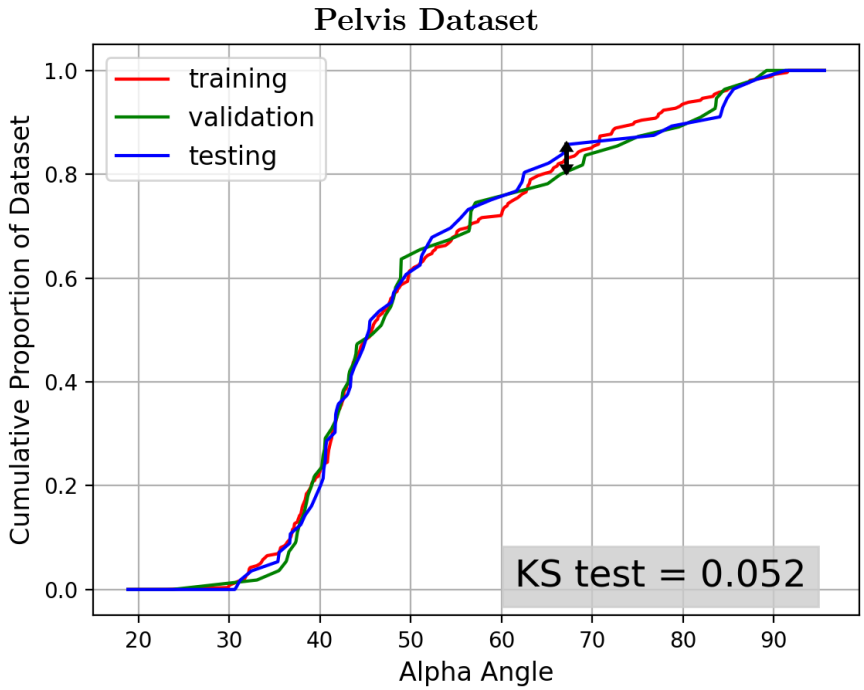


Figure 4.1: This graph shows the cumulative distributions of α-angle for the training, validation and test set with the lowest KS test score from 10^5 random allocations generated for the Pelvis dataset.

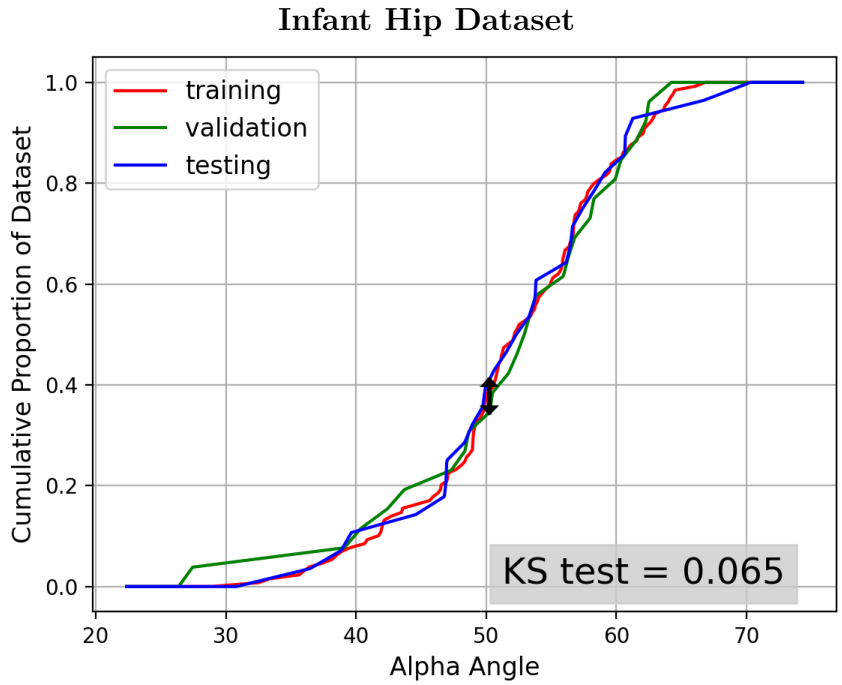
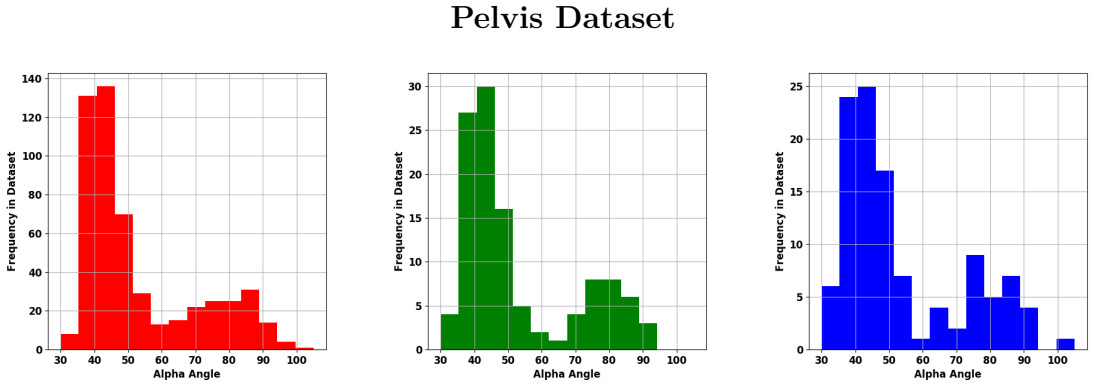
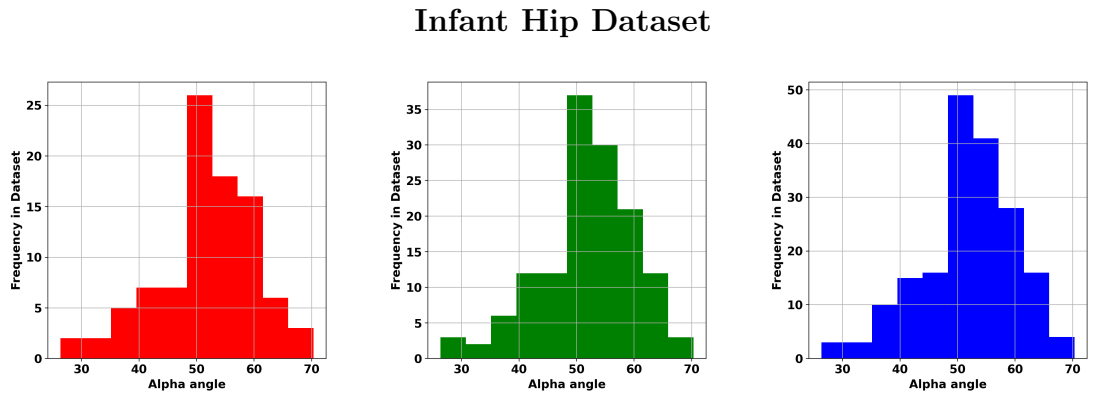


Figure 4.2: This graph shows the cumulative distributions of α-angle for the training, validation and test set with the lowest KS test score from 10^5 random allocations generated for the Infant Hip dataset.



(a) Training set **(b)** Validation set **(c)** Testing set

Figure 4.3: The distribution of α-angle for the training, validation and test sets with the lowest KS test score of the Pelvis dataset. They are visually similar and display a multimodal distribution consistent with what is seen in the general population [3].



(a) Training set **(b)** Validation set **(c)** Testing set

Figure 4.4: The distribution of α-angle over the training, validation and test sets with the lowest KS test score for the Infant Hip dataset. It can be seen that they are visually similar.

4.5 KS Test Results

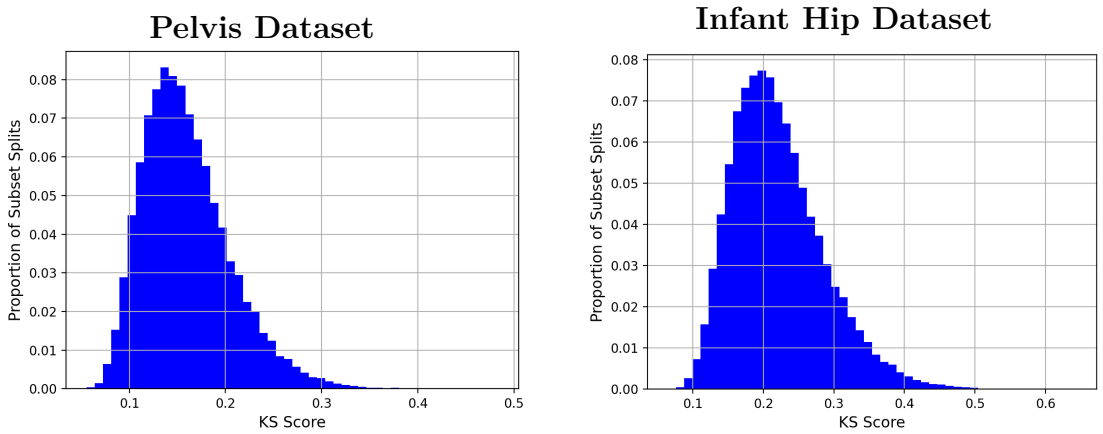
The cumulative distributions of α-angle for the random training, validation and test set allocation which achieved the lowest KS test score are shown in Figure 4.1 for the Pelvis dataset and Figure 4.2 for the Infant Hip dataset. These figures show that the lowest KS test score over the 10^5 random allocations for the Pelvis and Infant Hip datasets were 0.052 and 0.065 respectively.

Figure 4.3 and Figure 4.4 show the distribution of α-angles for these training, validation and test sets and it is visually clear that they were similar in both cases which proves the effectiveness of the KS method. The distribution of α-angles in the Pelvis dataset is multimodal with the local maxima at 45° representing hips with no conditions and the local maxima at 85° representing hips with FAI. This is consistent with the α-angle distribution seen in the general population [3].

Other researchers can reproduce these results exactly (assuming they have access to the datasets) if the random seed used to generate the random allocations is shared with them. Alternatively, they could simply be provided with a `json` file with which images are in each set and validate for themselves that a similar distribution of α-angle is present in each. Finally, if other researchers were to perform the same experiment with a different random seed it is likely that they would also obtain a highly representative training, validation and test set. This is explained in Section 4.5.1.

4.5.1 Distribution Of KS Test Scores

The KS test score distributions for the 10^5 random training, validation and test set allocations generated are plotted in Figure 4.5a and 4.5b for the Pelvis and Infant Hip datasets respectively. They show that the mean and standard deviation of the KS test score for a random training, validation and test set was 0.160 ± 0.046 for the Pelvis dataset and 0.221 ± 0.065 for the Infant Hip dataset. This shows that the average random allocation was significantly less representative than the best allocations found which had KS test scores of 0.052 and 0.065. These results show that generating 10^5 different random allocations was enough to find



(a) The distribution of KS test scores for the 10^5 random training, validation and test sets which were generated and tested for the Pelvis dataset. The mean and standard deviation of the KS test score was 0.160 ± 0.046.

(b) The distribution of KS test scores for the 10^5 random training, validation and test sets which were generated and tested for the Infant Hip dataset. The mean and standard deviation of the KS test score was 0.221 ± 0.065.

Figure 4.5: Histograms of the KS test scores for 10^5 random allocations for each dataset. In both cases, the mean KS test score was far greater than the optimal KS test shown in Figure 4.1 and Figure 4.2. This shows that a single random training, validation and test set is unlikely to have a similar distribution of α-angles in each set.

one with a KS test score at least two and a half standard deviations lower than the average for both datasets. This shows that choosing a random training, validation and test set is likely to inflate or deflate results for an experiment but the KS method ensures robust validation.

The distributions in these figures show that the probability of a random allocation having a significantly lower KS test score than the best found in this experiment is very low. This means that it is likely that orders of magnitude more random allocations need to be generated and tested to find one with a significantly lower KS test score. However, testing orders of magnitude more random training, validation and test sets requires time and computation power. These results further prove that the choice of generating 10^5 random allocations was enough.

In addition, these distributions show that if another researcher were to perform these experiments again there is a high chance that they would find a training, validation and test set with a similarly low KS test score. This is because the distributions appear roughly Gaussian (with a slightly positive skew). By treating these distributions as Gaussian it becomes possible to calculate an estimated

probability of encountering similarly low KS test scores if the experiment were to be repeated.

Approximating the Pelvis KS test score as a Gaussian distribution (with a mean and standard deviation of 0.160 ± 0.046) produces a probability of 99.59% that a training, validation and test set allocation with a KS test score of less than 0.06 will be found. Likewise approximating the Infant Hip KS test score as a Gaussian distribution (with a mean and standard deviation of 0.221 ± 0.065) produces a probability of 91.74% that an allocation with a KS test score of less than 0.07 will be found. Whilst the true KS test distributions are not exactly Gaussian these calculations show that finding representative training, validation and test sets using this method is reproducible.

4.6 Generalisability Of The KS Method

The KS method can be generalised to different datasets. For the Pelvis and Infant Hip datasets the α-angle measurement was used as the statistic used to compute the KS test score but this can be changed for other datasets. Any statistic which is a real number and is associated with, or calculated from, each image can be used. The statistic could be a function of the image's pixels, for example, image brightness, image blurriness or image contrast. The statistic could be a function of the landmark annotations such as average landmark spread. The KS method also works for non-landmark detection datasets. For example, the size (number of pixels) of a tumour present in an image could be used as the statistic for a tumour segmentation dataset.

In addition, the statistic could be clinical metadata attached to each image such as the height, weight or age of the patient captured in the image. It is even possible to encode discrete properties such as ethnicity to work with this method. An example of this is given in Figure 4.6. This figure shows a synthetic example with no real images attached to each value but it nevertheless illustrates how the KS method could be used. Different ethnic groups are encoded as a number and the KS test score is calculated over the cumulative distributions just like in Section 4.3. In this case, there is a large disparity between the demographics represented in the training and validation set compared with the test set. The test set has significantly fewer White British patients present. If different demographics present conditions in different ways, it is vital to train a model on a representative training set so that the model does not underperform for particular demographics. It is worth noting that future work should explore other methods, such as the chi-square goodness of fit test [24], for obtaining balanced training, validation and test sets with respect to discrete properties. This method has been used for statistical analysis in other non-machine learning publications [16].

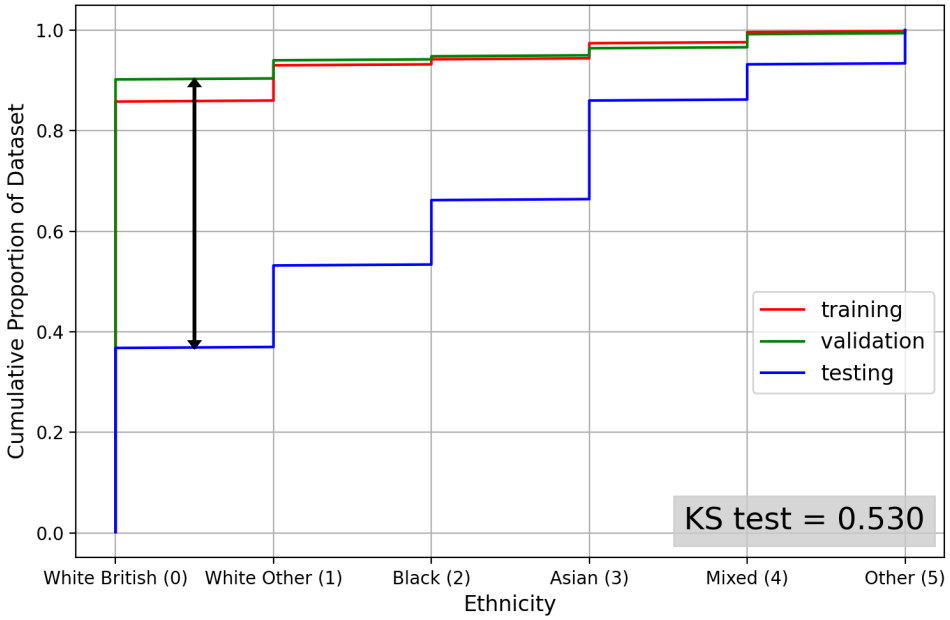


Figure 4.6: This figure shows how the KS method can be used for discrete statistics, such as ethnicity, associated with each image. Note that this is a dummy example with no real images associated with each set. Each demographic is encoded with a value, in this case White British (0), White Other (1), Black (2), Asian (3), Mixed (4), Other (5). Each image in the dataset is given one of these values depending on the ethnicity of the patient captured in the image. The cumulative distribution of these values is plotted for the training, validation and test sets and the KS score is calculated. In this example, there is a large disparity between the demographics represented in the training and validation set compared with the test set.

4.7 Conclusion

This section proposed a method for producing a representative training, validation and test set allocation by generating a large number of random allocations and then performing a KS test on each of them to calculate which allocation is the most representative with respect to a particular statistic. This method ensures that the training, validation and test sets are similarly distributed, unlike a random allocation, which leads to more robust testing of methods. The KS test method also improves the reproducibility of experiments because the representative training, validation and test set allocation generated using the KS method can be shared with, or recomputed by, other researchers so they can perform the same experiments.

Using the KS test method a representative training, validation and test set was computed for the Pelvis and Infant Hip datasets. For both datasets 10^5 random allocations were generated and the KS test was performed on each of these using the α-angle as the statistic. The allocations with the lowest KS test score for each dataset were selected as the final training, validation and test set and were used for all experiments described in the remainder of this thesis.

In addition, this section described how this method can be generalised to other datasets and statistics, including discrete statistics such as ethnicity. The KS method is especially useful for allocating images to training, validation and test sets with respect to properties which cannot be reproduced using data augmentation in the training process. Properties such as the severity of hip conditions or demographic information cannot be synthetically added to the training set so ensuring that those domains are represented in the training set leads to more robust models.

Future work can extend the KS method to work for multiple statistics at once. This can be achieved by calculating the KS test score for multiple different statistics for each training, validation and test set allocation generated. The final allocation chosen would be the one which achieves the lowest combined KS test score over all the statistics. This would ensure that training, validation and test sets are representative of demographic information, image information and clinical metadata at the same time.

5 Landmark Detection

Contents

5.1 Experimental Setup 87
5.1.1 Preprocessing 87
5.1.2 Post-processing 91
5.1.3 Validation Metrics 91
5.2 Hyperparameter Tuning 94
5.2.1 Experiments Performed 94
5.2.2 Data Augmentation 95
5.2.3 Heatmap Generation 99
5.2.4 Architectures 102
5.2.5 Loss Function 104
5.2.6 Hyperparameter Summary 105
5.3 Publicily Available Dataset Results 106
5.3.1 Cephalometric Dataset Results 106
5.3.2 Hand Dataset Results 109
5.3.3 Public Datasets Results Conclusion 111
5.4 Pelvis Dataset Results 112
5.4.1 Experiments Performed 112
5.4.2 Landmark Detection Results 112
5.4.3 LCE and Alpha Angle Results 116
5.4.4 FAI Diagnosis Results 118
5.5 Infant Hip Dataset Results 120
5.5.1 Experiments Performed 120
5.5.2 Landmark Detection Results 120
5.5.3 Alpha and Beta Angle Results 123
5.5.4 DDH Diagnosis Results 124
5.6 Discussion and Conclusion 125
5.6.1 Advantages of the Technology for Clinics 125
5.6.2 Landmark Detection and Angle Calculation 126
5.6.3 Diagnostic Performance 127

This chapter describes how deep CNN were trained to perform landmark detection on the Cephalometric, Hand, Pelvis and Infant Hip datasets. This section is split into three parts: In Section 5.1 the experimental setup is explained including the model training procedure and the image pre and post-processing steps taken. Section 5.2 describes how different parts of the training pipeline were optimised in a process called hyperparameter tuning. Finally Sections 5.3 to 5.5 report the results the final models achieved over the Cephalometric, Hand, Pelvis and Infant Hip test datasets.

The experiments described in this chapter used a supervised training scheme. Deep CNNs were trained using heatmaps centred at the ground truth landmark points. The model outputs a heatmap upon which a loss is calculated against the training heatmap and used to train the model. This way the model learns how to output heatmaps centred at the landmark points. This approach is visualised in Figure 5.1 and is a common approach in recent literature as discussed in Section 2.3. Although some very recent methods use a two-stage approach, as described in Section 2.4.3, recent work has also shown that often simpler networks, tuned correctly, can achieve SOTA results [59]. For this reason the experiments described in this section train a single, existing, network architecture (see Section 5.2.4). In addition, the major contributions in Chapter 6 utilize the output heatmaps of networks instead of proposing yet another architecture.

After the models were trained they were tested over each dataset's test. This process is visualised in Figure 5.2 and results for each model are reported in Section 5.3 to 5.5. The results for the Pelvis dataset, presented in Section 5.4, were published in the paper entitled *Automatically Diagnosing Hip Conditions from X-rays using Landmark Detection*, at the 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) [89]. In addition, the results for the Infant Hip dataset, presented in Section 5.5, were published in the paper entitled *Automatic identification of clinical landmarks and calculation of Graf angles in newborn hip screening: A pilot study* presented at the British Orthopaedic Association [120].

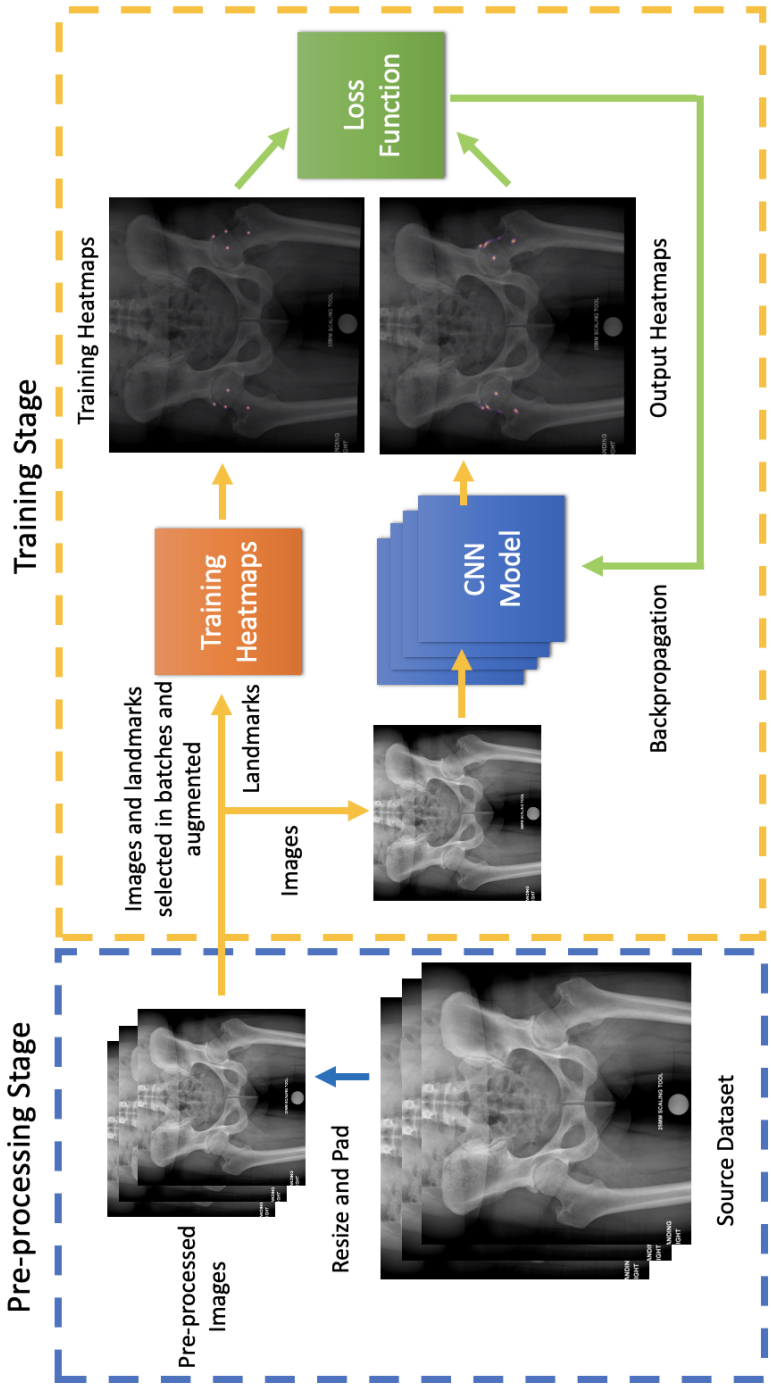


Figure 5.1: This figure shows the process which was used to train the landmark detection models described in this thesis. This specific figure uses images from the Pelvis dataset however the pipeline is the same for all datasets. Images were first resized and padded to make them have the same width and height. These images were then augmented and passed into the CNN model to produce output heatmaps. It is worth noting that there is a separate output heatmap for each landmark however these heatmaps have been compressed into a single image in this figure. A loss was then computed between these output heatmaps and the training heatmaps which were generated from the ground truth landmark positions. This loss was backpropagated to train the weights of the model.

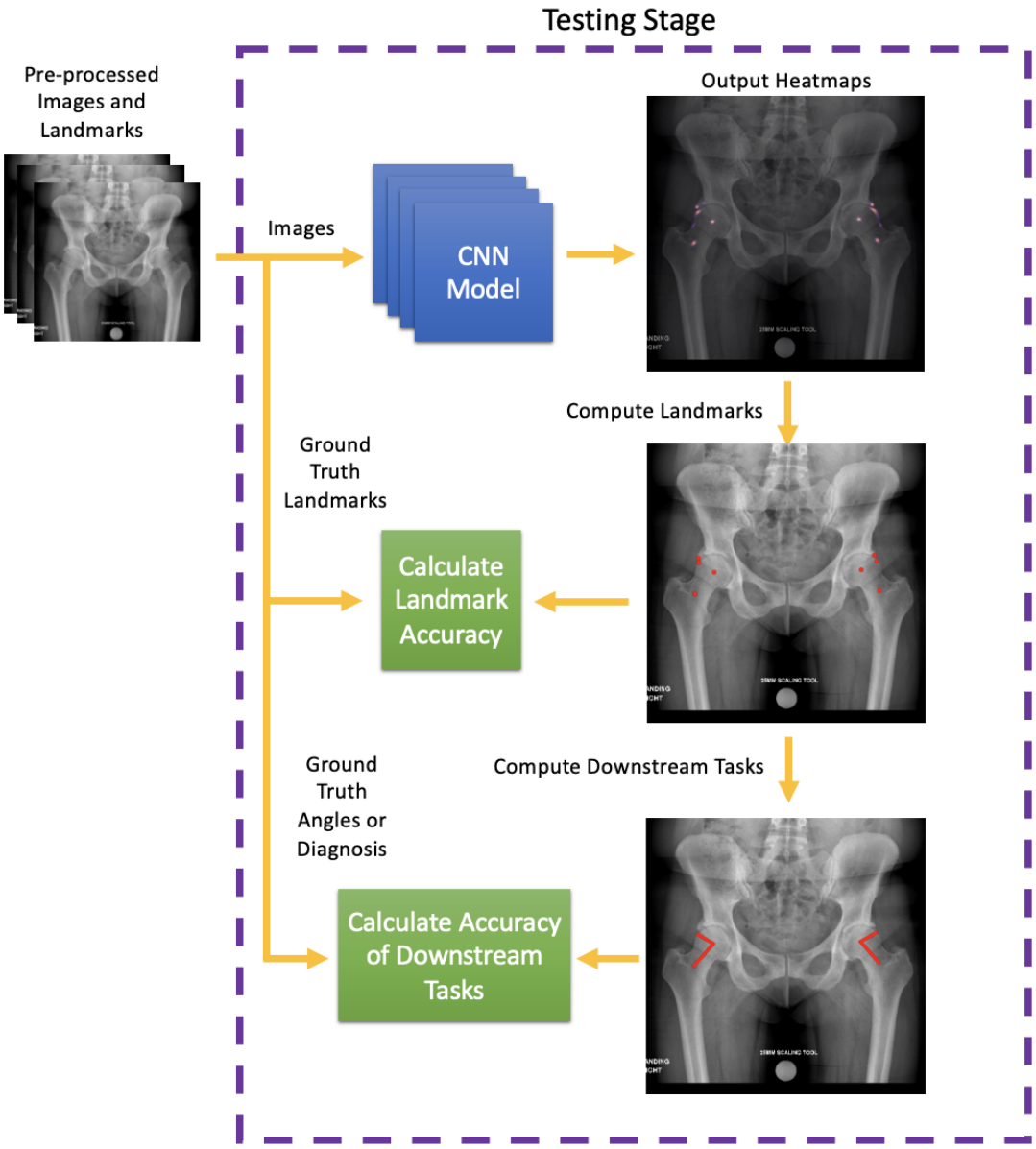


Figure 5.2: This figure shows the process which was used to test the landmark detection models described in this thesis. This specific figure uses images from the Pelvis dataset however the pipeline is the same for all datasets. Preprocessed images in the relevant test set were input into the trained CNN model to output heatmaps. It is worth noting that there is a separate output heatmap for each landmark however these heatmaps have been compressed into a single image in this figure. Predicted landmarks were then defined as the brightest points in each heatmap. The Euclidean distances between the predicted landmarks and the ground truth landmarks for that image were then computed (see Section 5.1.3 for specifics) and these results are reported for each dataset later in this chapter. In the case of the Pelvis and Infant Hip datasets angles and diagnosis were also computed from the predicted landmarks and compared against the ground truth.

5.1 Experimental Setup

All experiments were coded using the `PyTorch` library.[1] The code used for all experiments in this thesis is available online. The link and requirements are provided in Appendix E. The package `Weights and Biases` was used to keep track of the hyperparameters for each experiment and to log and store results.[2] Experiments were run on a node of Oxford University's Advanced Research Computing (ARC) service supercomputer equipped with a NVIDIA Tesla V100-SXM2 GPU with 32GB of memory. The effect different data augmentation methods, architectures, initial model weights and loss functions had on model accuracy is analysed in Section 5.2. However hyperparameters for the training procedure and gradient descent were preselected empirically and are listed here: The Adam optimizer was used for all experiments with a learning rate of 0.001 which was stepped down by a factor of 0.1 on epochs four, six and eight.[3] A batch size of six was used. All models were trained for ten epochs after observing that their performance over the validation set plateaued at this epoch during training. Ten epochs of training on the GPU node took between thirty minutes for the Infant Hip dataset to six hours for the Hand dataset.

The same network architecture was used for each dataset with the exception of the final layer which was changed to have the same number of channels as there are landmarks for each dataset. This made sure that the correct number of landmarks were predicted by each model. Although the same network architecture was used for all datasets, the model weights learned during training were naturally different.

5.1.1 Preprocessing

Before any network was trained, the images in each dataset were preprocessed. In this work, images were downsampled to a single consistent resolution for each dataset. The first reason for this is that the images in the Cephalometric, Hand and Pelvis datasets are too large (image resolutions stated in Table 3.2) to be passed

[1] `https://pytorch.org/`
[2] `https://wandb.ai/site`
[3] `https://pytorch.org/docs/stable/generated/torch.optim.Adam.html`

Dataset	Preprocessed Resolution
Cephalometric	(640, 800)
Hand	(608, 800)
Pelvis	(512, 512)
Infant Hip	(640, 480)

Table 5.1: This table shows the resolution the images in each dataset were resized to in the preprocessing step.

through a CNN without surpassing the memory constraints of the 32GB GPU used for experimentation. Secondly, images in the Hand, Pelvis and Infant Hip datasets vary in resolution and it was required that they be of uniform resolution to train the models with batches of images. Finally, all architectures used in this work have four downsampling layers each of which halves the resolution of an image in each dimension. For this reason, images were scaled to an input width and height which are each divisible by $2^4 = 16$.

The images were rescaled, maintaining their aspect ratio, to fit inside the chosen resolution shown in Table 5.1. Bicubic interpolation was used to rescale the images using the **`imgaug`** library function **`resize`**.[4] This process determines the intensity of every pixel in the downsampled image by first identifying its corresponding position in the original image. Then, subsequently, cubic interpolation is applied in both the horizontal and vertical directions based on the neighboring pixels in the original image [63].

The landmarks were also rescaled, relative to the top left corner of the image, so they were placed at the same anatomical positions on the rescaled image. Each rescaled image was then padded at either the bottom or the right side of the image with black pixels so all images in the dataset had the same resolution. The chosen input image resolutions in Table 5.1 were selected for two reasons: firstly, the less padding a preprocessed image has the better because constant value pixels are used for padding and give no more information to a model. Therefore, a resolution with a similar aspect ratio to the average aspect ratio of the original images was chosen. Secondly, images needed to be of high resolution to improve the localisation accuracy

[4] `https://imgaug.readthedocs.io/en/latest/source/api_augmenters_size.html#imgaug.augmenters.size.Resize`

of the model but also small enough such that a batch of six images could pass through our chosen architecture within the memory constraints of the 32GB GPU.

Although downsampling the images is required to train a deep CNN landmark detection model, downsampling does reduce the quality of the images and the resolution of key details in the image. An example of an original X-ray pelvis image, and its downsampled version, are shown in Figure 5.3. The small bump of the femoral head is harder to see in the downsampled image and this could affect the accuracy at which the models trained on these images can detect the lateral cam point used to measure the α-angle. It is worth noting that as mentioned in Chapter 3 images in all datasets have a bit-depth of eight. If this bit depth was greater then it is possible that this downsampling would have less of an effect on the quality of details in the image. Additionally, if the bit-depth was lower details like a small bump on the femoral head would be more difficult to spot in both the original and the downsampled images.

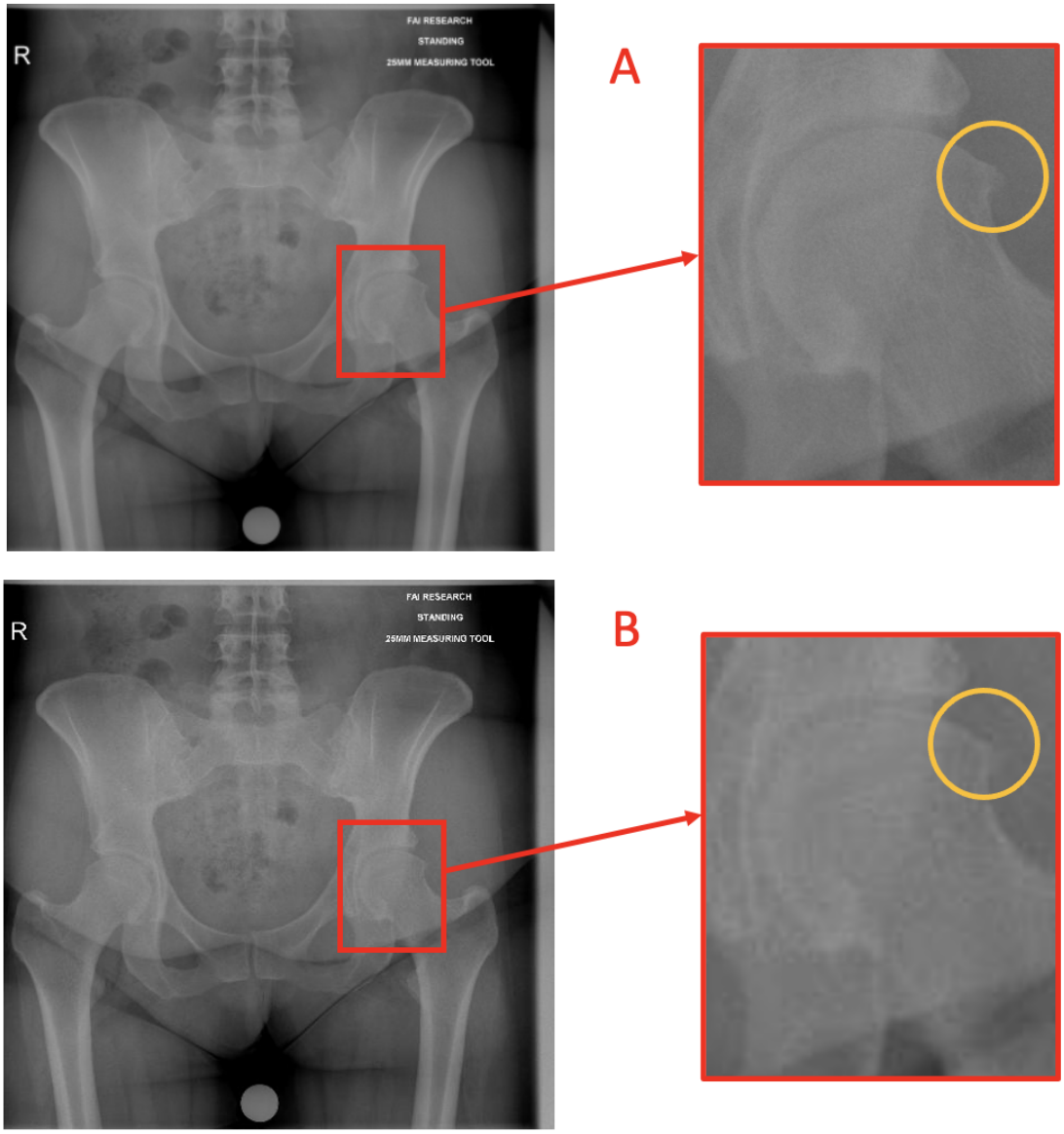


Figure 5.3: This figure shows the effect that downsampling the images has on the image quality. (A) shows a full resolution (2992 pixels by 2991 pixels) X-ray pelvis image and a cropped-out femoral head region. (B) shows the same image but downsampled to 512 pixels by 512 pixels which is the size of images upon which the models described in this section were trained. A corresponding cropped-out femoral head region is shown for the downsampled image. This image shows a femoral head with a small bump on it, highlighted by the orange circles. For the reasons described in Section 5.1.1 downsampling of images is essential for training deep CNN models. However, detail is lost in this process as shown by the clearer bump in the high-resolution image compared to the lower-resolution image. Detecting this bump is essential for calculating the lateral cam point landmark used to measure the α-angle.

5.1.2 Post-processing

The models described in this chapter were trained on heatmaps centred at the ground truth landmark positions. This means that the models learned how to output heatmaps centred on the landmark positions. At both training and test time, each output heatmap from the relevant model was passed through a two-dimensional softmax function which turned the output into a probability distribution over the domain of all pixels in the image. The reason for this was so a loss could be computed on the output heatmaps to train the model. Formally each network outputs a heatmap $\tilde{H}$ for each landmark l. $\tilde{H}_l \in \mathbb{R}^{w \times h}$ where w and h were the width and height of the input images. It is worth noting this is different from input training heatmaps which use the symbol H later in this chapter. The two-dimensional softmax function $S_l(\cdot, \cdot)$ worked on each heatmap independently and is defined by the following equation:

$$S_l(i,j) = \frac{e^{\tilde{H}_l(i,j)}}{\sum_{s=1}^{w} \sum_{t=1}^{h} e^{\tilde{H}_l(s,t)}} \tag{5.1}$$

where $i \in \{1, 2, 3, ..., w\}$ and $j \in \{1, 2, 3, ..., h\}$ are coordinates on the output heatmap $\tilde{H}$. After each heatmap was passed through the two-dimensional softmax function the pixels with the highest values in each were chosen as the final landmark predictions.

5.1.3 Validation Metrics

Two metrics, called Mean Radial Error (MRE) and Successful Detection Rate (SDR), are used to measure the performance of landmark detection models in past literature and for the work presented in this thesis. MRE is the mean average of the radial errors over a set of landmarks. Radial error is the Euclidean distance between a ground truth landmark placement and its corresponding predicted landmark placement measured in mm. This is defined formally as the following:

$$\text{MRE} = \frac{1}{n} \Sigma_{i=0}^{n} |x_i - y_i| \tag{5.2}$$

where n is the number of landmarks and $x_i, y_i \in \mathbb{R}^2$ are the predicted and ground truth landmarks respectively. $|x_i - y_i|$ is the Euclidean distance between the prediction and ground truth for the i-th landmark. The lower the MRE the more accurate the predictions.

SDR is defined as the percentage of predicted landmarks in a set with an MRE less than a selected value. Formally this is expressed as:

$$\text{SDR}_V = 100 \times \frac{|i \in [0...n] : |x_i - y_i| < V|}{n} \tag{5.3}$$

where V is an arbitrary value defined by the user. In the work presented in this thesis V was chosen based on what is deemed acceptable within a clinical setting, as well as on the values used in previous literature for a dataset so that a direct comparison to past approaches could be made.

Hyperparameter tuning described in Section 5.2 was performed by selecting which hyperparameters of a model led to the lowest MRE (most accurate) score over the validation set. Once hyperparameters had been tuned, models were retrained and tested on each dataset's test set. MRE and SDR results are shown over each dataset's test set in Section 5.3 to Section 5.5.

Intraclass Correlation Coefficient

As explained in Section 3.4.1 and Section 3.5.1, landmarks can be connected to form clinically useful angles used for the diagnosis of conditions. Section 5.4 and Section 5.5 present the results obtained when predicted landmarks are connected to calculate LCE-angles and α-angles for the Pelvis dataset and α-angle and β-angle for the Infant Hip dataset. To validate the accuracy of these angle predictions they were compared using metrics to ground truth angle predictions calculated from the ground truth landmark points.

As well as the mean and median average pairwise difference between the ground truth and predicted angle values, the Intraclass Correlation Coefficient was also used as a metric. The Intraclass Correlation Coefficient, originally proposed by Ronald Fisher [38], is a statistical measure used to assess the consistency of measurements

made by different individuals or methods on the same subjects. Within the context of this thesis, the Intraclass Correlation Coefficient (ICC) is a statistic which measures the correlation between the ground truth and the predicted values for a particular angle calculated over a set of images. Consider a set of images which, in total, capture m angles each of which has a ground truth value of $\theta_{i,1}$ and a predicted value of $\theta_{i,2}$ for $i = 1, ..., m$. The ICC is defined by:

$$\text{ICC} = \frac{1}{ms^2}\Sigma_{i=0}^{m}(\theta_{i,1} - \bar{\theta})(\theta_{i,2} - \bar{\theta}) \tag{5.4}$$

where:

$$\overline{\theta} = \frac{1}{2m}\Sigma_{i=0}^{m}(\theta_{i,1} + \theta_{i,2}) \tag{5.5}$$

$$s^2 = \frac{1}{2m}\left(\Sigma_{i=0}^{m}(\theta_{i,1} - \bar{\theta})^2 + \Sigma_{i=0}^{m}(\theta_{i,2} - \bar{\theta})^2\right) \tag{5.6}$$

The ICC value is always within the interval of $[-1, 1]$. The closer to 1 the ICC is, the better the correlation between the ground truth and predicted angles across the set of images.

5.2 Hyperparameter Tuning

5.2.1 Experiments Performed

This section describes how hyperparameter tuning was performed so that the best possible training pipeline could then be used to train models over the full training sets in Section 5.3 for Cephalometric and Hand datasets and in Section 5.4 and Section 5.5 for Pelvis and Infant Hip datasets.

The hyperparameters optimised were data augmentation methods, training heatmaps, network architectures and loss functions. The experimental structure for testing different values of each of these hyperparameters was the same. For each value of the hyperparameter being experimented with, three models were trained on the training set of each dataset and the average performance of these three models was reported over each dataset's validation set. The value of the hyperparameter which obtained the best performance was selected as final and went on to be used for all other experiments in this chapter.

Training models on each dataset for every permutation of hyperparameters was impractical due to the model training time and the number of hyperparameter permutations.[5] Instead, different values for each hyperparameter were experimented with in isolation whilst the values of all the other hyperparameters were fixed. For example, Section 5.2.2 reports the results that different data augmentation methods obtained over the validation set. For these experiments, all hyperparameters, excluding data augmentation, were kept the same. The 'default' hyperparameters, which were kept the same unless the experiment actively tested different values for them, were as follows: all data augmentation methods applied, models trained on Gaussian heatmaps with $\sigma = 3$, UNet++ architecture initialised with pre-trained weights and NLL loss function.

[5]In this section seven different data augmentation methods are tested, eight different heatmap generation methods, two network architectures, two weight initialisations and two loss functions. If three repetitions of each experiment were performed this would mean that $7 \times 8 \times 2 \times 2 \times 2 \times 3 = 1344$ different models would have to be trained for each dataset. This is practically infeasible when the training time for a model takes between thirty minutes and six hours depending on the dataset.

For the experiments described in this section 80% of the training sets for the Cephalometric and Hand datasets were used to train the models and the other 20% were used as validation sets on which results are reported. This avoided tuning hyperparameters on images which are part of the test sets which is bad practice. For the Pelvis and Infant Hip datasets the training and validation sets were chosen using the KS test, as described in Section 4.3.

5.2.2 Data Augmentation

Data augmentation is the process of augmenting input images, and their corresponding landmarks, to artificially inflate the training set size without making the model more likely to overfit. It has been shown in the literature to improve results [59, 69, 111]. In addition, prior knowledge about the task can be encoded in the data augmentation approach. For example, if images in a dataset are captured from a wide variety of angles, then rotation augmentation in the same plane can ensure that a model is robust to that variance.

Experiments Performed

The augmentations used during training for the models presented in this thesis were rotation, intensity scaling, geometric scaling, X and Y translation, elastic deformation and horizontal flipping. These augmentations were performed using the `imgaug` library.[6] Each of these augmentations has hyperparameters representing the amount of the augmention. These hyperparameters are listed in Table 5.2 and were chosen to make sure that the training sets match the characteristics of the test set for each dataset. It was not feasible to test every combination of data augmentation methods or test all amounts of data augmentations. Therefore, for each dataset, three models were trained with no data augmentation, then another three models were trained with one data augmentation method applied, then another three models were trained with two data augmentation methods applied etc.... until three models were trained with all data augmentation methods applied. All of these models were

[6] `https://imgaug.readthedocs.io/en/latest/`

then tested on the validation sets and the results were averaged and reported in each group of three models. The following list describes each data augmentation method, the hyperparameters associated with it and why they were chosen for each dataset.

Hyperparameter	Cephalometric	Hand	Pelvis	Infant Hip
Rotation Factor	0	(-15°, 15°)	(-3°, 3°)	(-5°, 5°)
Intensity Factor	(0.5, 1.5)	(0.5, 1.5)	(0.5, 1.5)	(0.5, 1.5)
Scale Factor	(0.9, 1.1)	(0.9, 1.1)	(0.9, 1.1)	(0.8, 1.2)
X Transform	(-3%, 3%)	(-8%, 8%)	(-2%, 2%)	(-5%, 5%)
Y Transform	(-3%, 3%)	(-4%, 4%)	(-7%, 7%)	(-10%, 10%)
Elastic Strength	(0, 500)	(0, 100)	(0, 500)	(0, 500)
Elastic Smoothness	50	25	50	50
Horizontal Flipping	0	0	0.5	0

Table 5.2: The data augmentation hyperparameters used for each dataset. Cells with two numbers mean a value was randomly chosen within that range each time that augmentation was performed.

1. **Rotation**: This augmentation rotates each image a random number of degrees clockwise or anticlockwise within a range. This augmentation was not used for the Cephalometric dataset because all the images in that dataset are aligned because a physical vertical guide was used and is visible in all images. The Hand dataset had the most rotation augmentation applied because people placed their hands at different angles on the X-ray slide. Therefore 15° of rotation was used both clockwise and anti-clockwise. The Pelvis dataset images are mostly vertically aligned so only 3° of rotation augmentation was used. The Infant Hip dataset images were taken from a similar angle as is required in clinical practice and 5° of augmentation was chosen for this dataset.

2. **Intensity Scaling**: This augmentation multiplies all pixels of each image by a random constant value within a range. Pixels were multiplied by values between 0.5 and 1.5 for all datasets. All pixels were represented by a value between 0 and 255 and if a pixel value exceeded 255 after intensity scaling it was set back down to 255. Intensity scaling was performed because there was variation in brightness between the images in each dataset. The underlying

anatomy and landmarks should always be recognisable and thus models should be robust to this augmentation.

3. **Geometric Scaling**: This augmentation scales up or down the image by a certain percentage of its original size about the centre of the image. The values used were chosen based on the mean and standard deviation of the landmark spread statistic given in Table 3.3. For example, the landmark spread of the Cephalometric dataset is on average 22.43% with a standard deviation of 2.81%. As 2.81 is approximately 10% of 22.43%, images were scaled up or down by at most 10% to artificially ensure that the model was robust to all possible landmark spreads which were likely to appear in the test set. It is worth noting that, if the image was scaled up, some pixels on the borders got cropped out of the image. Alternatively, if the image was scaled down, black pixels were added to the image border to maintain the image dimensions.

4. **Transformation**: This augmentation shifts all pixels of the image up, down, left, and right by randomly chosen amounts. Similar to the scaling augmentation, the hyperparameters for this augmentation were chosen based on values in Table 3.3 but for the variation of landmarks. For example, the position of the landmarks in the Infant Hip dataset varies by about 5% horizontally and 10% vertically so these figures were used as the x and y shift amount for this augmentation. This ensured that examples in the test set were represented in the training set.

5. **Elastic Deformation**: This augmentation shifts pixels in local fields around a set of points. Other literature has shown this to be effective on medical imaging tasks [59, 111]. This augmentation increased the amount of training data without anatomically incorrect examples being introduced. Setting elastic smoothness to 25 or 50 ensured that the bones captured still looked realistic in the datasets without becoming too jagged.

6. **Horizontal Flipping**: This augmentation flips the image, and the associated landmarks, about the y-axis such that the landmarks on the right of an image are projected to the left and vice-versa. This augmentation was only appropriate for the Pelvis dataset where images capture symmetrical anatomy on the left and right sides. For this dataset images were horizontally flipped with a random chance of 50%.

	MRE (mm) over the validation sets. The mean and std over three seperately trained models is given.			
Augmentation	Cephalometric	Hand	Pelvis	Infant Hip
No Augmentation	1.153±0.029	0.712±0.026	2.318±0.025	9.319±1.168
+ Rotation	1.153±0.029	**0.688±0.018**	2.160±0.031	9.013±0.734
+ Multiply	1.003±0.019	0.703±0.009	2.122±0.020	8.368±0.118
+ Scaling	1.053±0.073	0.690±0.010	2.156±0.061	8.556±0.585
+ Translation	**0.982±0.016**	0.707±0.014	2.190±0.032	8.208±0.051
+ Elastic Deformation	0.983±0.020	0.694±0.007	2.275±0.036	**8.074±0.097**
+ Horizontal Flipping	-	-	**2.094±0.086**	-

Table 5.3: This table shows how models trained with different augmentation methods performed over the validation sets. It was not tractable to test every combination of augmentation methods on each dataset so instead no augmentation was tested and then data augmentations were added cumultatively. The top row of the table shows the results when no augmentation was applied and then each subsequent row shows the results when a data augmentation method was added so that the final row contains results when all augmentation methods were applied. The standard deviation value in each cell is between the MRE for each of the three models and not over the true radial errors. The best average MRE over three models for each dataset is shown in bold.

Results

Results for each data augmentation method, with the parameters stated in Table 5.2, are given in Table 5.3. The top row in the table presents results over the validation set with no augmentation and each subsequent row reports results with a data augmentation added with the final row showing results when all data augmentation were applied.

For both the Pelvis and Infant Hip datasets, applying all augmentation techniques achieved the best MRE results over the validation sets. For Cephalometric all data augmentation methods except elastic deformation improved results. However, the results with and without elastic deformation are within 0.001mm of each other,

which is negligible. For the Hand dataset applying only rotation augmentation achieved the best result. This shows that because the Hand dataset is considerably larger than the other datasets, there was already a sufficient variety of training data.

It is worth emphasizing that the standard deviation in MRE for a specific augmentation method often encompasses the MRE obtained by other data augmentation methods for the same dataset. An example of this is that the mean and standard deviation MRE for models trained with rotation augmentation for the Hand dataset was 0.688±0.018 and the mean and standard deviation when all data augmentations were used was 0.694±0.007. Not only are the mean MRE values only 0.006mm apart but the standard deviation results show that there was one model trained using all data augmentation methods which obtained a better result than one of the models trained just using rotation augmentation. These results show that if these experiments were repeated it is likely that models trained using different data augmentation methods would have performed the best.

On balance, these results show that augmentation during training is beneficial. It is difficult to measure the impact of individual data augmentation methods due to the inherent randomness of the training process. However, applying all augmentation methods either obtained the best results for a dataset, or results which were a negligible difference of 0.001mm or 0.006mm from the best for the Cephalometric and Hand datasets respectively. In all cases using all data augmentation methods achieved considerably better results than no augmentation. For these reasons, all data augmentation methods were used in all other experiments for all datasets.

5.2.3 Heatmap Generation

As stated in the introduction to this chapter the models train by computing a loss against heatmaps centred at the ground truth landmark points. This approach was chosen because, as described in Section 2.3, it has been shown to achieve SOTA results. The most frequently used approach in recent literature is to generate heatmaps using a Gaussian function like in the original Tompson *et al.* paper [129] (Examples of other methods in literature using this approach



Figure 5.4: This figure shows that training heatmaps for an image in the Hand dataset. In this example the heatmaps were generated using a Gaussian function with σ set to five. It is worth noting that there is a separate training heatmap for each landmark however they are merged into one image to show this figure.

were given in Section 2.4.2). These heatmaps are constructed using Gaussian distributions centred at the ground truth landmarks with a standard deviation σ, mathematically defined as:

$$H(i,j) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{|(i,j)-(x,y)|}{\sigma}\right)} \tag{5.7}$$

where x and y represent the coordinates of the ground truth landmark position and i and j are any coordinate of the output heatmap H. $|(i,j)-(x,y)|$ represents the Euclidean distance between the given coordinates i and j and the ground truth landmark position. An example of heatmaps generated using this function for an image in the Hand dataset is shown in Figure 5.4.

An alternative to Gaussian heatmaps is one-hot heatmaps. One-hot heatmaps are defined as a two-dimensional array where every element is zero except for a value of one at the location of the ground truth. This can be viewed as a way of

encoding a landmark and is equivalent to modelling the problem as a classification task where each pixel is a class and the correct pixel classification is represented as a one-hot vector. Using one-hot heatmaps for the purpose of training landmark detection models has not been experimented with in previous literature and is novel to this thesis. One-hot heatmaps are defined as:

$$H(i,j) = \begin{cases} 1, & \text{if } i = x \text{ and } j = y \\ 0, & \text{otherwise} \end{cases} \tag{5.8}$$

Experiments Performed

Three models were trained on each dataset's training set for each heatmap generation method. The heatmap generation methods tested were Gaussian heatmaps with σ set from one to seven and also one-hot heatmaps.

	MRE (mm) over the validation sets. The mean and std over three seperately trained models is given.			
σ	Cephalometric	Hand	Pelvis	Infant Hip
One-hot	1.059±0.076	0.741±0.024	2.199±0.044	9.150±0.513
1	1.009±0.060	0.698±0.009	2.137±0.061	9.495±1.100
2	1.005±0.010	0.707±0.005	2.109±0.010	8.640±0.642
3	**0.983±0.020**	**0.694±0.007**	**2.094±0.086**	**8.074±0.097**
4	1.029±0.014	0.724±0.022	2.130±0.142	8.610±0.469
5	1.037±0.013	0.738±0.016	2.138±0.044	8.804±0.415
6	1.076±0.026	0.751±0.009	2.191±0.035	8.186±0.051
7	1.126±0.063	0.794±0.016	2.288±0.027	8.645±0.454

Table 5.4: The average MRE for each dataset when trained on Gaussian heatmaps with different σ in equation 5.7 over three models. The best average MRE over three models for each dataset is shown in bold.

Results

Results over the validation set for models trained on Gaussian heatmaps generated with different heatmaps are shown in Table 5.4. The mean and the standard deviation of MRE over three separate models are given in the table. These results show that setting $\sigma = 3$ produced the best results on average for all datasets. Therefore, Gaussian heatmaps with $\sigma = 3$ were used for all experiments presented in

the remainder of this chapter. Similarly to the data augmentation results reported in Section 5.2.2, the standard deviation in MRE results for a specific σ frequently encompassed the MRE obtained for various different σ values. For example, the mean and standard deviation for models trained with $\sigma = 3$ heatmaps for the Hand dataset was 0.694±0.007 and the mean and standard deviation for model trained with $\sigma = 1$ heatmaps was 0.698±0.009. These standard deviation results show that there was one model trained with $\sigma = 1$ heatmaps which obtained a better result than one of the models trained with $\sigma = 3$ heatmaps. These results show that if these experiments were run again it is likely that models with different values of σ would have performed the best and, therefore, different values of σ would have been chosen.

5.2.4 Architectures

The aim of this work was not to construct new model architectures for the purpose of landmark detection but instead to show that existing, widely used, architectures could achieves SOTA results. To this end the U-Net architecture [111] and the UNet++ [149] architecture were tested.

Experiments Performed

Three models with the U-Net architecture [111] and three models with the UNet++ [149] architecture were trained over each dataset's training set and then tested on each dataset's validation set. Both these networks used a ResNet34 architecture as the encoder. These architectures were chosen because they are fast to train and can be implemented using existing libraries such as `pytorch segmentation models`.[7].

An addition experiment was conducted where the ResNet34 encoder with random weight initialisation was tested against initialising with weights transferred from the ImageNet task. This means that before these models were trained they were initialised with some object detection knowledge. Pre-trained weights can be easily implemented using the `pytorch segmentation models` library.

	MRE (mm) over the validation sets. The mean and std over three seperately trained models is given.			
Model	Cephalometric	Hand	Pelvis	Infant Hip
U-Net	1.079±0.082	0.712±0.010	2.151±0.022	8.566±0.721
UNet++	**0.983±0.020**	**0.694±0.007**	**2.094±0.086**	**8.074±0.097**

Table 5.5: This table compares the results obtained using a U-Net [111] to a UNet++ [149]. The best average MRE over three models for each dataset is shown in bold. The UNet++ outperformed the U-Net on every dataset.

		MRE (mm) over the validation sets. The mean and std over three seperately trained models is given.			
Encoder	Model	Cephalometric	Hand	Pelvis	Infant Hip
Not pre-trained	UNet++	1.362±0.139	0.821±0.040	2.414±0.031	8.901±0.388
Pre-trained	UNet++	**0.983±0.020**	**0.694±0.007**	**2.094±0.086**	**8.074±0.097**

Table 5.6: This table compares results for the UNet++ architecture with and without initialisation with pre-trained weights. An identical UNet++ with a Resnet34 [55] encoder was used in both cases, however, the latter used pre-trained weights for the Resnet34 encoder tuned on the ImageNet classification task. The best average MRE over three models for each dataset is shown in bold.

Results

Table 5.5 shows that the UNet++ performed better across all datasets when compared to the U-Net. Table 5.6 shows that initialising the encoder of the UNet++ with weights trained on ImageNet produced MRE results between 18% and 28% more accurate than that of random weight initialisation which is a remarkable improvement. Due to the superior performance obtained when using pre-trained weights, these were used for all experiments presented in the remainder of this thesis.

[7] `https://github.com/qubvel/segmentation_models.pytorch`

5.2.5 Loss Function

Binary Cross Entropy (BCE) loss and Negative Log Likelihood (NLL) loss are two loss functions used by landmark detection methods in the literature. Both produce a loss on the output of the network after a two-dimensional softmax activation function. BCE loss for a particular landmark l and pixel (i, j) is defined as:

$$\text{BCE}_l(i,j) = H_l(i,j) \cdot \log\left(S_l(i,j)\right) + (1 - H_l(i,j)) \cdot (1 - \log\left(S_l(i,j)\right)) \tag{5.9}$$

where the value $H_l(i, j)$ is the training heatmap H for landmark l at position (i, j). The value $S_l(i, j)$ is the output heatmap after being passed through a two-dimensional softmax function for landmark l at position (i, j).

NLL loss is defined as:

$$\text{NLL}_l(i,j) = -H_l(i,j) \cdot \log\left(S_l(i,j)\right) \tag{5.10}$$

For both BCE and NLL loss functions the final loss value is the mean average of the BCE and NLL values over all landmarks l for all pixels (i, j).

Experiments Performed

Three models were trained on each dataset's training set using BCE and NLL loss and the average of their results over each dataset's validation set is reported.

	MRE (mm) over the validation sets. The mean and std over three seperately trained models is given.			
Loss Function	Cephalometric	Hand	Pelvis	Infant Hip
Binary Cross Entropy	0.993±0.009	0.696±0.004	**2.052±0.055**	9.783±0.217
Negative Log Loss	**0.983±0.020**	**0.694±0.007**	2.094±0.086	**8.074±0.097**

Table 5.7: This table compares BCE loss to NLL loss. The best average MRE over three models for each dataset is shown in bold.

Results

Table 5.7 shows that NLL loss narrowly outperformed BCE loss on three of the four datasets, with the exception being the Pelvis dataset where BCE loss outperformed NLL by a negligible 0.042mm. Based on these results, and practical reasons including experimental and code efficiency, NLL was the loss function used for all experiments described in the remainder of this thesis.

5.2.6 Hyperparameter Summary

The experiments performed in Section 5.2 provide useful insights into what hyper-parameters are most impactful for landmark detection model training on medical imaging datasets. These findings are useful to the community, since not all have been analysed in existing literature. Whilst other publications have reported the results that different network architectures obtained over a dataset, none have reported the effect that using pre-trained weights had on landmark detection accuracy. In addition, no existing publication provides results for networks trained with different data augmentation methods to show if any are particularly effective. Finally, the effect that the σ value of the training Gaussian heatmaps has on model performance is not well documented in the literature and, in the vast majority of papers, is simply stated without explanation. An exception to this is in Payer *et al.* [105], however in the method described by this paper σ is a learned parameter.

The findings in this section can be summarised as follows: data augmentation improved results for all datasets. However, it is hard to analyse which specific augmentations produced the biggest improvement in MRE. The UNet++ achieved consistently better results than the U-Net across all datasets. Initialising the encoder of the UNet++ with weights trained on ImageNet produced MRE results between 18% and 28% more accurate than that of random weight initialisation which is a significant improvement. Using Gaussian training heatmaps with $\sigma = 3$ was found to produce the best results for the datasets in this thesis however these values could be different if these experiments were performed again. Finally, the NLL loss function achieved a small performance boost over BCE loss on three out of four of the datasets.

5.3 Publicly Available Dataset Results

Section 5.2 described how hyperparameters were optimised. Based on the findings presented in that section it was concluded that the UNet++ architecture should be used with pre-trained weights and trained using NLL loss. In addition, it was established that all data augmentation experimented with should be applied during training and that Gaussian training heatmaps with $\sigma = 3$ should be used for the Cephalometric and Hand datasets. With these hyperparameters selected the final models were trained and tested.

For the Cephalometric dataset, the three final models were trained on the whole training set, as opposed to 80% of it in Section 5.2. These models were then tested on Test Sets 1 and 2 and their average results are reported in Section 5.3.1.

The Hand dataset has three pre-defined training and test set allocations given by the original authors of the dataset [103]. There is no overlap in the images contained within each of the training and test set allocations. Three models were trained on each of these three training sets and then each was tested on the corresponding test set. The results on each test set were averaged and reported in Section 5.3.2.

5.3.1 Cephalometric Dataset Results

Method	Test Set 1					Test Set 2				
	MRE (mm)	SDR (%)				MRE (mm)	SDR (%)			
		2mm	2.5mm	3mm	4mm		2mm	2.5mm	3mm	4mm
Ibragimov et al[58]	1.87	71.70	77.40	81.90	88.00	-	62.74	70.47	76.53	85.11
Lindner et al[81]	1.67	74.95	80.28	84.56	89.68	-	66.11	72.00	77.63	87.42
Arik et al[8]	-	75.37	80.91	84.32	88.25	-	67.68	74.16	79.11	84.63
Yao et al[141]	1.24	84.84	90.52	93.75	97.40	1.61	71.89	80.63	86.36	93.68
Chen et al[19]	1.17	86.67	**92.67**	**95.54**	**98.53**	1.48	75.05	82.84	88.53	**95.05**
Zhong et al[148]	**1.12**	86.91	91.82	94.88	97.90	1.42	76.00	82.90	88.74	94.32
Proposed Method	**1.12**	**87.47**	92.25	95.33	98.05	**1.41**	**76.40**	**84.23**	**89.28**	94.84

Table 5.8: Results for the hyperparameter tuned UNet++ trained on the full Cephalometric training set. An average of three separately trained models is reported. Lower MRE is better, higher SDR percentages are better. All results from past methods are copied from literature and not reproduced.

The average MRE and SDR results the three separately trained models achieved over Test Set 1 and Test Set 2 are reported in Table 5.8. The hyperparameter-tuned

UNet++ achieved SOTA MRE accuracy and the only past approach which achieved superior SDR results is Chen et al [19]. This paper achieved an SDR less than 1% higher over the 2.5mm, 3.0mm and 4.0mm thresholds.

Previous approaches that used the Cephalometric dataset for experimentation were not always robust in their setup. It is not always clear what validation set they used and repeat experiments were not completed. There was no data leakage between training and tests set in the experiments described in this thesis and three separate models were trained and average results reported. This makes the results reported in this section more reliable than those in previous literature.

Examples of landmark predictions for images in Test Set 2 are shown in Figure 5.5. Figure 5.5A shows a cropped region containing four landmarks all predicted within 2mm of the ground truth. The majority of the landmarks were similarly accurately predicted as shown in Table 5.8 by the SDR for 2mm being 87.47% and 76.40% for Test Set 1 and 2. Figure 5.5B and Figure 5.5C show examples where a landmark was inaccurately detected with an error of over 6mm. Figure 5.5B shows a high radial error for Landmark 16 on the chin of the person captured. In Test Set 2 the ground truth for this point is consistently placed higher than in the training set and Test Set 1 as discussed in Section 3.2. For this reason, all predictions for this landmark had a systematic radial error of at least 3mm in Test Set 2 which caused an increase in the MRE from 1.11mm over Test Set 1 to 1.42mm over Test Set 2. For this reason, to improve results Landmark 16 should be relabelled either in the training set and Test Set 1 or in Test Set 2 so that it has a consistent placement.

Figure 5.5C shows the largest prediction error in Test Set 2 caused by predicting Landmark four, by the temple, incorrectly. This landmark's ground truth is marked slightly lower in this example than in other images. In this case, an additional expert clinician is needed to confirm whether this is a labelling error or something the model has not correctly detected. Further landmark detection examples are given in Figure B.1 to Figure B.3 in Appendix B.

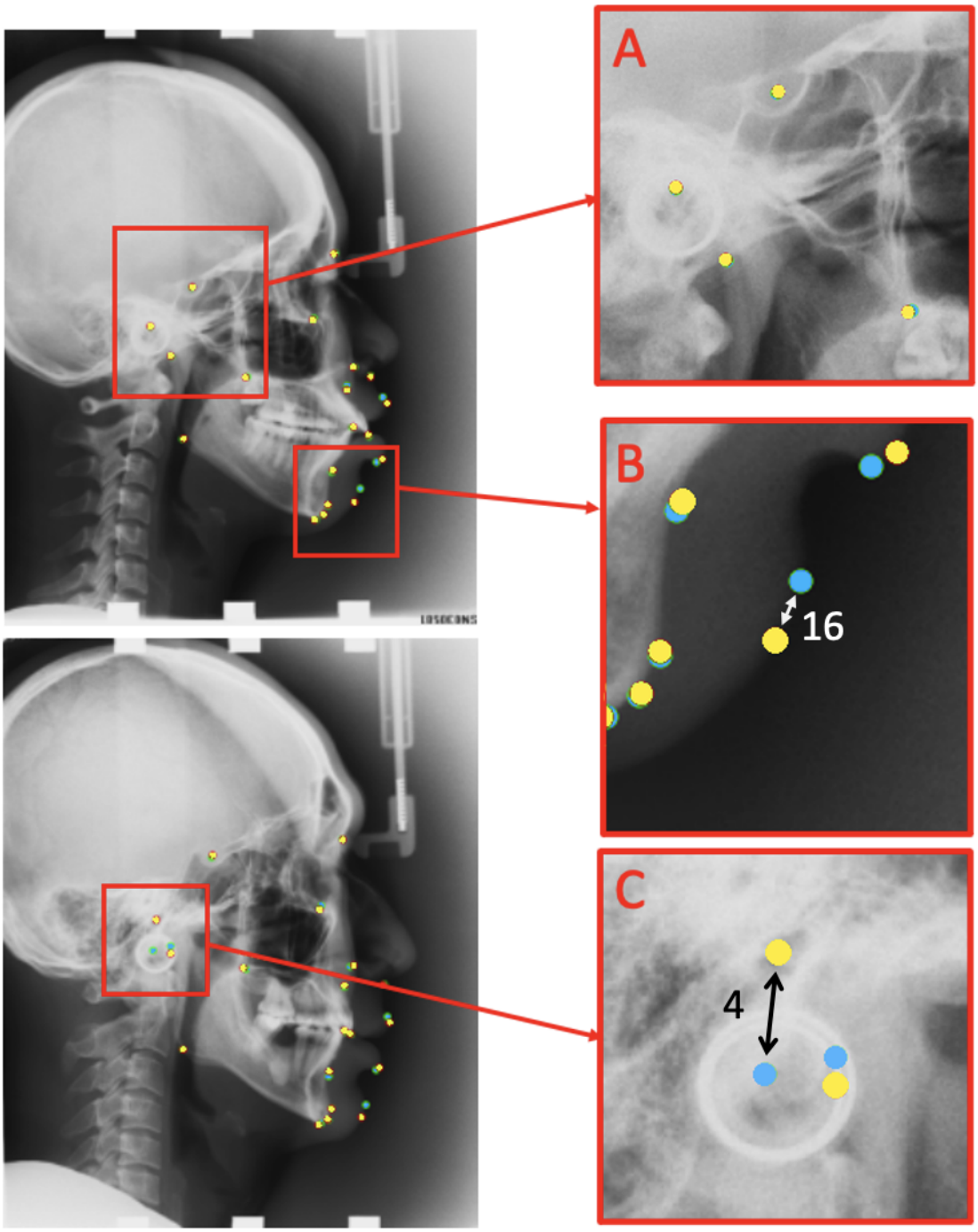


Figure 5.5: Two full images of localisations on the Cephalometric dataset are shown in this figure with three expanded regions shown in A, B and C. The blue dots are the ground truth and the yellow dots are the model's predictions. Four accurate localisations are shown in A. A large radial error between the ground truth and the prediction for the landmark is shown in the centre of region B. This is Landmark 16 and, as discussed in Section 3.2, this radial error is due to inconsistent labelling. C shows the largest localisation error in Test Set 2 where Landmark 4 has been predicted significantly above its ground truth placement. Expert guidance is required to know if this is a labelling mistake or a prediction error.

5.3.2 Hand Dataset Results

Method	MRE (mm)	SDR (%)		
		2mm	4mm	10mm
Payer et al (2016) [103]	1.13	87.60	98.66	99.96
Ebner et al [36]	0.97	91.60	97.84	99.31
Stern et al [123]	0.80	92.20	98.45	99.95
Urschler et al [131]	0.80	92.19	98.46	99.95
Lindner et al [81]	0.85	93.68	98.95	99.94
Payer et al (2019) [104]	**0.66**	94.99	99.27	**99.99**
Proposed Method	0.67	**95.97**	**99.48**	**99.99**

Table 5.9: Results for the hyperparameter tuned UNet++ compared with past literature. Lower MRE is better, higher SDR percentages are better. Results are the average of nine runs with three individual models trained with the same hyperparameters for each of the three training sets proposed by Payer et al (2016) [103]. The SDR ranges of 2mm, 4mm and 10mm were chosen because they are used in previous literature. All results from past methods are copied from literature and not reproduced.

Results for the Hand dataset are shown in Table 5.9. The hyperparameter-tuned UNet++ model achieved higher SDR statistics than all previous literature and achieved only 0.01mm higher than the SOTA MRE. Figure 5.6A shows an example where all landmarks were located within 2mm of the ground truth. The majority of landmarks were accurately predicted in this dataset as shown by the SDR of 2mm being 95.97% over the test set. Figure 5.6B shows a Landmark 19 prediction that had a radial error of 3mm, which was one of the highest in the test set. The ground truth has been placed to the right of its correct position in the centre of the gap between the finger bones. The radial error for Landmark 7 in the centre of Figure 5.6C was the highest radial error in the dataset at 9mm. This error has been caused by the model which incorrectly predicted that the landmark should be on the corner of a different bone which is higher up the palm.

Despite this error, the MRE for landmark predictions in the Hand dataset were lower than that of the Cephalometric dataset. This is due to the large number of images in the Hand dataset's training sets and consistent landmark definitions. Future work should obtain intra and inter-operator variability for the ground truth such that the bound for model accuracy can be determined for this dataset. Further landmark detection examples are given in Figure B.4 to Figure B.6 in Appendix B.

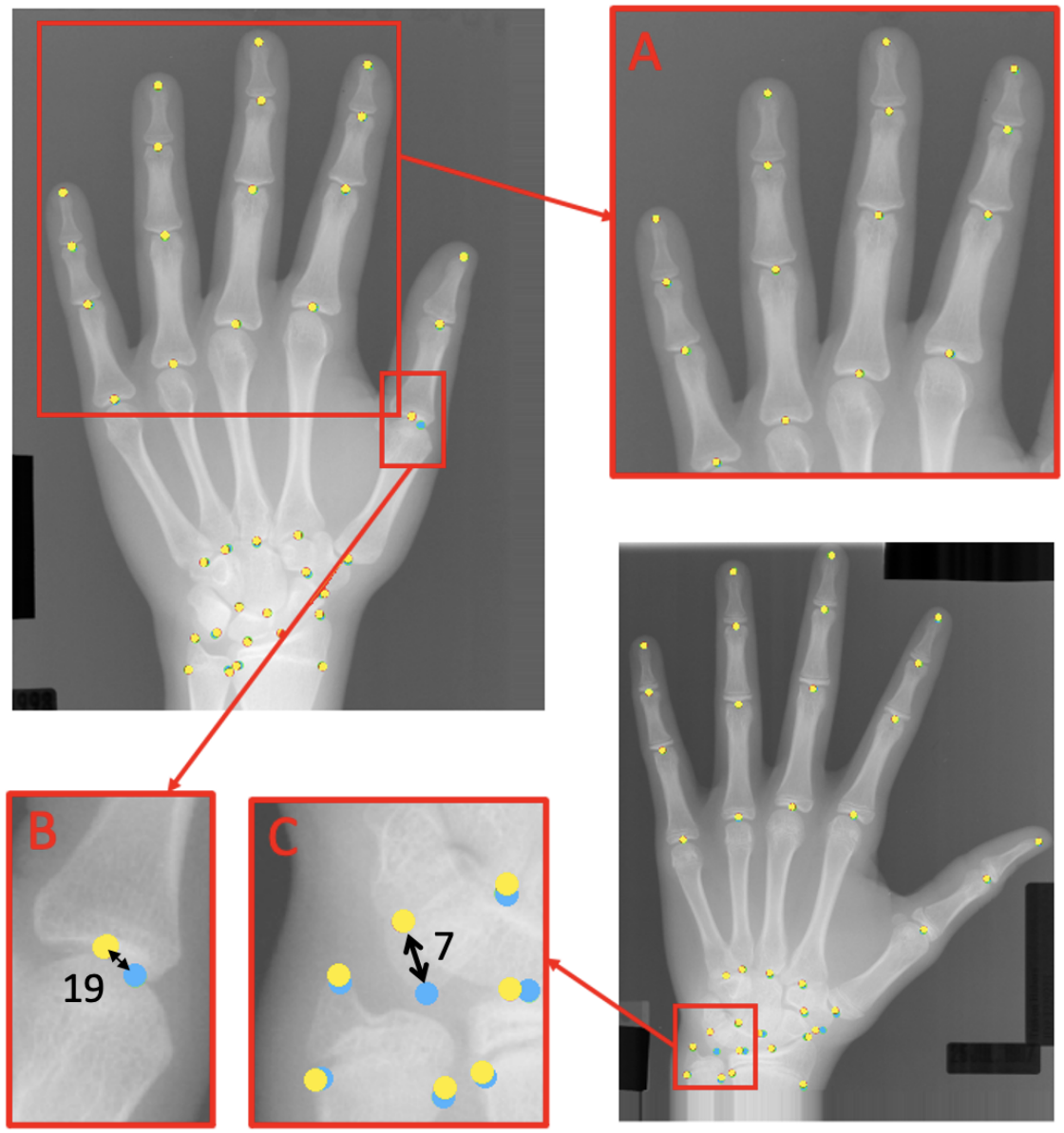


Figure 5.6: Two full images of localisations on the Hand dataset are shown in this figure with three expanded regions shown in A, B and C. The blue dots are the ground truth and the yellow dots are the model's predictions. All localisations shown in A are highly accurate which is very common in this dataset as shown by the high SDRs in Table 5.9. B shows a difference between the ground truth and the prediction for landmark 19. As discussed in Section 3.3 this is due to inconsistent labelling. C shows the largest localisation error in the test set which is caused by the model erroneously placing Landmark 7 at the corner of the wrong bone.

5.3.3 Public Datasets Results Conclusion

The UNet++, with tuned hyperparameters, achieved either superior MRE and SDR results than past approaches, or results within 0.01mm of the SOTA, for both the Cephalometric and Hand dataset. This outcome is significant for the community because the variations among the reported results from different methods and publications are only on the order of hundredths of a millimetre. Given this negligible performance difference, the emphasis shifts to the heightened importance of ensuring the experiment's reproducibility and robustness. The method proposed here is easily reproducible because it used the existing UNet++ architecture which can be implemented using open source libraries such as **`pytorch segmentation models`**. This is in contrast to the complex two-stage methods proposed by Zhong *et al.* [148] and Chen *et al.* [19] which would have to be implemented by other researchers from scratch. Additionally, this result further proves the phenomenon described in the No New U-Net paper by Isensee *et al.* [59] that existing networks can achieve SOTA performance when their hyperparameters are tuned correctly.

In regards to the performance of the models, the majority of the landmark detection errors were a consequence of ground truth labelling errors or ambiguity, for example, Landmark 16 in the Cephalometric dataset. Section 3.2 stated that the inter-rater variability between the two clinicians who labelled the Cephalometric dataset was 1.38mm. When this is compared to the 1.41mm MRE achieved by the model described in Section 5.3.1 it is clear that the model is close to achieving results within the range of clinicians. This suggests that the accuracy of landmark detection methods on these datasets has reached a stage where it is bounded by the clarity of the landmark definitions instead of the computing methodology although inter-rater statistics need to be gathered for the Hand dataset to confirm this theory.

It is important to note that these datasets were collected specifically for research purposes. This means that it is likely the images were more precisely collected, and contained fewer patients with unusual presentations, than a dataset of images collected during routine hospital practice. Future work should focus on obtaining and sharing datasets which represent images collected during routine practice.

5.4 Pelvis Dataset Results

5.4.1 Experiments Performed

In Section 5.2 three UNet++ models were trained using pre-trained weights, NLL loss, all data augmentations methods active and using Gaussian heatmaps with $\sigma = 3$ for the Pelvis dataset. These models were trained on the Pelvis training set and achieved the lowest MRE results of all the models tested over the Pelvis validation set. Of those three models trained the one with the lowest MRE over the validation set was then tested on the Pelvis test set which was selected by the KS test method as described in Section 4.4. The landmark prediction results, as well as the α-angle, LCE-angle and FAI diagnosis results this model achieved over the test set are reported in the following Sections 5.4.2 to 5.4.4.

5.4.2 Landmark Detection Results

Landmark	Radial Error (mm)	
	Mean ± Std	Median
Lateral Acetabulum Point	1.319±0.810	1.114
Femoral Head Center	1.415±0.886	1.275
Neck Orientation Point	2.427±1.588	2.066
Lateral Cam Point	3.871±5.400	1.870
All Landmarks	2.258±3.041	1.576

Table 5.10: This table reports the mean, standard deviation and median radial error the model achieved over the landmarks in Pelvis test set. The radial errors for the left and right side landmarks are combined into a single list upon which the mean, standard deviation and median are calculated.

The results reported in Table 5.10 show that the MRE for all landmarks predicted was 2.258mm. Examples of landmarks which were detected within 2.258mm are shown in Figure 5.7. Further landmark detection examples are given in Figure B.7 and Figure B.8 in Appendix B. The results in Table 5.10 show that the lateral acetabulum points were the most accurately detected, followed by the femoral head centres, then neck orientation points, and finally, lateral cam points. For reference on the definitions of these landmarks see Section 3.4. These results show that

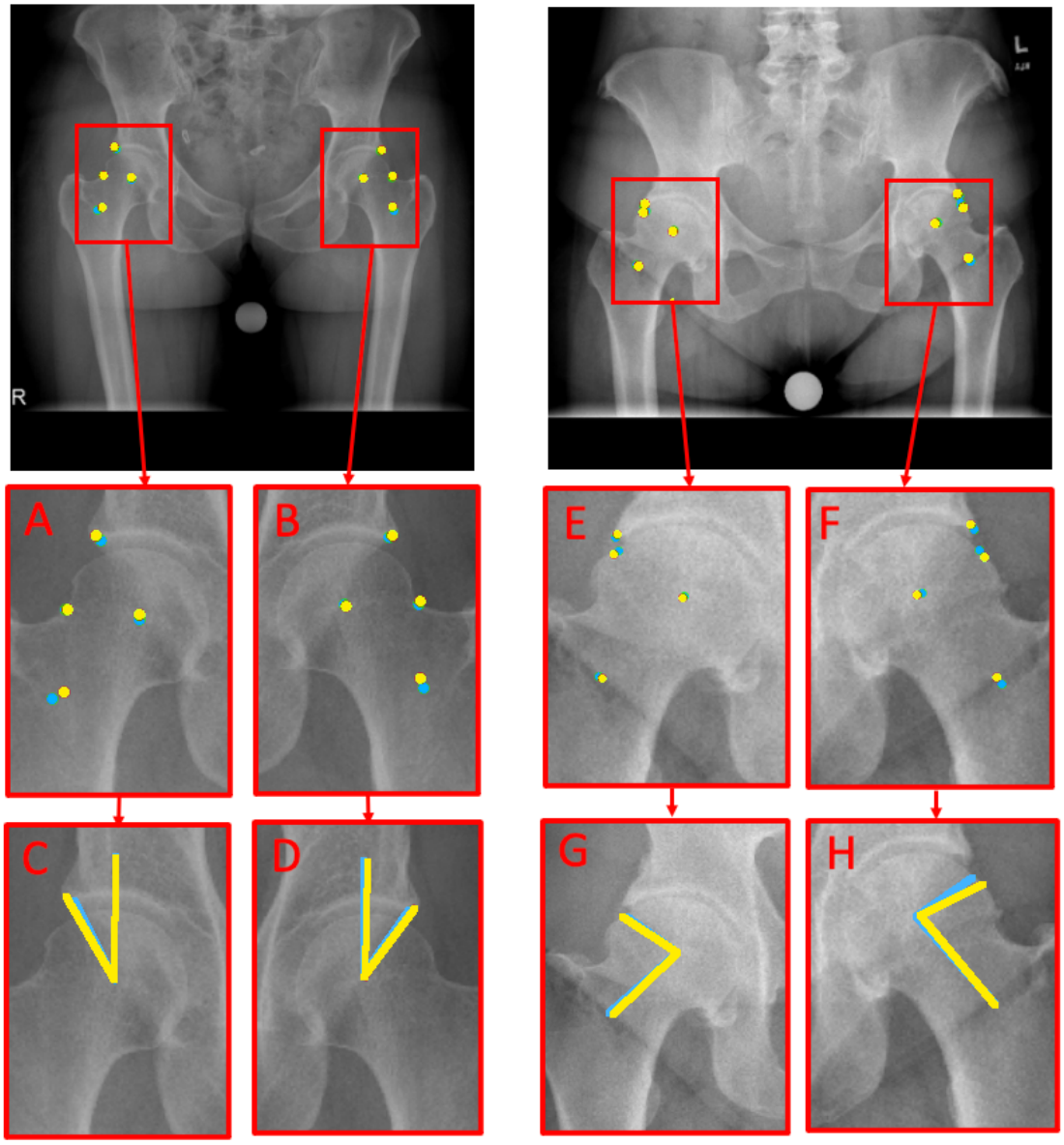


Figure 5.7: Two images from the Pelvis dataset are shown in this figure with four expanded regions shown in A, B, E and F. The relevant ground truth and predicted landmarks in A and B are connected to show the LCE-angle comparison in C and D. Similarly the relevant landmarks are connected in E and F to show the α-angle comparison in G and H. The blue dots and lines are the ground truth and the yellow dots and lines are the model's predictions. All predictions shown in this figure were highly accurate which led to less than 3° difference between all the ground truth and predicted angles shown in the bottom row. This is particularly impressive because E and F capture images of femurs with pistol-grip type cam impingements.

landmarks which are defined to be on the corners of anatomical structures, such as the lateral acetabulum point, are more accurately predicted than those located within bones and defined using more complicated rules. For example, where the centre of the femoral head is located is ambiguous because the femoral head is not a perfect circle. Similarly, the neck orientation point is defined as a point along the neck axis of the femur but the orientation of the neck axis, and how far along the axis the landmark should be placed, are ambiguous. It therefore makes intuitive sense why predictions for these landmarks have higher radial errors than the lateral acetabulum point. It is likely a human rater would also have a greater variability when placing these landmarks as their exact locations are ambiguous.

This is also the reason that the lateral cam point was the hardest for the model to predict. Section 3.4.2 discussed how the clinician's α-angle, which relies on the lateral cam point placement, can differ from the geometrically correct placement. For these reasons, it is often ambiguous where the lateral cam point should be placed. An additional reason why the lateral cam point is less accurately detected than the other landmarks is because the placement of this landmark depends largely on whether a small bump is present on the femur. If a small bump is present then the lateral cam point should be much higher up the femor however the model does not always spot these small bumps. This is supported by the results in Table 5.10 which show that the MRE is significantly higher than the median radial error for the lateral cam point. This shows that the lateral cam points were accurately detected on (median) average, however, some very large errors on a small number of landmarks skewed the mean. These large errors were caused by the model failing to spot small bumps on the femur head. Examples of deformities which were difficult to spot on the femoral head are shown in Figure 5.8B and D. These are discussed more thoroughly in Section 5.4.3.

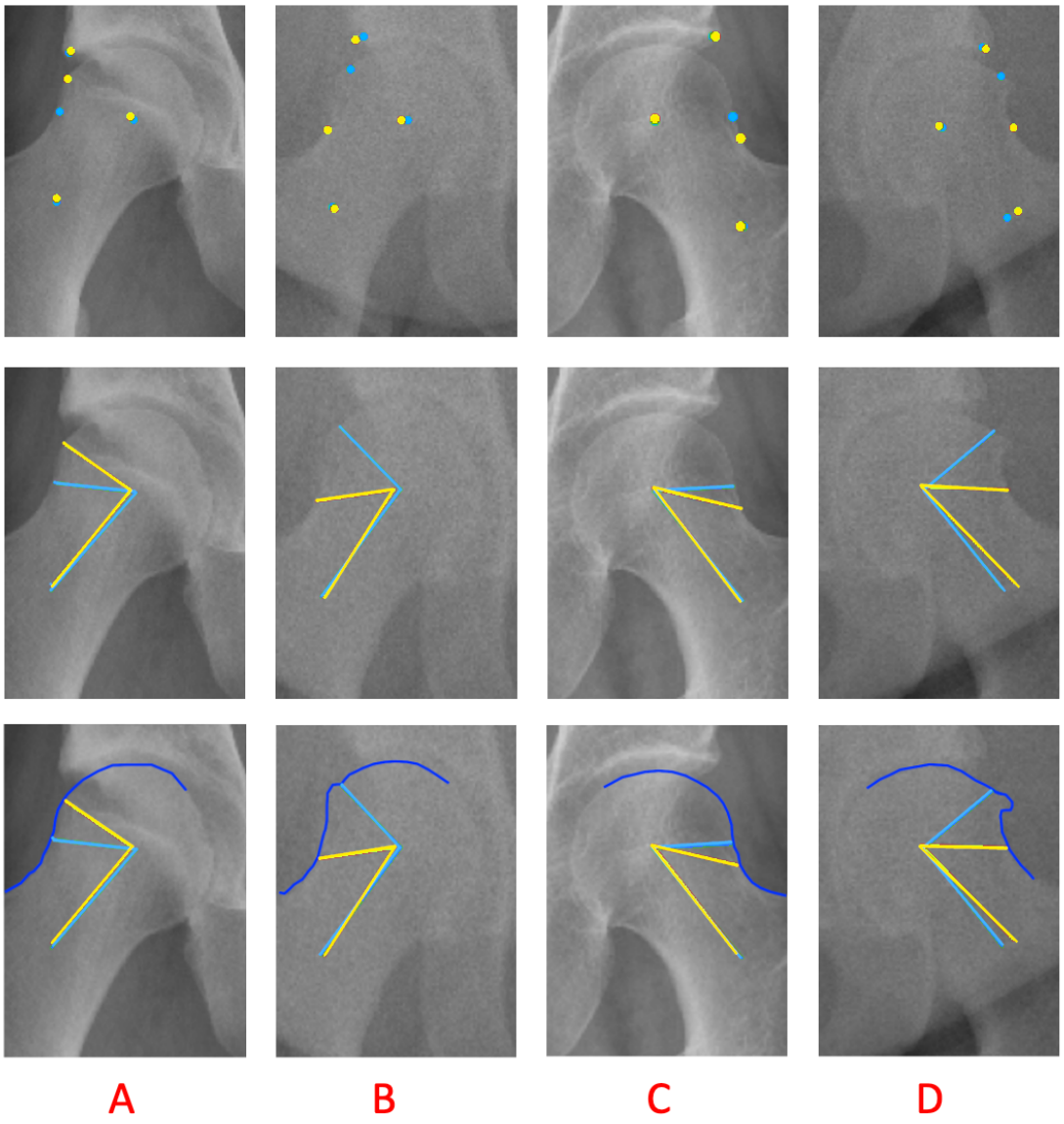


Figure 5.8: The figure shows four examples of inaccurate α-angle predictions. The first row shows the ground truth landmark positions in blue and the predicted landmark positions in yellow. The second row shows their corresponding α-angle measurements in the corresponding colours. The third row adds a blue contour highlighting the true shape of the femoral head in each case. Example A shows the only example in the test set where the model predicted a larger α-angle than the ground truth. This was a model misprediction because the femoral head is smooth and a smaller α-angle should have been predicted. In the hips displayed in B and D, there exists a small bump on the femoral head which the model did not detect. These examples show the largest α-angle errors in the test set. In C a 10° error was made which caused the patient to be mispredicted as having no cam impingement.

5.4.3 LCE and Alpha Angle Results

Measurement	Degrees Difference (°)		ICC
	Mean ± Std	Median	
LCE-angle	2.241° ± 1.768°	1.855°	0.919
α-angle	8.170° ± 12.320°	3.461°	0.563

Table 5.11: This table reports the LCE-angle and α-angle results the model achieved over the images in Pelvis test set. The mean, standard deviation, and median in the table are computed over the difference between the ground truth angles and the predicted angles. The ICC statistic is a measure of how correlated the ground truth and predicted angles were. Both the ground truth angles and the predicted angles are produced by connecting the relevant ground truth and predicted landmarks with line segments and calculating the angles between them as shown in Figure 5.7. Results are an average between left and right hips.

Table 5.11 shows the mean, standard deviation and median angle difference between the predicted angles (derived from the predicted landmarks) and the ground truth angles (derived from the ground-truth landmarks). The LCE-angle was more accurately predicted than the α-angle on average with a mean angle difference of 2.241° and 8.170° respectively.

In addition, the ICC is given between the two sets of angles (see Section 5.1.3 for ICC definition). The ICC is a statistic that measures how correlated two sets of values are, with one being a perfect correlation. Agricola *et al.* [3] found an ICC between clinicians of 0.73 for measurement of the α-angle. The ICC score between the model's α-angle and the ground truth was 0.563. These results are aligned to the landmark prediction results reported in Section 5.4.2. Those results showed the lateral cam point was difficult to detect in agreement with the clinician and this error was propagated through to the α-angle measurement. Comparatively, the model's LCE-angle predictions achieved an ICC score of 0.919 against the ground truth, showing it can be predicted with much more consistent agreement with the clinician.

Four examples of inaccurate α-angle predictions are shown in Figure 5.8. Figure 5.8A shows the *only* example in the test dataset where the model predicted a larger α-angle than the ground truth. This is an example of a model prediction mistake because the femoral head appears smooth in this image and the lateral cam

point should have been predicted lower to produce a lower α-angle. Figure 5.8B and Figure 5.8D show examples where the model did not spot a small bump on the femoral head which caused an error above 30.0° in the α-angle measurement in both case. In both images the bump on the femoral head is small and the image quality is fuzzy. These examples were the largest α-angle errors in the test set and they show that the model did not perform well on images of sub-optimal quality and did not detect small bumps on the femoral head accurately.

A reason for the latter is that the model took as input the full X-ray image and was not given any information about where in the image it should pay attention to when predicting the lateral cam point. This is a problem because the small bumps on the femoral head are only represented by a very small number of pixels relative to the whole image and it is likely that the model did not learn the significance of these pixels during the training process. Future work could add an attention mechanism to the model to enable it to focus on the femoral head. This method would first detect a Region of Interest (RoI), which is the area of an image the model needs to pay attention to, and then feed it into a landmark detection model. This has the additional benefit that the RoI could be cropped from the original image, not the downsampled version, which would increase the resolution of key features in the image as discussed in Section 5.1.1. There are several methods for detecting a RoI. Firstly, a separate CNN model could be trained to detect the RoI. Alternatively, the GRAD-CAM [116] method could be applied to find which parts of an image the model is paying attention to when making landmark predictions. These parts could then be used as a RoI. Finally, the area covered by the landmark's output heatmap (see Chapter 6) could be used as a RoI. In addition, it could also be possible to create an end-to-end trainable solution where an attention mechanism which uses queries, keys and values [137] is built into a landmark detection CNN. This solution requires more in-depth research.

Further examples of LCE and α-angle model predictions are given in Appendix C.

5.4.4 FAI Diagnosis Results

		Predicted Pincer FAI	
		Positive	Negative
True Pincer FAI	Positive	26	6
	Negative	3	77

Accuracy: 91.96% Sensitivity: 81.25% Specificity: 96.25%

Table 5.12: This table shows the confusion matrix for pincer impingement type FAI comparison between model and clinician where an LCE angle threshold of 40° was used over the test set. Hips with a higher value than the threshold were positively diagnosed.

		Predicted Cam FAI	
		Positive	Negative
True Cam FAI	Positive	14	12
	Negative	2	84

Accuracy: 87.50% Sensitivity: 53.85% Specificity: 97.67%

Table 5.13: This table shows the confusion matrix for cam impingement type FAI comparison between model and clinician where an α-angle threshold of 65° was used over the test set. Hips with a higher value than the threshold were positively diagnosed.

Tables 5.12 and 5.13 show results for diagnosing pincer type and cam-type FAI automatically. Pincer impingement was diagnosed when the predicted LCE-angle was above 40°. The model achieved an accuracy of 91.96% when detecting pincer impingement, with a sensitivity of 81.25% and a specificity of 96.25%. These results show that pincer impingement type FAI can be reliably detected in Pelvis X-rays using this method. The misdiagnoses the model made for LCE-angle were all on patients with LCE-angles within 3° of the threshold of 40°. This shows that even with a highly accurate model mispredictions will occur when there is a single threshold for a positive diagnosis. This is because an angle prediction could only be a single degree different from the ground truth angle but if that prediction is the other side of the threshold then it is still counted as a misdiagnosis. Overall the results for LCE-angle prediction are good and to date, this is the only automatic LCE-angle prediction approach in the literature.

Cam impingement was diagnosed when the predicted α-angle was above 65°. The model achieved an accuracy of 87.50%, with a sensitivity of 53.85% and a

specificity of 97.67%. This shows that the model was very accurate at classifying patients with no cam impingement however it often underestimated the α-angle for patients with cam impingement which led to nearly half of these patients being incorrectly diagnosed with no cam impingement (false negative). The reason for this is that the algorithm often missed small bumps on the femoral head as discussed in Section 5.4.3 and shown in Figure 5.8B and D. In addition, similarly to LCE-angle, sometimes patients with α-angles very close to the threshold of 65° were misdiagnosed even though the α-angle error was small. Figure 5.8C shows a relatively small α-angle error however because the ground truth α-angle is 70° but the predicted α-angle was 60° this patient was incorrectly diagnosed as having no cam impingement. This shows that small errors close to the cam-impingement threshold of 65° caused some of the incorrect classifications reported in Table 5.13.

5.5 Infant Hip Dataset Results

5.5.1 Experiments Performed

In Section 5.2 three UNet++ models were trained using pre-trained weights, NLL loss, all data augmentations methods active and using Gaussian heatmaps with $\sigma = 3$ for the Infant Hip dataset. These models were trained on the Infant Hip training set and achieved the lowest MRE results of all the models tested over the Infant Hip validation set. Of those three models trained the one with the lowest MRE over the validation set was then tested on the Infant Hip test set which was selected by the KS test method as described in Section 4.4. The landmark detection results, as well as the α-angle, β-angle and DDH diagnosis results this model achieved over the test set are reported in the following Sections 5.5.2 to 5.5.4.

5.5.2 Landmark Detection Results

Landmark	Radial Error (pixels)	
	Mean ± Std	Median
Ilium Point 1	9.310±6.970	7.307
Ilium Point 2	7.261±5.805	5.269
Turning Point	5.590±5.947	4.070
Lower Limb Point	5.397±12.202	3.346
Labrum Point	7.328±5.602	5.842
All Landmarks	8.998±6.865	4.936

Table 5.14: This table reports the mean, standard deviation and median radial error the model achieved over the landmarks in Infant Hip test set.

The results for each landmark in the Infant Hip dataset are given in Table 5.14 and show that the MRE for all landmarks in the test set was 8.998 pixels. The MRE is given in pixels here because no pixel-to-mm ratio was given with the images in the dataset. The results show that ilium point 1 predictions had the most radial error followed by the labrum point. For definitions of these landmarks see Section 3.4. Similarly to the landmark prediction errors seen in Section 5.4.2 for the Pelvis the reason these landmarks have higher prediction radial errors is because of unclear landmark definitions. In the case of the ilium point 1, there is no clear definition of

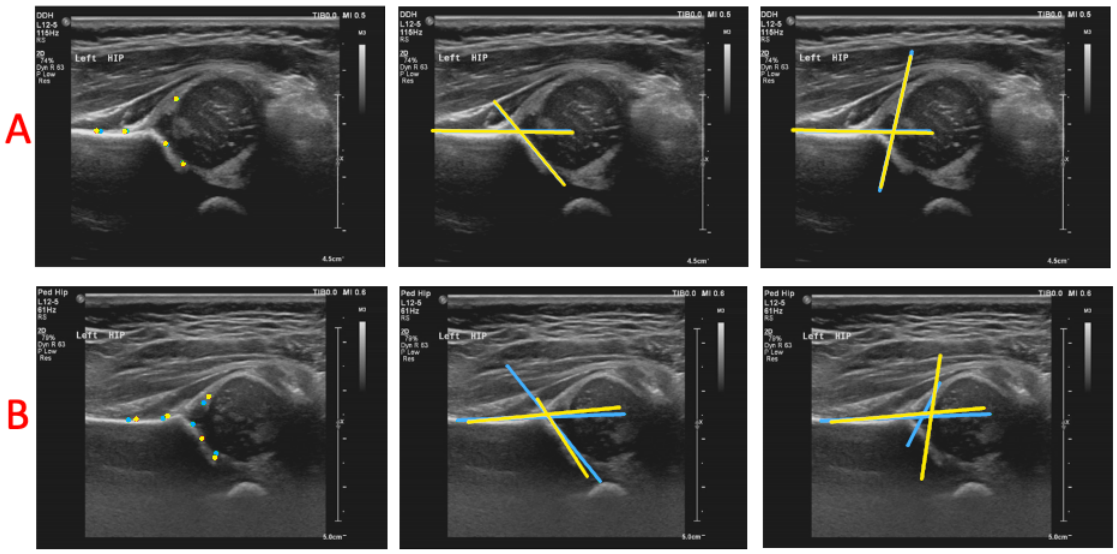


Figure 5.9: This figure shows two images from Infant Hip test set. In the first column the ground truth and predicted landmarks are shown. In the second column, the ilium points and the turning and lower limb points are connected in line segments to show the α-angle. In the third column, the turning and labrum points are connected to find the β-angle. The ground truth landmarks and lines are shown in blue and the predicted landmarks and lines are shown in yellow. A shows an image where all landmarks were predicted accurately and both α-angle and β-angle were within 2° of the clinician's measurement. B shows the most inaccurate landmark predictions in the test set with all landmarks, including the turning point which had an error of 35 pixels, being inaccurately detected leading to an α-angle error of 7.99° and a β-angle error of 14.76°.

how far this landmark should be placed along the ilium bone. The labrum point also does not have a strong definition, it is described as being in the centre of the labrum cartilage however the borders of the labrum can be hard to distinguish and the centre of an irregular shape is ambiguous.

Two examples of landmark predictions and the α-angles and β-angles calculated from them are shown in Figure 5.9. Figure 5.9A shows an example where all landmarks were predicted accurately leading to an α-angle and β-angle prediction within 2° of the ground-truth. Figure 5.9B shows the least accurate landmark predictions in the test set where all landmarks were detected with greater than 10 pixels of error, including the turning point which had a radial error of 35 pixels. Further landmark detection examples are given in Figure B.9 and Figure B.10 in Appendix B.

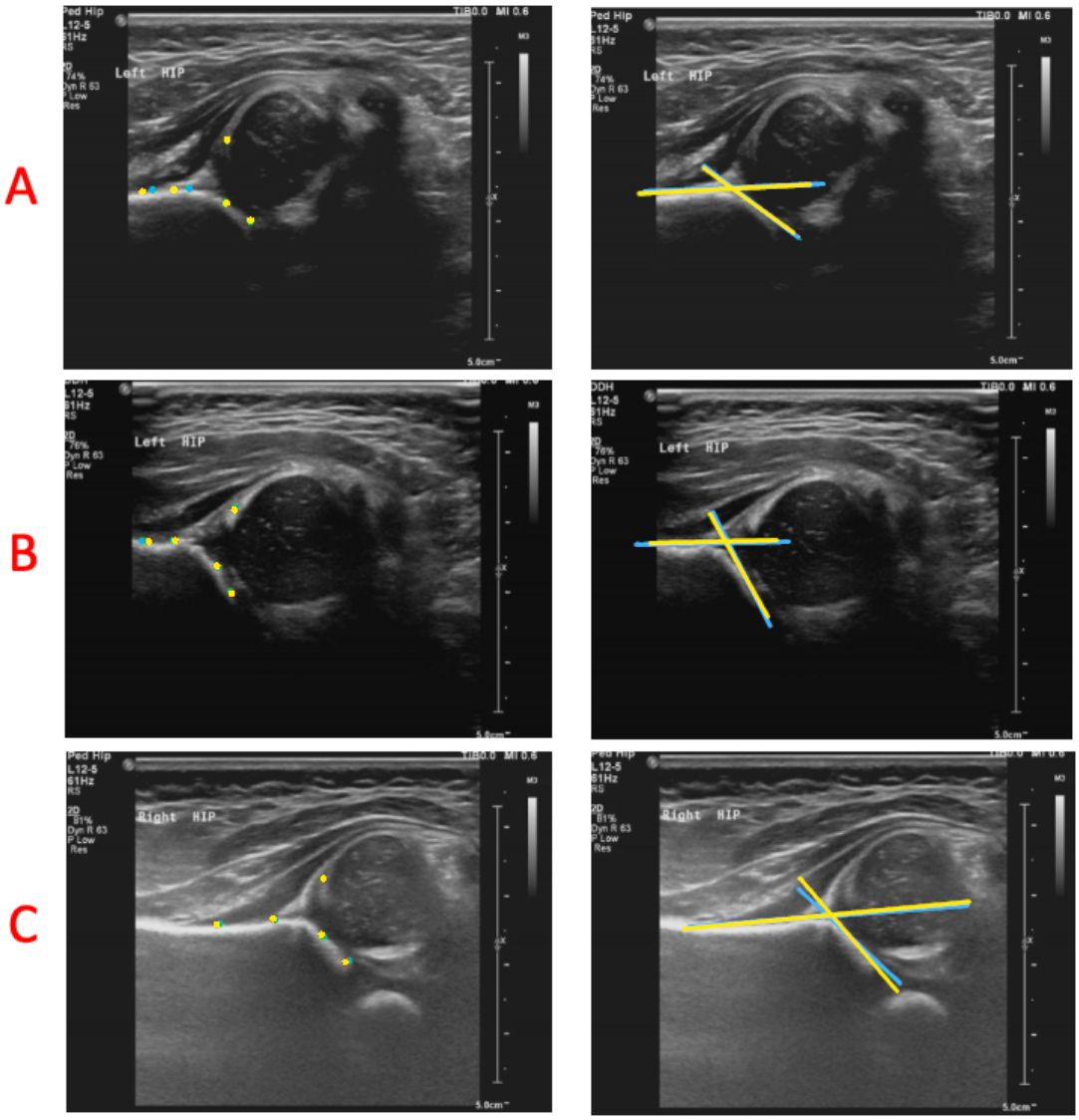


Figure 5.10: This figure shows three Infant Hip images from the test set. In the first column ground truth and predicted landmarks are shown. In the second column, their associated α-angle measurements are shown. The ground truth landmarks and lines are shown in blue and the predicted landmarks and lines are shown in yellow. A shows an example of a dysplastic hip with an α-angle $< 50°$, which was correctly identified by the model. B shows an example of a healthy hip which was correctly identified by the model. C shows an example where the model predicted the image as non-dysplastic when it is dysplastic. In this example, all landmarks were predicted within four pixels of the corresponding ground truth which demonstrates that even accurate predictions can cause a misclassification in diagnosis when there is a single threshold value for a positive diagnosis.

5.5.3 Alpha and Beta Angle Results

Measurement	Degrees Difference (°)		ICC
	Mean ± Std	Median	
α-angle	$3.167° \pm 2.536°$	$2.503°$	0.873
β-angle	$4.512° \pm 3.800°$	$3.298°$	0.883

Table 5.15: This table reports the α-angle and β-angle results the model achieved over the images in Infant Hip test set. The mean, standard deviation, and median in the table are computed over the difference between the ground truth angles and the predicted angles. The ICC statistic is a measure of how correlated the ground truth and predicted angles were. Both the ground truth angles and the predicted angles are produced by connecting the relevant ground truth and predicted landmarks with line segments and calculating the angles between them as shown in Figure 5.9.

Table 5.15 shows the mean, standard deviation and median angle difference between the predicted angles (derived from the predicted landmarks) and the ground truth angles (derived from the ground truth landmarks for the Infant Hip test dataset. The model achieved a mean angle difference of 3.167° and 4.512° for the α-angle and β-angle respectively. In addition, the ICC is given between the two sets of angles (see Section 5.1.3 for the definition of ICC). The ICC scores were 0.873 and 0.883 for the α-angle and β-angle respectively. These results show that there is a strong correlation between the predicted and ground truth angles although unfortunately there are limited clinician intra-rater and inter-rater variability ICC scores for these measurements to compare against.

Notably, the high radial errors for the ilium point 1 and the labrum point, shown in Table 5.14, did not translate into large α-angle differences. This is because, although these landmarks had high prediction errors in some cases, the landmarks were still predicted on the same axis as the line segment used in the α-angle calculation. As the ilium points were mispredicted along the ilium bone the α-angle predictions were still close to the ground truth. This demonstrates that the radial error of a landmark prediction is a reductive measurement and does not necessarily translate into bad performance on downstream tasks. Further examples of α-angle and β-angle model predictions are given in Appendix D.

5.5.4 DDH Diagnosis Results

		Predicted DDH	
		Positive	Negative
True DDH	Positive	12	7
	Negative	1	27

Accuracy: 82.98% Sensitivity: 63.16% Specificity: 96.42%

Table 5.16: This table shows the confusion matrix for DDH diagnosis between model and clinician using an α-angle threshold of 50° over the test set, where hips with a value lower than the threshold are positively diagnosed.

As discussed in Section 3.5.2, DDH is diagnosed by evaluating only the α-angle. In this work an α-angle of less than 50° indicates DDH. Table 5.16 shows that the model achieved an accuracy of 82.98% when diagnosing DDH, with a sensitivity of 63.16% and a specificity of 96.42%. These results show that overall the accuracy of this method was good, although it tended to overestimate the α-angle and thus often accurately classified healthy patients but missed DDH in seven out of the nineteen cases (36.84%). It is worth noting that the test set for the Infant Hip dataset is only 47 scans. A larger study should be completed to draw definitive conclusions.

Three examples of α-angle detection are shown in Figure 5.10. Figure 5.10A shows an example of a dysplastic hip with an α-angle less than 50°. Despite the ilium points each having a localisation error greater than ten pixels the α-angle was correctly calculated within 2° for this image. Figure 5.10B shows an example of a non-dysplastic hip with an α-angle $> 50°$, which was correctly identified by the model with all landmarks localised within four pixels. Figure 5.10C also shows an example where all landmarks were predicted within four pixels. However, despite the accurate landmark predictions this is an example of an image where the model incorrectly predicted it as non-dysplastic when it is dysplastic. This happened because the predicted α-angle was 48° but the ground-truth α-angle is 51°. Similarly to the α-angle and LCE-angle results in Section 5.4.3, these results show that when a defined threshold value is used to indicate a positive diagnosis there will always be mispredictions even when the landmark accuracy for a model is very high.

5.6 Discussion and Conclusion

5.6.1 Advantages of the Technology for Clinics

Introducing automatic landmark detection technology to clinics would have several benefits. Firstly this technology will help clinicians triage patients more accurately by improving the way patient images are assessed. Currently, measurements in clinics are often not being taken in a standardised way, not as accurately as they could be, or not being taken at all. This is because they are taken using simple drawing and measurement tools in the hospitals' PACS software with no artificial intelligence functionality or any automatic guidance. This makes the process of taking these measurements a slow and laborious task. In contrast, a trained UNet++ model predicts all landmarks, corresponding angle measurements and suggests a diagnosis for an image in less than five seconds on an average laptop. This technology therefore has the potential to increase the efficiency and accuracy that these measurements are taken in clinics. As a consequence, this software can help clinicians avoid fatigue whilst taking such routine measurements.

Furthermore, using this technology could eliminate the necessity for a consultant clinician to take the measurements, allowing a more junior resident (or even a midwife in the case of DDH) to take on this responsibility. Moreover, patient images are often sent to a radiologist team who reviews them and sends back the images to doctors with measurements taken. Automated landmark technology has the potential to enable clinicians to swiftly take these measurements themselves, thereby reducing the time required to deliver results to patients. Automatic landmark detection technology even has the potential to help with image acquisition by guiding clinicians to align their scanners or probes so that the most useful view of a patient's anatomy is obtained.

5.6.2 Landmark Detection and Angle Calculation

The models trained in this section for the Cephalometric and Hand datasets produced SOTA landmark detection results. This shows how effective the UNet++ is at detecting landmarks when the hyperparameters of the training pipeline are tuned correctly. The trained models for the Pelvis and Infant Hip datasets produced landmark detection results with a median radial error of 1.576mm and 4.936 pixels respectively.

The radial errors of the predictions for the Pelvis and Infant Hip datasets were caused by two reasons. The first reason is ambiguous landmark definitions. The radial errors for the femoral head centre and neck orientation point for the Pelvis dataset are caused by this because there is no clear point at which these landmarks should be placed. This is also the case for the ilium point 1 in the Infant Hip dataset. As there is no unique placement for these points it is likely there is also large human inter-rater variability for these landmarks. The second reason is model prediction mistakes. An example of this is the lateral cam point in the Pelvis dataset. The trained model often missed small bumps on the femoral head especially when the images are of low quality. This produced the higher radial errors seen in the test set. Future work could reduce model errors by implementing an attention mechanism to force the model to focus on the femoral head when performing landmark detection as discussed in Section 5.4.3.

For the Pelvis dataset the average radial error the model achieved for specific landmarks propagated through to the accuracy of the LCE-angle and α-angle measurements. The LCE-angle was accurately predicted with an ICC score of 0.919 because the lateral acetabulum point has a clear definition and was the most accurately predicted landmark. On the other hand, the radial errors for the lateral cam point produce some incorrect α-angle predictions causing an ICC score of 0.563 for that angle. It was a different story for the Infant Hip dataset where although the ilium point 1 landmark had a high radial error that did not translate to high α-angle or β-angle errors. This was because the radial error for that landmark was along the axis used to take the angle measurements.

5.6.3 Diagnostic Performance

In regards to using these models for the separate task of diagnosing FAI and DDH conditions the results were mixed. Whilst both models achieved a specificity of over 96% when diagnosing conditions, their sensitivity was lower at 53.85% for cam-type impingement and 63.16% for DDH. This means that patients with the conditions present were often misdiagnosed. Cam-type impingement cases were missed because the model failed to detect small bumps on the femoral head. In addition, both the Pelvis and Infant Hip datasets contained more healthy patients than patients with these conditions. This likely biased the model to predict fewer patients with conditions. Finally, the models made mistakes when the ground truth was close to the angle threshold for that patient. In these cases, both false positive and false negative diagnostic predictions were made. The models often made relatively accurate landmark predictions, but if the predicted angle was the wrong side of the diagnostic threshold, even by a few degrees, it would be classified as a misprediction. This effect could be measured in future by having a third category of diagnosis called 'borderline' containing all predictions within a couple of degrees of the threshold.

The consequence of missing these cases in a clinic would be a reduction in patient care quality. Missing patients with cam impingement will mean that these patients will continue to experience discomfort with hip movement and are more likely to develop osteoarthritis. Missing patients with DDH would mean that these children will have to have more surgeries to fix their hips because the later the condition is spotted the more treatment is required. These negative effects highlight that diagnosis in clinics should not be provided solely by these models. Any automatic landmark detection system deployed in the clinic should have a mechanism where the clinician can check and edit the automatically placed landmarks to ensure they are clinically acceptable. Additionally, a failsafe mechanism should be built into landmark detection models which can alert the user when an incorrect prediction is likely to have been made. This would help draw the clinician's attention to these landmarks and suggest it is likely they will need to be edited. A novel method for achieving this is presented in Chapter 6.

6

Utilizing Output Heatmaps

Contents

6.1 Visual Characteristics of Output Heatmaps 135
6.1.1 Observations 135
6.1.2 Ideas 135
6.2 Reliability Diagrams For Landmark Detection 138
6.2.1 Method 138
6.2.2 Experiments Performed 138
6.2.3 Results and Conclusion 141
6.3 Temperature Scaling 144
6.3.1 Experiments Performed 144
6.3.2 Temperature Scaling Results and Conclusions 147
6.4 Expected Radial Error (ERE) 149
6.4.1 Definition 149
6.4.2 Examples 156
6.4.3 Correlation 156
6.4.4 Comparison To Highest Heatmap Value 160
6.4.5 ERE Throughout Training 161
6.4.6 ERE Conclusions 163
6.5 Using ERE To Flag Potentially Unacceptable Predictions 164
6.5.1 Method 164
6.5.2 Experiments Performed 165
6.5.3 ROC Results 167
6.5.4 Quantitative and Qualitative ERE Results 167
6.5.5 Conclusion 173
6.6 Ensembling 174
6.6.1 Formal Definitions 174
6.6.2 Experiments Performed 175
6.6.3 Ensembling Results 178
6.6.4 Uncertainty in an Ensemble 179
6.6.5 Ensembling Conclusions 184
6.7 Hip Studio 185

The start of Chapter 5 explained how the models described in this thesis were trained on Gaussian heatmaps centred at the ground truth landmark positions. As a result, these models output heatmaps from which final landmark predictions were obtained by selecting the highest value pixel in each output heatmap. This was shown in Figure 5.2. Chapter 5 did not present output heatmaps in detail and they were simply described as a method for obtaining landmark predictions. This aligns with the majority of landmark detection literature, that similarly refrain from analysing the output heatmaps of models. However, this chapter proposes a new idea. The hypothesis behind the methods proposed in this chapter is that information stored in the output heatmaps of models can be used to obtain uncertainty measurements for each landmark prediction. This idea is powerful because output heatmaps are already a feature of landmark detection models so if they can be utilized to provide an uncertainty measurement then no new architecture implementation or model training would be required to obtain this benefit. This uncertainty measurement could then be used to flag potentially inaccurate landmark predictions to clinicians.

This idea that the output heatmaps of landmark detection models can be used to provide an uncertainty measurement has been explored concurrently with the work presented in this chapter by Payer *et al.* [105], Thaler *et al.* [128] and Schobs *et al.* [115]. However, as detailed in Section 2.6.4, these methods have weaknesses. Their methods require a new model to be trained and they do not demonstrate that their uncertainty measurement can flag inaccurate predictions to clinicians. In this chapter, the visual characteristics of output heatmaps are examined in Section 6.1 and then quantitatively analysed in Section 6.2. Based on the observations made in these sections a novel method for obtaining an uncertainty measurement is proposed in Section 6.4. The work presented in Section 6.1, Section 6.2 and Section 6.4 has been published in the paper entitled *Contour-Hugging Heatmaps for Landmark Detection* in the proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2022 [88].

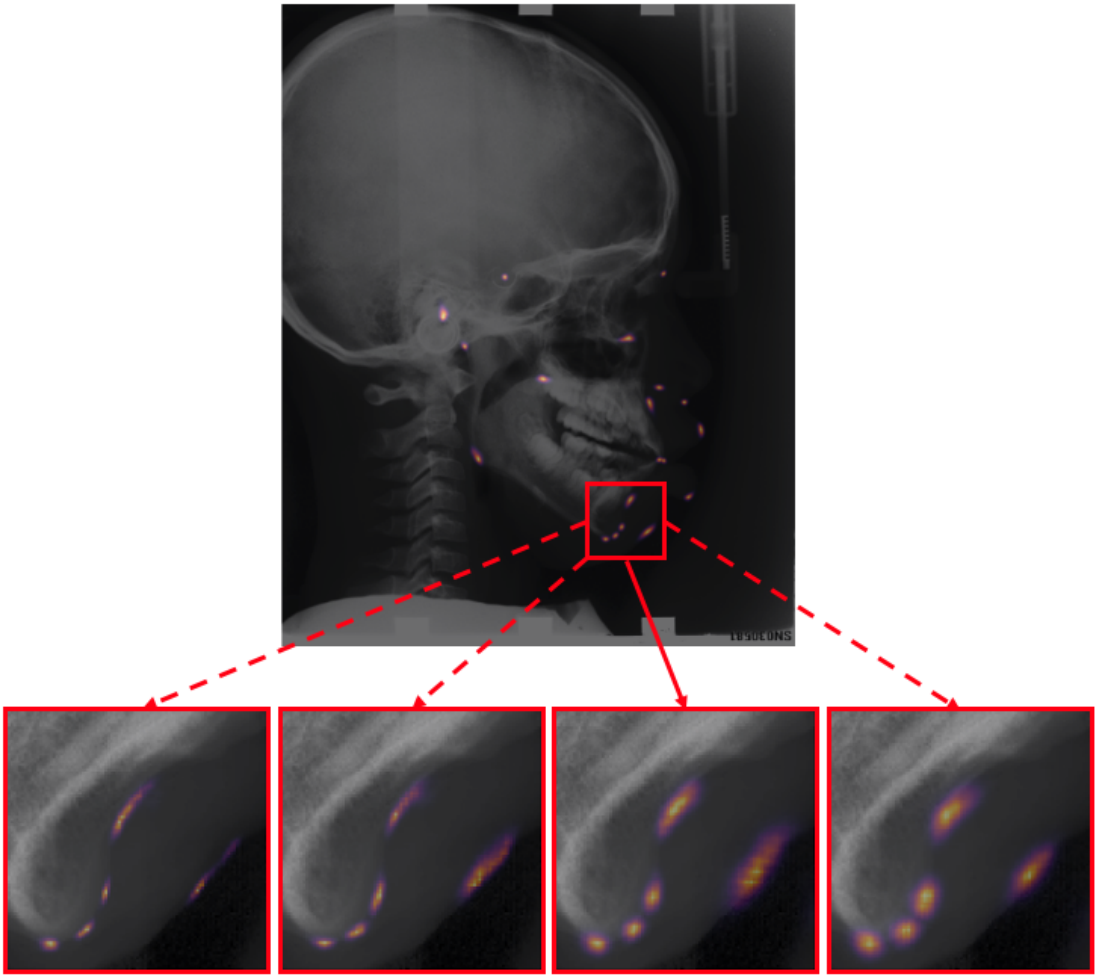
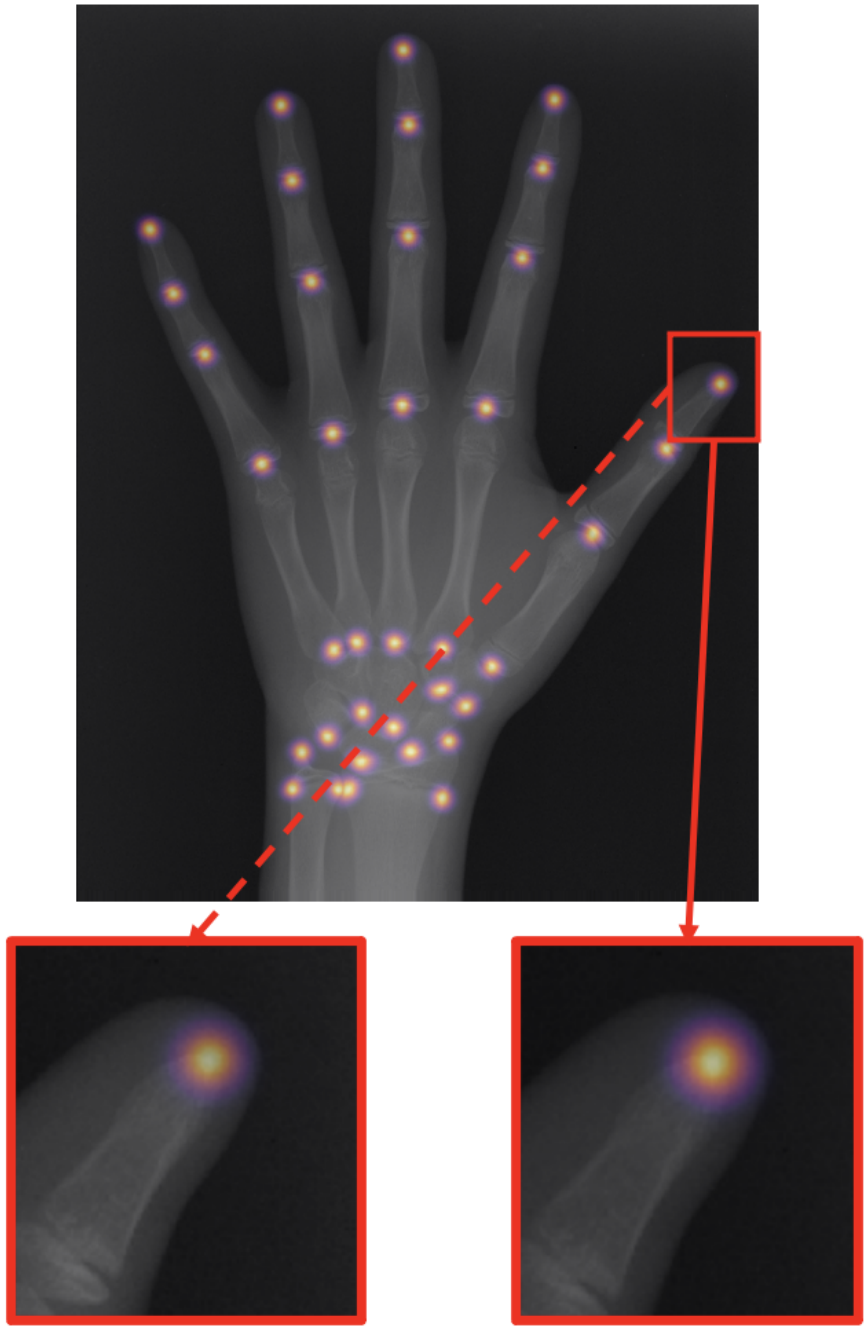


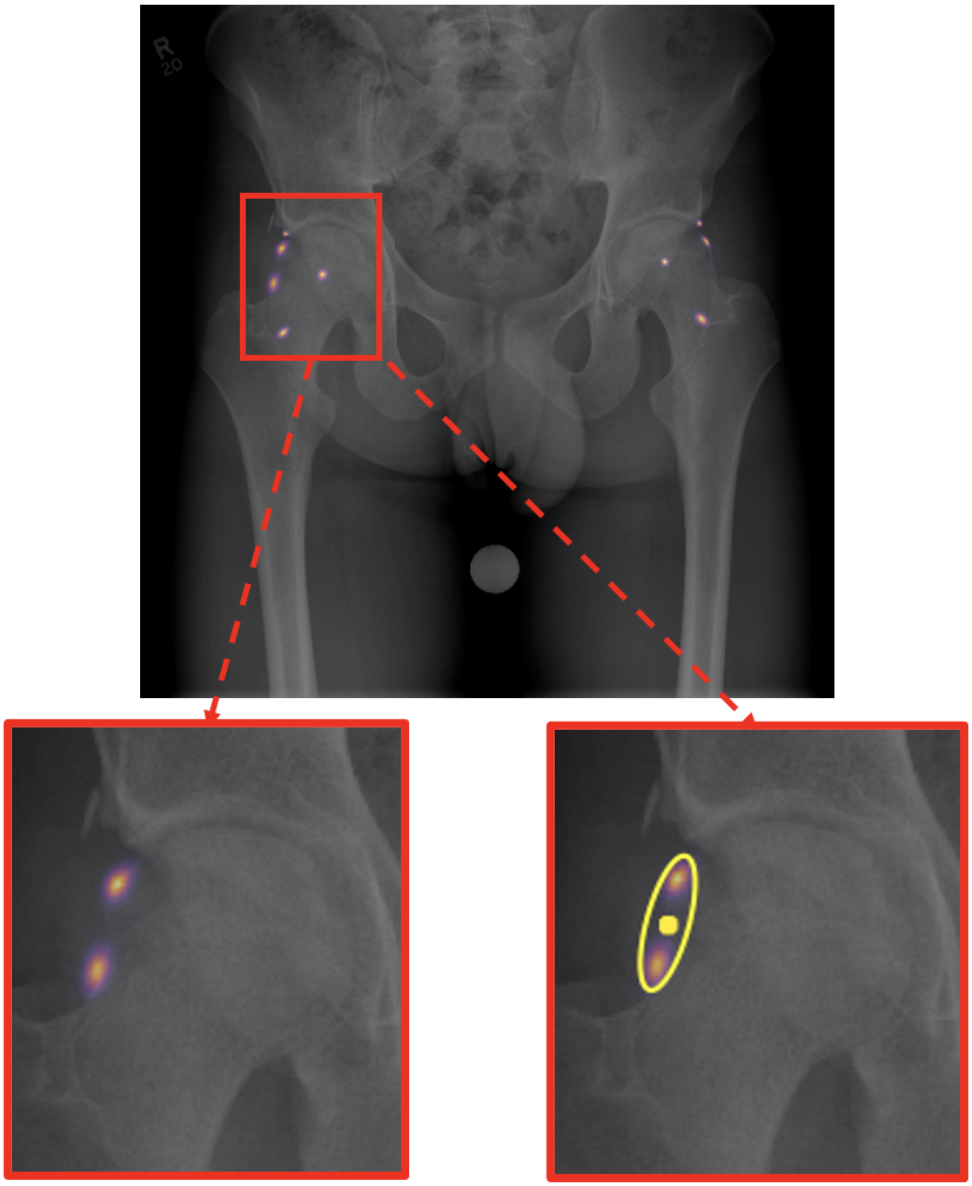
Figure 6.1: Sub-figures (a) to (d) visualise output heatmaps for five landmarks on the chin of a patient captured in the Cephalometric dataset. Each area of bright pixels is the heatmap for a particular landmark and five heatmaps have been compressed into a single image to produce these figures. The brighter (more white) the pixel in the heatmap is the higher the value of that pixel in the heatmap. For the experiments described throughout this thesis, the hottest point (pixel with the highest value) of each heatmap is selected as the final landmark prediction. The captions for each sub-figure show what value of σ (or one-hot) was used to produce the Gaussian training heatmaps for that model. The whole cephalometric image with overlayed heatmaps is shown at the top of the figure. The heatmaps on the whole image were generated with a model trained on $\sigma = 2$ Gaussian heatmaps hence why the arrow to sub-figure (c) is solid and the others are dotted. Subfigures (a), (b) and (d) were still generated on the same patch of the image shown. The models used to produce these output heatmaps are the same models which were trained for hyperparameter tuning as described in Section 5.2.3. It can be seen that the output heatmaps have higher variance (larger spread) the larger the σ value of the Gaussian training heatmap. In addition, models trained using Gaussian heatmaps with lower σ produce output heatmaps that follow the anatomical contours of the chin more closely.



(a) A training Gaussian heatmap with $\sigma = 3$ for a landmark at the end of a thumb in the Hand dataset.

(b) An output heatmap for the same landmark.

Figure 6.2: These figures show the training and output heatmaps for a landmark at the end of a digit in the Hand dataset. The model used to output the heatmap in (b) is the same model which was described in Section 5.5. This figure shows that training and output heatmaps have similar variance. In addition, it is clear from this figure that neither of these heatmaps represents the true probability distribution of where this landmark is likely to be placed because it is unlikely a clinician would place the landmark either within the finger bone or outside the skin. However, these points are still assigned some probability in the training and output heatmap.



(a) This figure shows a single multi-modal output heatmap for a lateral cam point on a Pelvis.

(b) This figure shows the Gaussian approximation of this output heatmap. The yellow ellipse represents the landmark's uncertainty.

Figure 6.3: (a) shows an output heatmap for a single lateral cam point landmark produced by the model described in Section 5.4. The heatmap is multi-modal as shown by the two separate areas of bright pixels. This output heatmap shows that the model is uncertain whether the landmark should be placed high up the side of this patient's femoral head or lower on the femoral neck. This is caused by the model being unsure if a small bump exists on this femoral head, a failure case which was discussed in Section 5.4. (b) shows a visualisation of the Gaussian distribution calculated from the heatmap. It is clear from this example that using the Gaussian distribution as an approximation will not work in this case because the predicted landmark (yellow dot) is not placed at either mode of the heatmap thus this placement is unlikely to be accurate. In addition, the Gaussian uncertainty distribution does not encode the multi-modality of the output heatmap because a higher probability will be given to pixels close to the yellow dot which is different to the output heatmap.

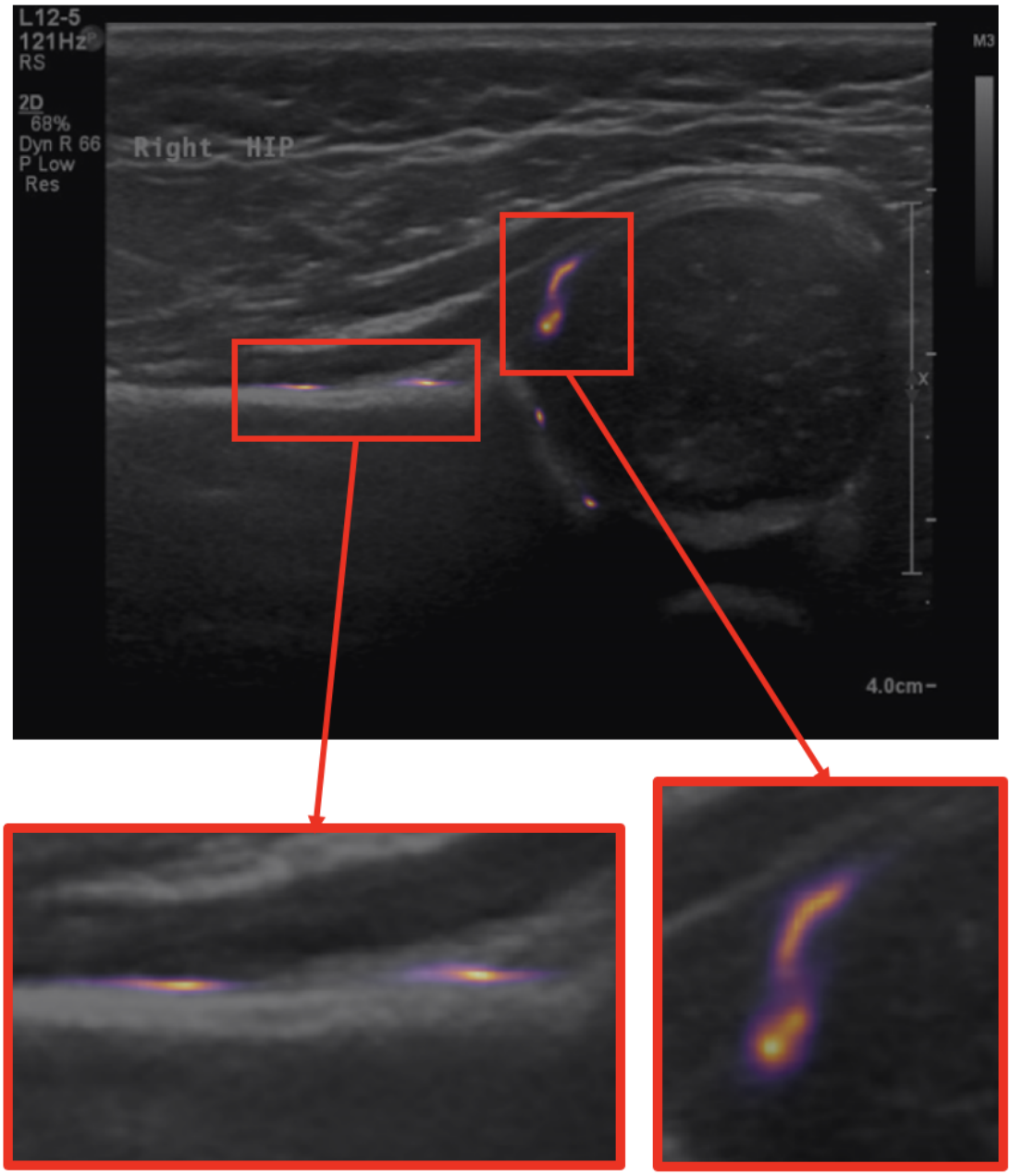


Figure 6.4: An example of an output heatmap produced by the model described in Section 5.5 on an Infant Hip image. Heatmaps for four of the landmarks, including the two in the lower left panel, have a single mode and are approximately Gaussian. In contrast, the output heatmap for the labrum landmark is multimodal as shown in the lower right panel.

6.1 Visual Characteristics of Output Heatmaps

6.1.1 Observations

Figures 6.1 to 6.4 show output heatmaps of the models described in the previous chapter. From these figures the following observations can be made:

1. **Contour Hugging** - Despite Gaussian training heatmaps being perfectly symmetrical, the output heatmaps are asymmetrical and often follow the contours of anatomical edges. For example, the heatmaps shown in Figure 6.1 follow the outline of the patient's chin. This is especially visible for models trained with Gaussian heatmaps with low σ values.

2. **Multi-Modal** - Output heatmaps can be multi-modal. Example of this include the output heatmap for the lateral cam point on a Pelvis image shown in Figure 6.3a and the output heatmap for the labrum point on an Infant Hip image in Figure 6.4.

3. **Relationship Between Training Heatmaps and Output Heatmaps** - The larger the σ of the training Gaussian heatmaps the higher the variance of the output distributions as shown in Figure 6.1.

6.1.2 Ideas

None of these observations have been made in previous literature. These observations lead to the following ideas:

1. **Heatmaps as Discrete Distributions** - Observing that the output heatmaps are contour hugging suggests that the heatmaps represent where a clinician would place the landmark. In addition, the multi-modal distributions, such as in Figure 6.3a, exist because clinicians have placed that landmark around either of the modes in the training set. In the case of the example this is because clinicians often place the lateral cam point either just above a small bump of the femoral head or at a point on the femoral neck. These observations lead to the idea that it is possible to interpret output heatmaps as

discrete distributions where the value at each pixel represents the probability of a clinician placing the landmark at that point.

2. **Assessing Calibration** - Guo *et al.* [50] proposed using reliability diagrams and the 'Expected Calibration Error' (ECE) statistic to assess the calibration of the output distributions of a classification network as discussed in Section 2.6.1. This poses the question as to whether it is possible to adapt this method to landmark detection to assess the calibration of the heatmaps output from models. A further idea is that if calibration can be assessed then an optimal method for generating training heatmaps, which produces the most calibrated output heatmaps, can be found. Figure 6.1 shows that the higher the standard deviation of the training heatmaps, the higher the variance of the output heatmaps. Intuitively, training a model with Gaussian heatmaps will result in uncalibrated models. This is highlighted in Figure 6.2 where (a) shows a training heatmap and (b) shows an output heatmap. Neither of these heatmaps is likely to be an accurate representation of where a person would place this landmark because they would not place it so far into the bone or outside the skin when it is intended to be at the tip of a digit. On the other hand, the calibration of one-hot heatmaps (as defined in Section 5.2.3) could also be tested. These training heatmaps only assign a probability to the precise pixel of the ground truth meaning that they do not assume what the probabilities are of the surrounding pixels.

3. **Limitations of Interpreting Output Heatmaps as Gaussian** - Assuming output heatmaps are Gaussian distributions means that the final landmark predictions are taken as the mean of each output heatmap and the landmark uncertainty as the variance of each output heatmap. This approach is used in Payer *et al.* [105] and Thaler *et al.* [128]. However, this is not a good enough assumption when the output heatmaps are multi-modal. This is highlighted in Figure 6.3. Figure 6.3a shows a multi-modal output heatmap for the lateral cam point and Figure 6.3b shows a Gaussian approximation

of the heatmaps. In this figure, the yellow dot represents the expected value of the heatmap and the ellipse represents the covariance matrix. If the predicted point were placed where the yellow dot is in this figure it would be erroneous as the point is highly unlikely to be placed there by a clinician. This qualitative example provides evidence towards using the mode (brightest pixel) as the predicted landmark point. Additionally, it shows that an uncertainty measurement should calculate how likely the mode point is to be accurate, as opposed to how likely the mean point is to be accurate.

6.2 Reliability Diagrams For Landmark Detection

Motivated by the idea that the calibration of landmark detection models can be assessed, this section describes how reliability diagrams were used to quantify whether output heatmaps represent a distribution of where a clinician would place each landmark. The experiments described in this section test the following hypothesis: firstly, whether the calibration of output heatmaps can be tested. Secondly, whether training models with Gaussian heatmaps or one-hot heatmaps produce more calibrated output heatmaps. Finally, in Section 6.3, it is tested whether the Temperature Scaling method can be used to produce output heatmaps which are more calibrated. Reliability graphs and temperature scaling have not been applied to landmark detection in the current literature.

6.2.1 Method

Reliability diagrams are adapted for landmark detection by formulating landmark detection as a classification problem. This is done by flattening the output heatmap (after a two-dimensional softmax has been applied) for each landmark into a one-dimensional vector where each value in the vector represents the network's confidence of the landmark being placed at a particular pixel. Once landmark detection has been formulated this way the reliability diagram method described in Section 2.6.1 is used. If a landmark detection model is perfectly calibrated then the confidence of the highest confidence pixel in the output heatmap is equal to the probability of the landmark being placed there. More tangibly, if the brightest pixel in the heatmap has a value of 0.2 then the ground truth landmark should be located at that pixel 20% of the time.

6.2.2 Experiments Performed

The reliability diagrams produced in this section were created by using models trained with one-hot heatmaps and Gaussian heatmaps with $\sigma = 3$. The value $\sigma = 3$ was chosen because models trained with this σ achieved the best MRE results across all datasets in Section 5.2.3. Reliability diagrams are not given for models

trained using Gaussian heatmaps with other σ values because this section focuses on the difference in calibration between models trained with one-hot heatmaps and models trained with Gaussian heatmaps.

For the Pelvis and Infant Hip datasets the one-hot trained models used for the experiments described in this section are the same as those described in Section 5.2.3. For the Cephalometric and Hand datasets new one-hot models were trained on all images in these dataset's respective training sets because the models described in Section 5.2.3 were only trained on 80% training set images. The hyperparameters of the rest of the training pipeline remained the same as described in that section. The Gaussian heatmap trained models used for the experiments described in this section are the same models described in Section 5.3 to Section 5.5.

For each dataset both the one-hot trained, and the Gaussian heatmap trained, models were run over the dataset's validation set and reliability diagrams were produced. For the Pelvis and Infant Hip datasets the validation set was computed using the KS test in Chapter 4. For the Cephalometric dataset, Test Set 1 was used as the validation set, and for the Hand dataset, the test set of the first training/test split was used as a validation set. Reliability diagrams were computed by aggregating the landmark predictions in the validation sets into ten bins based on their confidence scores. Ten bins were chosen because that number is used in the original reliability diagram paper by Guo *et al.* [50].

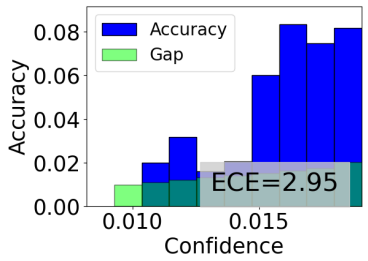
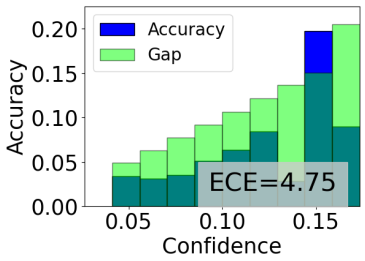
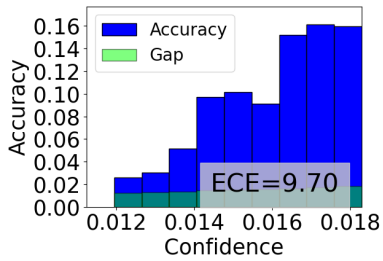
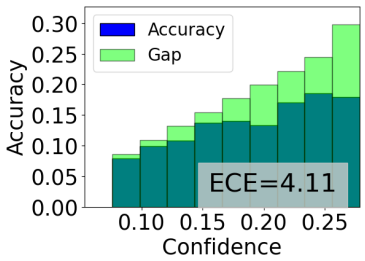
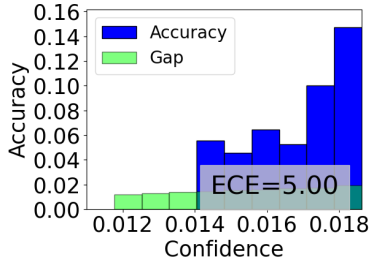
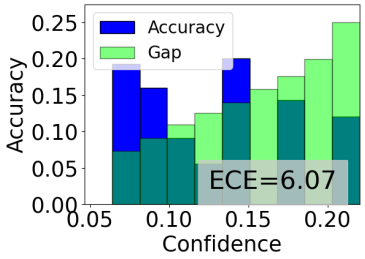
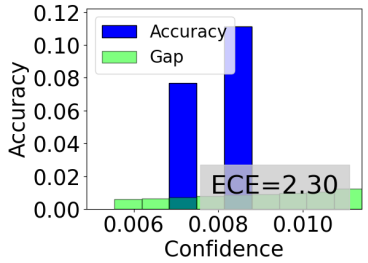
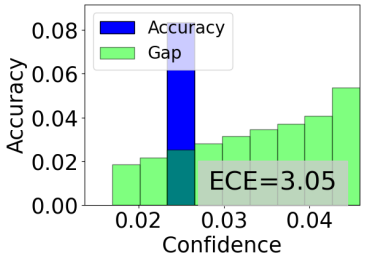
Dataset	Gaussian heatmaps with $\sigma = 3$	One-hot heatmaps
Cephalometric		
Hand		
Pelvis		
Infant Hip		

Figure 6.5: Reliability diagrams for each dataset with models trained on heatmaps generated using the Gaussian function with $\sigma = 3$ and one-hot heatmaps respectively. If the top part of a bar is blue that means the accuracy of predictions was higher than the confidence, i.e. the model was under-confident. If the top part of a bar is green that means the model was overconfident. The 'gap' caption in the legend refers to the gap between the (lower) accuracy of the model and the (higher) confidence of the model. These results show that using $\sigma = 3$ training heatmaps resulted in underconfident predictions whereas using one-hot heatmaps resulted in more calibrated predictions. However, results for Pelvis and Infant Hip datasets are calculated on smaller datasets and as such they produced unreliable graphs where individual highly accurate predictions skewed results for some bars. In addition, it can be seen that the ECE score shown in each plot does not correlate with how calibrated each reliability diagram looks visually. For example, the reliability diagram for the model trained with one-hot heatmaps on the Cephalometric dataset looks more calibrated, and has a higher ECE score than the model trained with Gaussian heatmaps.

6.2.3 Results and Conclusion

Observing the reliability diagrams in Figure 6.5 we see that the models trained using Gaussian heatmaps with $\sigma = 3$ for the Cephalometric and Hand datasets were very under-confident for all predictions. This is shown by the blue bars being significantly higher than the green bars in these figures. On the other hand, the models trained with one-hot heatmaps were on average slightly overconfident in their predictions. This is shown by the comparatively small lime green bars on the figures on the right of Figure 6.5. It makes intuitive sense that a model trained with Gaussian heatmaps would be under-confident. Training a model with Gaussian heatmaps assumes the landmark placement probability reduces equally in all directions from the ground-truth placement. This assumption is untrue in many cases such as for landmarks on the edge of bones as shown in Figure 6.2. Using one-hot heatmaps does not make this assumption and, as such, the models trained with one-hot heatmaps were more calibrated.

For the Pelvis dataset, the model trained with one-hot heatmaps was also underconfident, however, there was a more mixed picture for the model trained with $\sigma = 3$ heatmaps. It was under-confident for lower confidence predictions and mostly overconfident for higher confidence predictions. The first reason for this is because the lateral cam point landmarks were significantly more difficult to predict than the other landmarks in the Pelvis dataset which led to there being little correlation between confidence and accuracy when all landmarks were placed into the same set of bins. Secondly, the validation set for the Pelvis dataset was comparatively small compared to the Cephalometric and Hand datasets. The chosen validation sets for Cephalometric and Hand datasets contained 150 and 298 images per set with 19 and 37 landmarks per image. This means that these validation sets contained $150 \times 19 = 2850$ and $298 \times 37 = 11026$ landmarks respectively. This was significantly more than the Pelvis dataset whose validation set has only $56 \times 8 = 448$ landmarks. The fewer landmarks, the less meaningful the reliability diagrams were, because there were fewer landmarks in each bin. For instance, since there were only ten landmarks with a confidence of 0.15-0.17 this made the

average confidence and accuracy for that bin heavily sensitive to individual landmark predictions. The accuracy rate would have changed from 10% to 30% if only two more of those landmarks (of any type) were predicted exactly on the pixel of the ground truth. This effect was even more pronounced in the Infant Hip dataset, whose validation set only contained 48 images with five landmarks each for a total of 240 landmarks. In the case of the model trained with one-hot heatmaps, there were only two landmarks within that set which were located pixel-perfectly hence the two blue bars on that figure. For those landmarks, the reliability diagram shows that the model was highly under-confident however this was only because there were so few (8-12) landmarks in those bins.

The problem of small sets is amplified by the fact that the class domain for the landmark detection problem is all the pixels in the image, which in the case of the Infant Hip dataset was $640 \times 480 = 307,200$ pixels. This makes a pixel-perfect prediction highly unlikely. When this probability is so low, in addition to a small dataset, reliability diagrams do not show calibration reliably. Nevertheless, the method proposed is sound and provides a useful picture of calibration for larger datasets.

Finally, as in Guo *et al.* [50], the ECE statistic is printed on each reliability diagram (ECE formula described in Section 2.6.1). The ECE results for the Pelvis and the Infant Hip datasets were unreliable because of the small validation set. For the Hand dataset, the ECE was lowest for the one-hot heatmap trained model whereas, for the Cephalometric dataset, it was lowest for the Gaussian heatmaps trained model. This latter result does not correlate with the visualization of calibration given by the reliability diagram. The reliability diagram for the model trained with one-hot heatmaps looks more calibrated than the model trained with Gaussian heatmaps for the Cephalometric dataset. The reason the ECE score for Gaussian heatmaps was lower is that a disproportionate number of the landmark predictions had very low confidence and were placed in the lowest confidence bin. This meant that, for these landmarks, both the accuracy and confidence were very low and thus the ECE score was low. This shows that the ECE score is only

reliable when there is an equal number of predictions in each bin. However, that is not the case for landmark detection. This makes the ECE score an unreliable measure of model calibration in landmark detection.

Future work should explore how reliability diagrams and the ECE can be modified to provide useful readings for landmark detection models. For example, using bins of non-uniform width (width meaning the interval of prediction confidence scores which are assigned to a particular bin) could ensure that an equal number of landmarks are placed into each bin which would mean that no bin would be particularly sensitive to individual predictions. Another improvement could be made by grouping small patches of pixels and then computing reliability diagrams using the sum of confidences and the sum of accuracy within a patch rather than a single pixel. This would lead to more predictions being counted as accurate which would provide more useful reliability diagrams for datasets with a low number of total predictions. Once these changes are made the ECE score, calculated from the reliability diagram, should provide a better measurement of calibration.

6.3 Temperature Scaling

This section describes how Temperature Scaling, as described in Section 2.6.2, can be adapted to landmark detection. This section then describes the experiments which were performed to test whether models trained for landmark detection can be made more calibrated using Temperature Scaling.

To adapt Temperature Scaling to landmark detection Eq.2.4 given in Section 2.6.2 needs to be adapted. Firstly, instead of dividing all values in the model's output vectors by a constant scalar, T becomes a vector of length n where n is the number of landmarks such that $T \in \mathbb{R}^n$. The heatmap for each landmark l where $l \in \{1 \dots n\}$, is scaled by T_l. This means that each landmark has independent scaling and calibration. In addition, to adapt Temperature Scaling to landmark detection the division by T_l takes place over two-dimensional heatmaps before a two-dimensional softmax is applied as opposed to over one-dimensional vectors before a one-dimensional softmax is applied. The Temperature Scaling formula for landmark detection is defined mathematically as:

$$S_l(i,j) = \frac{e^{\tilde{H}_l(i,j)/T_l}}{\sum_{s=1}^{w}\sum_{t=1}^{h} e^{\tilde{H}_l(s,t)/T_l}} \tag{6.1}$$

where $S_l(i,j)$ is the value of the output heatmap at coordinates i and j after Temperature Scaling and two-dimensional softmax have been applied. $\tilde{H}_l(\cdot,\cdot)$ is the value of the output heatmap for landmark l at a particular coordinate. The width and height of the output heatmap are represented by w and h.

6.3.1 Experiments Performed

To find the optimal values of the vector T the images in each dataset's validation set (same validation sets as used for the reliability diagrams in Section 6.2.2) were allocated into a training and a test set (within the validation set) in a split ratio of 50:50. All the parameters of the network (excluding T) were then frozen and the models described in Section 6.2.2 were trained further on this training set to minimise NLL loss over the test set (50% of the validation set). Training was

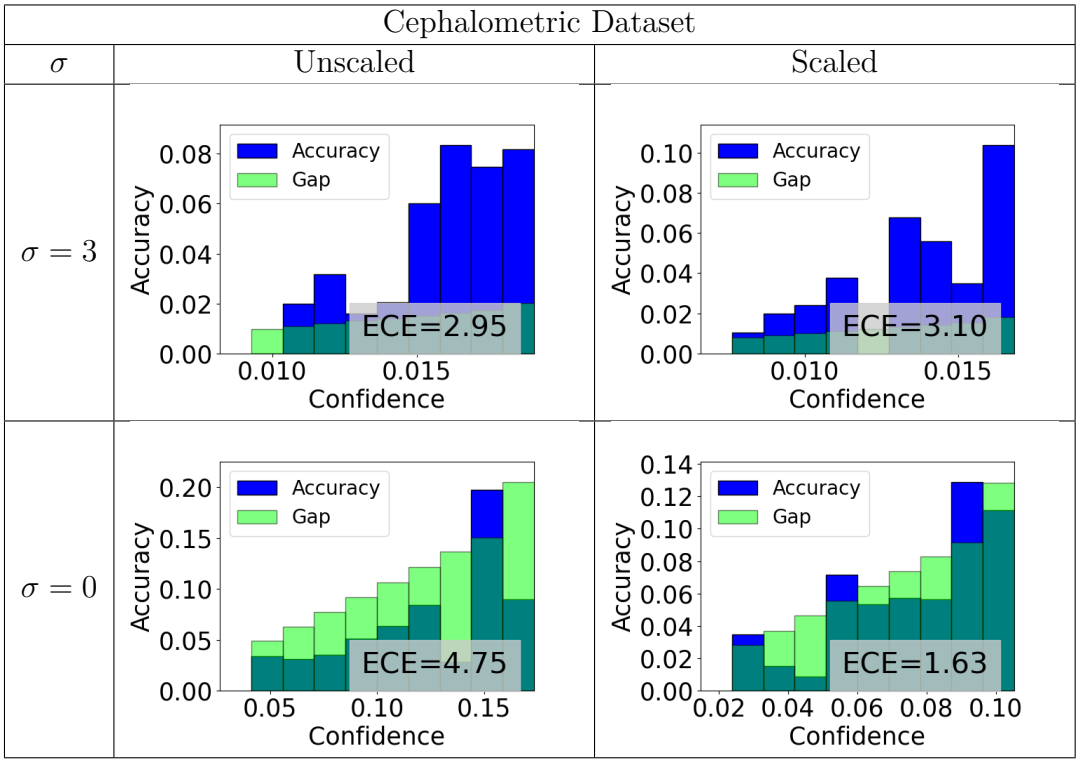


Figure 6.6: This figure shows how Temperature Scaling affected the calibration of models trained on the Cephalometric dataset. The top row shows unscaled vs. scaled for a model trained with Gaussian heatmaps with $\sigma = 3$ and the bottom rows show unscaled vs. scaled for a model trained with one-hot heatmaps. There was no difference in the calibration between the unscaled and scaled model for $\sigma = 3$ models. However, Temperature Scaling did increase calibration for models trained with one-hot heatmaps. This is because Temperature Scaling is developed to calibrate classification networks and one-hot heatmaps formulate landmark detection as a classification problem, as opposed to using Gaussian heatmaps.

conducted in batches of six images and was stopped at epoch 20 because it was empirically observed regularly that ECE does not reduce beyond that point for these datasets (see Figure 6.7). The ECE reported at each epoch of training for each dataset was recorded and is reported in the results Section 6.3.2. In addition, the reliability graphs before and after Temperature Scaling for the Cephalometric dataset are shown in Figure 6.6.

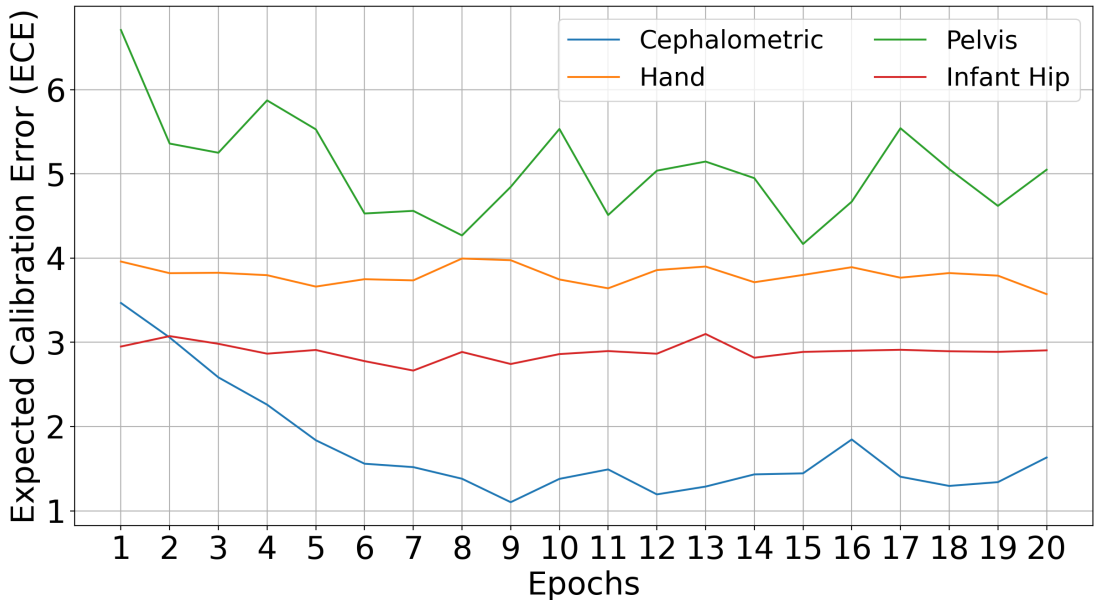


Figure 6.7: This graph shows how the ECE score changed with the number of Temperature Scaling training epochs for models trained with one-hot heatmaps for each dataset. The ECE score for the Cephalometric model reduced during Temperature Scaling however the other models did not see a meaningful reduction.

6.3.2 Temperature Scaling Results and Conclusions

Figure 6.6 shows how the reliability diagram and ECE score changed before and after Temperature Scaling for the Cephalometric dataset. Reliability diagrams showing the effect of Temperature Scaling for the other datasets are omitted for brevity. Instead, Figure 6.7 is given, which shows how the ECE score changed as Temperature Scaling progressed for each dataset. Figure 6.6 shows that Temperature Scaling only worked for the Cephalometric model trained on one-hot heatmaps. This result makes sense because one-hot heatmaps encode landmark detection as a classification problem and Temperature Scaling has only been shown in the literature to work on classification problems. For the Cephalometric dataset, the ECE score reduced after Temperature Scaling and the model became more calibrated, as shown in the bottom row of Figure 6.6. This can also be seen in Figure 6.7, which shows the ECE dropped as Temperature Scaling progressed.

However, this was not the case for the other datasets. The Hand model calibrated slightly with Temperature Scaling, as did the Pelvis model. However, the ECE score for the latter fluctuated heavily throughout training. The Infant Hip model stayed similarly calibrated throughout Temperature Scaling. This was because, as shown in the Infant Hip one-hot reliability diagram in Figure 6.5, there was only one pixel-perfect landmark detection in the Infant Hip set. This caused the single large blue bar in the reliability diagram. It was not possible to calibrate a model based on only a single correct detection.

As discussed in Section 6.2.3, the ECE score is not a reliable indicator of calibration for small datasets such as the Pelvis and Infant Hip datasets and therefore the ECE scores in Figure 6.7 were not to representative of the calibration of the networks. Regarding the Cephalometric and Hand datasets which do have significant numbers of landmarks one reason Temperature Scaling worked more effectively for the Cephalometric dataset is that there was a larger difference between how accurately different landmarks were predicted by the model. In the case of the Cephalometric dataset, Landmark 16 was regularly mispredicted because of labelling inconsistencies whereas the other landmarks were located relatively accurately.

This meant Temperature Scaling had a big effect at normalizing the calibration across these separate landmarks. Alternatively, all Hand dataset landmarks were consistently located accurately and Temperature Scaling, therefore, had limited potential to normalize ECE across the landmarks.

To conclude, Temperature Scaling improves the calibration of landmark detection models trained on one-hot heatmaps for datasets where a large number of images are available and there are differences in how accurately different landmarks are predicted. However, to confirm this conclusion experiments on other datasets with these characteristics should be performed and ECE should be modified to be more applicable to landmark detection tasks as explained at the end of Section 6.2.3.

6.4 Expected Radial Error (ERE)

The motivation for understanding the uncertainty of landmark detections is so that a system that can flag, or alert, unacceptable landmark predictions made by a model could be produced as suggested at the end of Section 5.6. In Chapter 5, it was shown how deep CNN models predict landmarks quickly and often within a clinically acceptable error. However, it was also shown that all the landmark detection models made mispredictions either because of ambiguous landmark definitions or because of model prediction failures. If a system that flags unacceptable landmark predictions automatically could be created then, when a model is deployed, a clinician could correct the unacceptable landmark predictions in the clinic before measurements or diagnoses of patients are made based on unreliable calculations. Having a system that flags potentially unacceptable landmarks would boost clinician confidence in the software and lead to its improved and more reliable adoption.

This section describes a proposed statistic called Expected Radial Error (ERE), a function of the output heatmap, which is a measure of the uncertainty of a landmark prediction.

6.4.1 Definition

The ERE statistic is calculated using the following equation:

$$ERE_l = C \cdot \sum_{i=1}^{w} \sum_{j=1}^{h} S_l(i,j) \sqrt{(i-\tilde{x})^2 + (j-\tilde{y})^2} \tag{6.2}$$

where w and h are the width and height of the output heatmap; $S_l(\cdot,\cdot)$ is the value of the output heatmap at a specific coordinate after it has been passed through a softmax function. $l \in [1..n]$ where n is the number of landmarks. $(\tilde{x}, \tilde{y})$ are the coordinates of the predicted landmark defined as the mode (hottest point) of S_l. C is the pixel-to-mm ratio for the particular image. This is used to scale ERE so it can approximate the radial error which is also given in mm.

This proposed statistic is called Expected Radial Error because assuming $S_l(i,j)$ represents the true probability that a clinician placed the landmark on pixel i,j

then ERE is the expected (Euclidean) radial error of that prediction. Section 6.2 demonstrated that calibrated landmark detection heatmaps are hard to obtain with existing methods. The name Expected Radial Error will still be used throughout the remainder of this thesis because it is the most descriptive of its function.

A graphical explanation of ERE is given in Figure 6.8. The blue box in the figure moves over each pixel in the output heatmap with the distance from the hottest point and the value of the heatmap multiplied for each pixel and summed to produce the final ERE value. If the output heatmap is more spread out with proportionally lower values near the hottest point, then the ERE score will be higher; whereas the ERE score will be lower if there are fewer bright pixels in the output heatmap (i.e. if the heatmap is more concentrated on one location).

The ERE statistic can be calculated on any landmark detection output heatmap. This means that it can be 'bolted-on' to any existing landmark detection network which uses heatmaps to provide an uncertainty measure with landmark predictions. No new network architecture is required for ERE to work. In addition, no retraining of existing landmark detection models is required as opposed to existing uncertainty measures in literature such as Payer *et al.* [105]. This is an advantage because implementing new architectures and retraining models takes significant time and computational power.

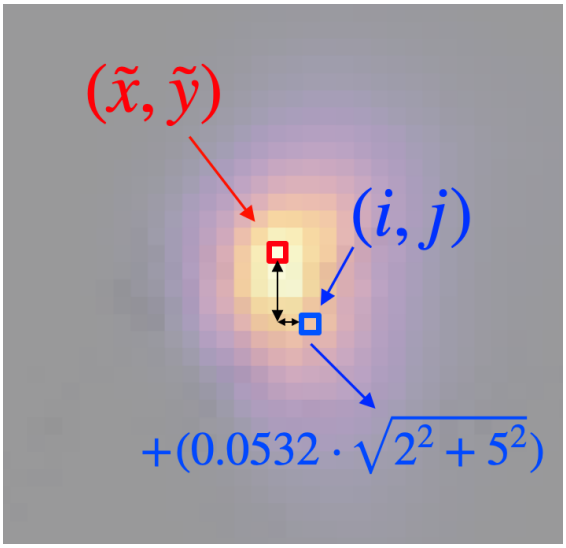


Figure 6.8: This figure shows an output heatmap for a landmark. The red box represents the final landmark prediction at the hottest point on the heatmap. The blue box represents pixel (i, j) which iterates over the whole output heatmap. In this case, the pixel is two pixel coordinates away on the horizontal axis and five coordinates away on the vertical axis from the predicted landmark position. The value at (i, j) in the heatmap in this example is 0.0532. Therefore $0.0532 * \sqrt{2^2 + 5^2}$ is the term added to the total sum of the ERE score for (i, j).

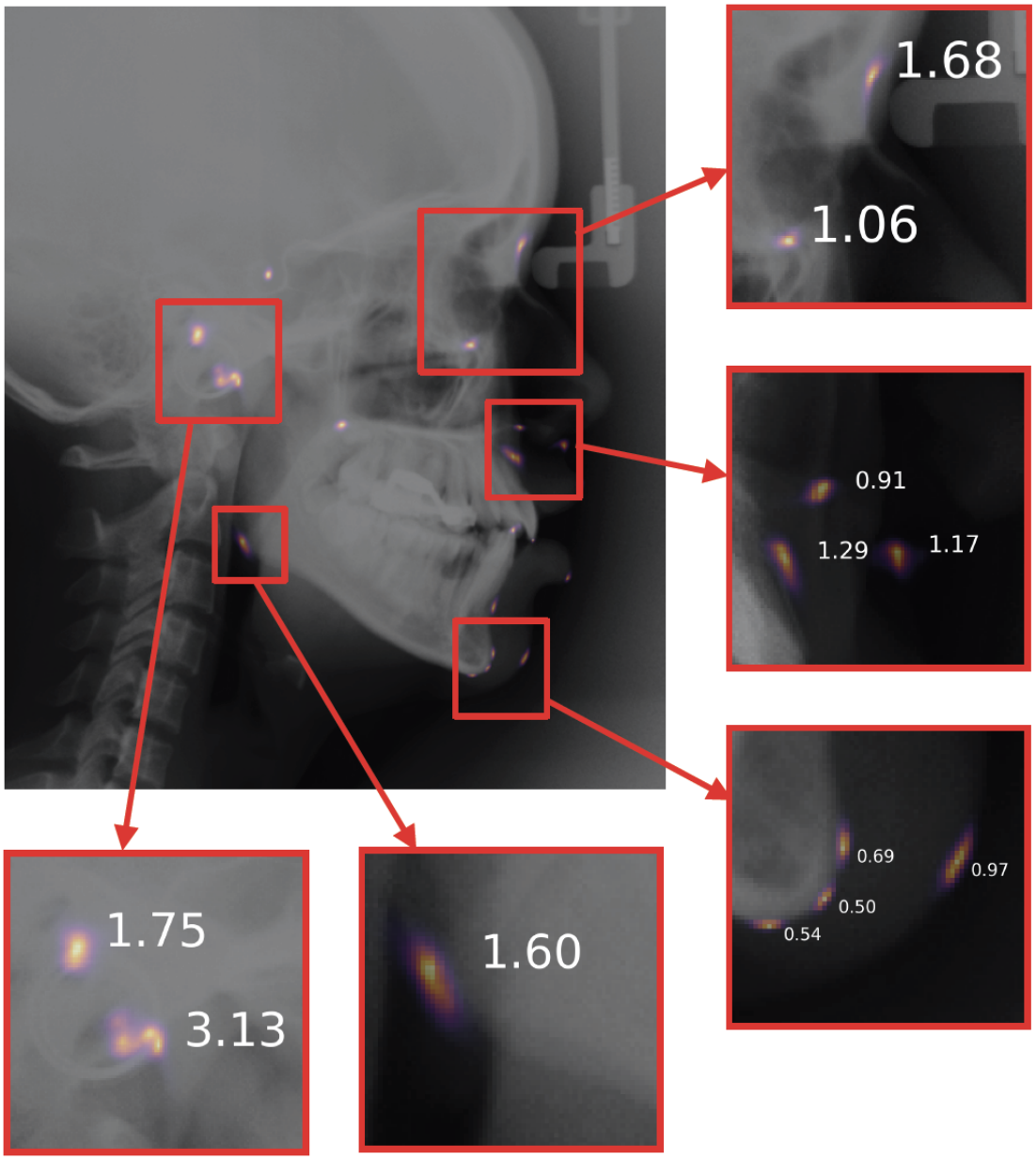


Figure 6.9: This figure shows output heatmaps and their corresponding ERE scores for a model trained with one-hot heatmaps for a Cephalometric image. The model which produced these heatmaps is the same as was described in Section 6.2.2. Landmarks with less spread-out heatmaps, such as those on the chin, have lower ERE scores when compared to landmarks with more spread out heatmaps such as those around the temple.

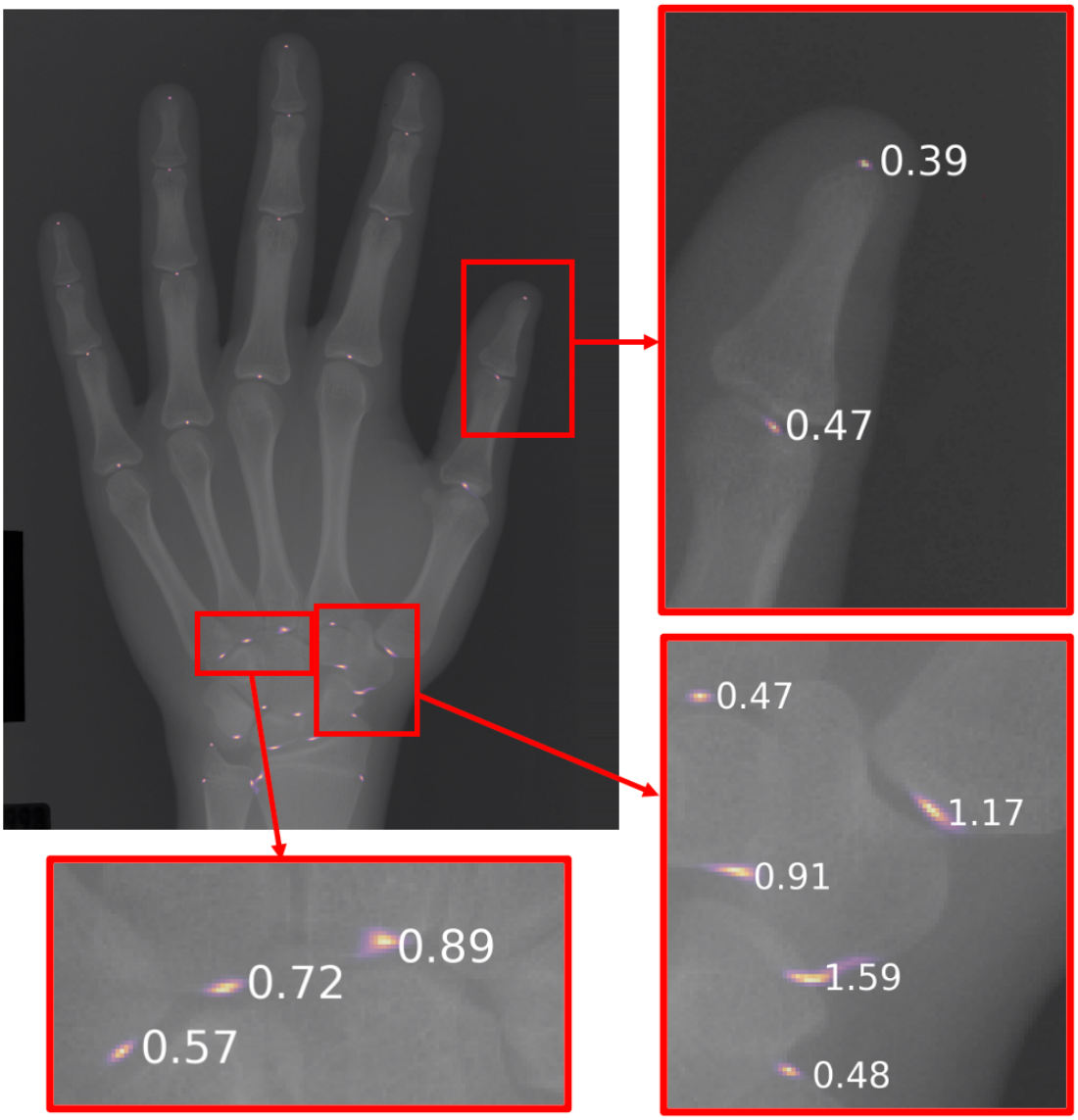


Figure 6.10: This figure shows output heatmaps and their corresponding ERE scores for a model trained with one-hot heatmaps for a Hand image. The model which produced these heatmaps is the same as was described in Section 6.2.2. Hand landmarks were on average predicted relatively accurately compared to the other datasets which means that the ERE scores are also low for this dataset. However, it can still be observed that landmarks with a greater heatmap spread have larger ERE scores in the output heatmaps shown in the bottom right panel.

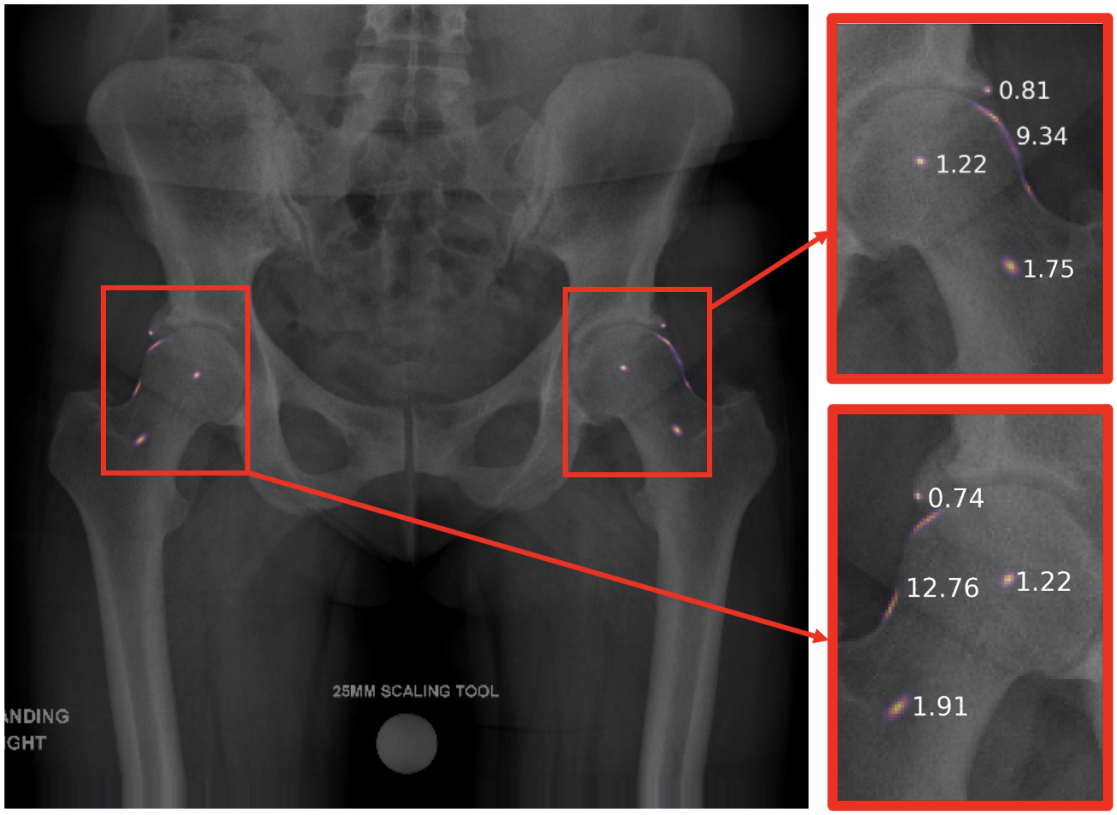


Figure 6.11: This figure shows output heatmaps and their corresponding ERE scores for a model trained with one-hot heatmaps for a Pelvis image. The model which produced these heatmaps is the same as was described in Section 6.2.2. The output heatmaps for the lateral cam points on each hip have significantly higher ERE scores than the other landmarks. These heatmaps are multi-modal and curve around the femoral head which produce high ERE scores.

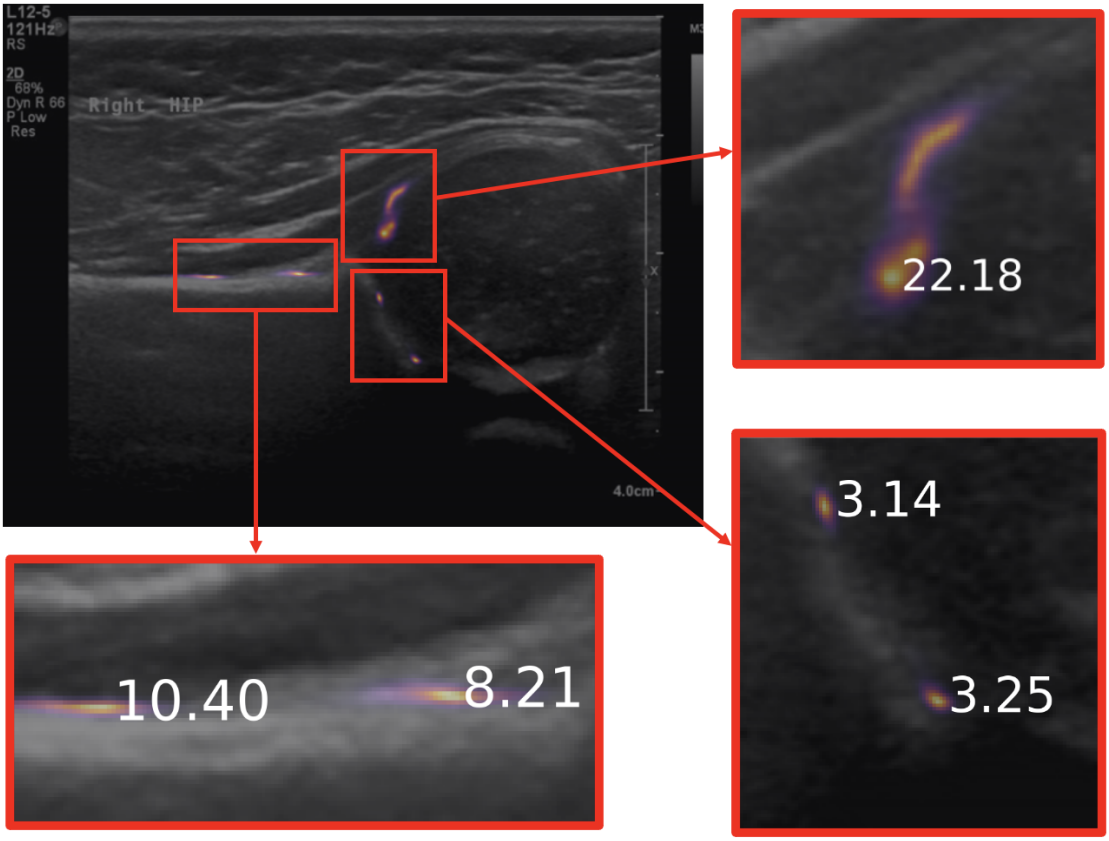


Figure 6.12: This figure shows output heatmaps and their corresponding ERE scores for a model trained with one-hot heatmaps for a Infant Hip image. The model which produced these heatmaps is the same as was described in Section 6.2.2. The largest ERE score is exhibited by the labrum point which has a multi-modal heatmap. The landmarks along the ilium also have relatively high ERE scores but have heatmaps that are similar to Gaussian. This shows that the uncertainty Gaussian statistics provided in past work are also represented by the ERE in this work.

6.4.2 Examples

Figures 6.9 to 6.12 show ERE scores next to the heatmaps upon which they were calculated for the different datasets. The output heatmaps were generated using models trained with one-hot heatmaps for each dataset. The same models which were described in Section 6.2.2. It can be seen from the heatmaps in Figure 6.9 that the larger the spread or variance of the heatmaps, the larger the ERE score. The two heatmaps shown in the lower left region in the figure have a larger variance and higher ERE scores than the landmarks on the chin that have relatively low ERE scores. In addition, it is observed that output heatmaps that qualitatively appear to be similar to Gaussians have a corresponding ERE score relative to their variance, such as those on the ilium in Figure 6.12.

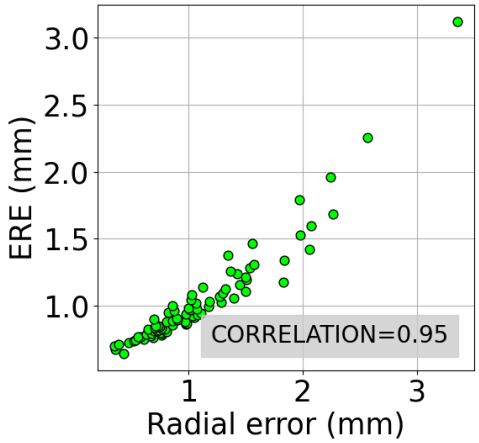
A landmark shown in Figure 6.9 has a multi-modal heatmap and a corresponding ERE score of 3.13, the highest in the image, which shows that the ERE statistic also captured the uncertainty of multi-modal distributions. Large ERE scores are also observed in the multi-modal output heatmaps for the lateral cam point in Figure 6.11 and the labrum point in Figure 6.12. The correlation between the variance of the output heatmaps and the ERE score can be seen qualitatively in all figures.

6.4.3 Correlation

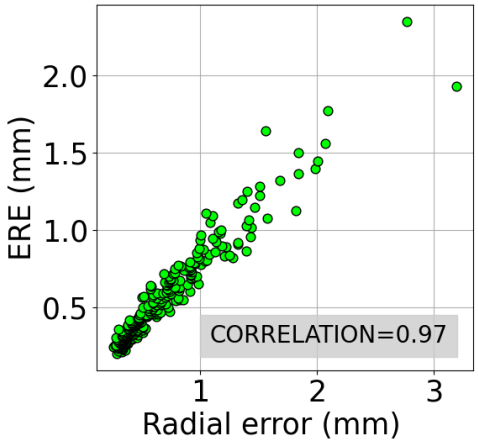
For the ERE score to be used to flag potentially inaccurately predicted landmarks, the ERE must correlate with the radial error of the landmark predictions. This sub-section describes how it was validated that ERE does correlate with radial error.

Experiments Performed

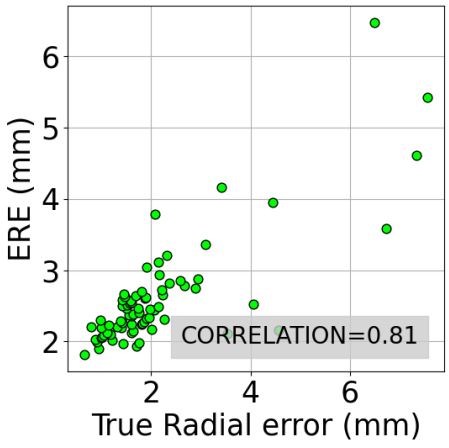
Models using different heatmap generation methods were trained for each dataset. The heatmap generation methods used were Gaussian heatmaps with σ set from one to seven and one-hot heatmaps. The Pelvis and Infant Hip models used were the same as those described in Section 5.2.3. New methods were trained for the Cephalometric and Hand datasets for each heatmap generation method because the models described in Section 5.2.3 were only trained on 80% of the training images



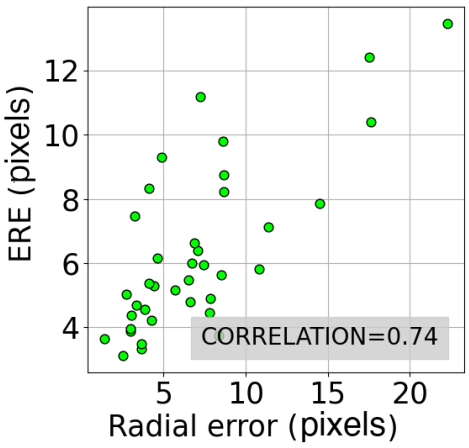
(a) Shows the correlation, with a bin size of 36, for the Cephalometric dataset for a model trained with $\sigma = 2$ heatmaps.



(b) Shows the correlation, with a bin size of 36, for the Hand dataset for a model trained with one-hot heatmaps.



(c) Shows the correlation, with a bin size of 6, for the Pelvis dataset for a model trained with $\sigma = 2$ heatmaps.



(d) Shows the correlation, with a bin size of 6, for the Infant Hip dataset for a model trained with $\sigma = 1$ heatmaps. The radial error and ERE are given in pixels here because no pixel-to-mm ratio was given with the Infant Hip images.

Figure 6.13: These graphs show the correlation between ERE, on the y-axis, and radial error, on the x-axis. Each dot represents a bin of landmarks.

for these datasets. All other hyperparameters of the training pipeline remained the same as the models described in that section.

Once these models were trained each was tested on the corresponding dataset's validation set. For these experiments, the same validation sets described in Section 6.2.2 were used. The radial errors and the ERE scores were recorded for each landmark prediction these models made in each validation set. To calculate the correlation between the ERE score and the radial errors of a model the same method as Kumar *et al.* [71] was used. Firstly each predicted landmark was put into a bin depending on its ERE score. Each bin had an arbitrary number of landmarks in it. Kumar *et al.* [71] chose each bin to contain 36 landmarks so that is the bin size which was used for Cephalometric and Hand datasets in this section. A bin size of six landmarks was chosen for the Pelvis and Infant Hip datasets because there were fewer landmarks in these datasets and it is important to have a significant number of bins. It is worth noting that the more landmarks in each bin the more likely it is that there will be correlation. This is because the more landmarks are aggregated in each bin the more the effect of outlier landmarks is masked. Once landmarks were put into bins the average ERE score and the average radial error were calculated for each bin and plotted as points on the x-axis and y-axis respectively. The correlation was then calculated for these points.

Results

Figure 6.13 shows the correlation graphs for landmarks in the validation set of each dataset as well as the overall correlation coefficient. The plots shown are for predictions made with models trained on Gaussian heatmaps with σ values of 2, 2 and 1 for Cephalometric, Pelvis and Infant Hip datasets respectively. The plot for the Hand dataset was for a model trained with one-hot heatmaps. Figure 6.13 shows correlation graphs for these models in particular because they produced the highest correlation between ERE and radial error of all the models tested. Each graph shows that there was a correlation, ranging from 0.74 to 0.97, between the ERE and

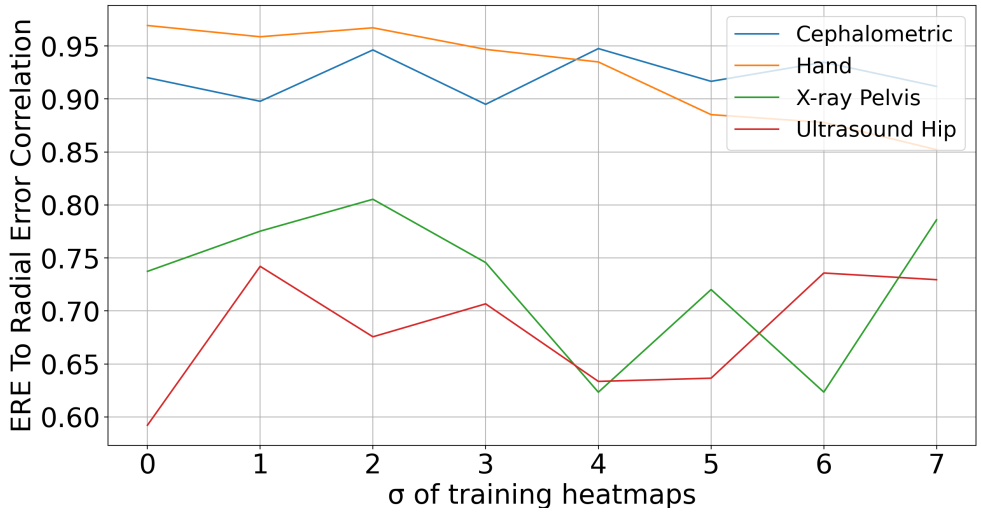


Figure 6.14: This graph shows how the ERE to radial error correlation (as calculated by the graphs in Figure 6.13) varied depending on the σ of the training heatmaps for each dataset.

radial error. Although the Pelvis and Infant Hip datasets had the lowest correlation this was also influenced by the smaller bin size of six used for those datasets.

It is worth noting that although there was a correlation between ERE and radial error, the ERE did not directly predict the radial error. This is shown by how the domain of the x and y axes in the correlation graphs is not the same. This is because output heatmaps for models trained on Gaussian heatmaps are more spread out the greater the σ used and produce higher ERE scores.

The σ of the Gaussian training heatmaps is plotted against the correlation between ERE and radial error in Figure 6.14. It can be seen that Cephalometric and Hand datasets consistently had the highest correlation scores regardless of the σ value. The other datasets consistently had lower correlations and fluctuated more depending on the σ value. One reason for this is that there were fewer landmarks in each bin so the correlation score was more sensitive to individual landmark errors.

Overall these results show that ERE is a good predictor for radial error across the different datasets. Although radial error will not be available when the model is deployed (because there will be no ground truth) as long as the training set is representative of real-world data then ERE can indicate potential landmark accuracy.

6.4.4 Comparison To Highest Heatmap Value

Previous landmark detection uncertainty literature focuses on approximating output heatmaps as Gaussians and uses the values in the covariance matrices as a pseudo-measure for uncertainty in the landmark prediction [105, 128]. This method cannot be easily compared to the ERE statistic because Gaussian approximation assumes the mean point is chosen as the landmark position whereas ERE uses the mode point for the reasons given under 'Limitations Of Interpreting Output Heatmaps As Gaussian' in Section 6.1.1 .

One past approach which can be compared directly with ERE is Schobs *et al.* [115] which uses the highest value pixel in the output heatmap (in other words the confidence of the predicted pixel) as a pseudo-measure of uncertainty. The theory behind this method is that the higher the value the more likely there is to be less radial error for the prediction. This subsection describes how this method was tested against ERE as a measure of uncertainty.

Experiments Performed

The highest value in each predicted landmark's output heatmap produced by models trained on Gaussian heatmaps with σ from one to seven, and one-hot heatmaps, over each dataset's validation set was recorded. These landmark predictions were then grouped into bins, with respect to the highest value in each prediction's output heatmap. The correlation between the average radial error of each bin and the average highest value of each bin was then computed and recorded. This is the same method that Section 6.4.3 used to produce the graphs in Figure 6.13 but using the highest value measurement instead of ERE. In the following results section, the average correlation for the highest value method across all models is reported and compared to the correlation score ERE achieved. Averaging across models trained on different heatmaps ensures that the conclusion drawn from the results is not skewed by a particular model.

Results

Table 6.1 compares the correlation scores of ERE, against the correlation scores for the highest value method. The results show that ERE is a consistently better predictor of the radial error than taking the highest value.

	Mean correlation between the uncertainty statistic and radial error across models trained with σ values from one to seven and one-hot heatmaps.	
Statistic	ERE	Hottest Value
Cephalometric	**0.921**	0.843
Hand	**0.924**	0.903
Pelvis	**0.727**	0.624
Infant Hip	**0.682**	0.660

Table 6.1: This table shows the mean and standard deviation values of the correlation scores between different uncertainty statistics and the radial error. The correlation value was calculated in the same way as Figure 6.13. The uncertainty measures tested are the ERE score proposed in this thesis and the highest value in the output heatmap proposed by Schobs *et al.* [115]

.

6.4.5 ERE Throughout Training

Experiments Performed

As it has been shown in Section 6.4.3 that ERE is correlated with the radial error of predictions and is therefore a measure of confidence in a model's predictions an experiment was performed to test how the confidence of a model changes during training. After every training epoch the models described in Section 6.4.3 were run over the corresponding validation set and the average ERE score for the predictions over that set were recorded. The particular models used were trained on Gaussian heatmaps with σ values of 2, 2 and 1 for Cephalometric, Pelvis and Infant Hip datasets respectively and one-hot heatmaps for the hand model because these achieved the highest correlation scores as shown in Figure 6.14. The models were trained for ten epochs because it was empirically found that the accuracy of the models plateaued after this point (as stated in Section 5.1).

Results

Figure 6.15 shows a plot of the average ERE score over the validation set vs. the epoch of training for each dataset. The ERE score on the y-axis is scaled between zero and one for each dataset because the ERE score is calculated using the pixel-to-millimeter constant which is different for each dataset and thus scaling was performed to fit each dataset into the same domain. It can be observed that ERE reduces for all datasets during training. This result is remarkable because it shows that the model gains confidence in its predictions as it trains. This is a novel result detailed in this thesis, which has not been discussed in previous landmark detection literature.

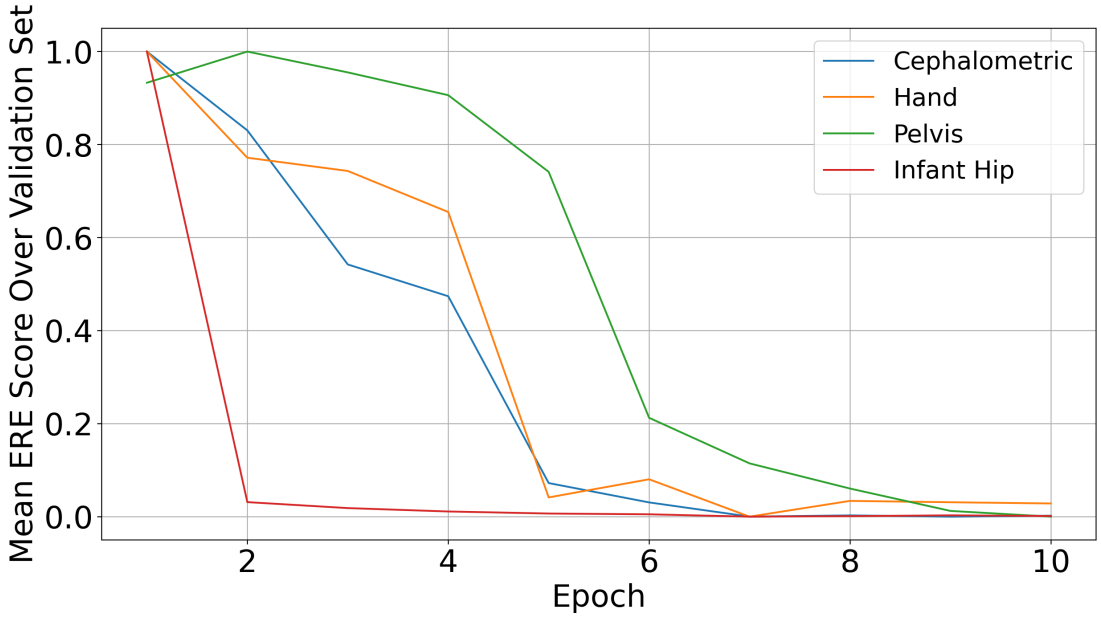


Figure 6.15: This graph shows that the average ERE score across the validation set reduces during training for each dataset as the model gains confidence. The y-axis on this plot is scaled to between zero and one for each dataset so each curve can be seen over the same y domain. Ten epochs of training were chosen because it was empirically found that the accuracy of the models plateaus after this point as stated in Section 5.1. The number of epochs of training does not affect the speed of inference of the model.

6.4.6 ERE Conclusions

The most important discovery reported in Section 6.4 is that the ERE score correlates with the radial error on all datasets, and has a higher correlation than previous approaches. This is an important discovery because the ERE score is easy to calculate, and does not require a new model architecture like previous approaches. This makes the ERE score the most effective and lightweight method for estimating the accuracy of a landmark detection model, improving upon Schobs *et al.* [115]. In addition, ERE provides a way to visualise and measure how a model gains confidence whilst training. However, it is worth noting that the validation of ERE to radial error correlation is limited for small datasets such as the Pelvis and Infant Hip datasets. This is because landmarks must be grouped into small bins before calculating correlation which makes the correlation measurement sensitive to individual landmark errors.

Future work could explore extending the concept that uncertainty can be obtained from values in the output heatmaps of deep learning models to other machine learning tasks. For example, for the task of segmentation an uncertainty measurement could be approximated from the output of a segmentation model. The correlation between this measurement and a segmentation validation metric such as Dice score [122] could then be computed, similarly to how ERE is compared to Euclean radial error in Section 6.4.3.

6.5 Using ERE To Flag Potentially Unacceptable Predictions

Since ERE correlates with the radial error it can be used to flag or alert potentially unacceptable landmark predictions, which is especially useful in the medical field. An alerting system would mean landmark detection algorithms could be adopted into clinic with the confidence that inaccurate predictions will be flagged to clinicians.

6.5.1 Method

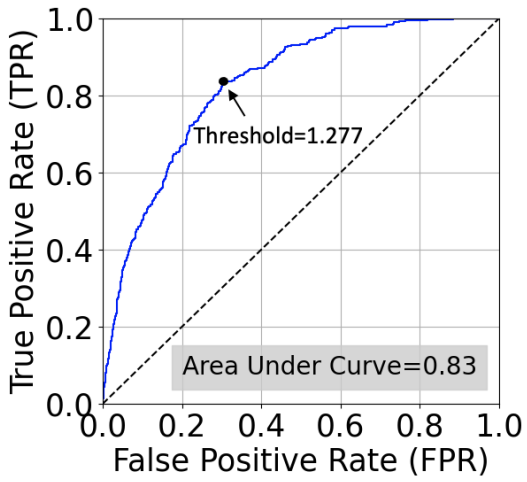
Firstly, what an acceptable landmark prediction is for a particular medical use case is defined. Different applications and medical images have different levels of tolerance for how far predictions can be away from ground truth landmark positions. Once a tolerance is decided a trained landmark detection model is run over a validation set of images and the ERE scores and radial errors computed. Then ERE scores and radial errors are used to plot Receiver Operating Characteristic (ROC) graphs. ROC graphs are used to display and calculate the effect that different threshold values would have on the true positive and false positive rates. In this context, this threshold is different from the tolerance threshold. This threshold represents which ERE scores indicate an unacceptable landmark prediction. In other words, this threshold is a cutoff where predictions with ERE values more than the threshold are classed as unacceptable.

Additionally an Area Under Curve (AUC) score is computed from ROC graphs. The AUC score is a measure of the performance of a classifier. If a threshold exists such that all predictions with an ERE score above that were truly acceptable landmark predictions and all predictions with an ERE below the threshold are truly unacceptable then the AUC score would be one. A baseline AUC for a classifier to achieve is 0.5 because this represents a random classifier with no predictive ability. This approach has the advantage that the user can choose a low threshold for which there will be a low false positive rate (FPR) at the expense of a low true positive rate (TPR), or alternatively, a high threshold for which there is a high TPR at the expense of a high FPR.

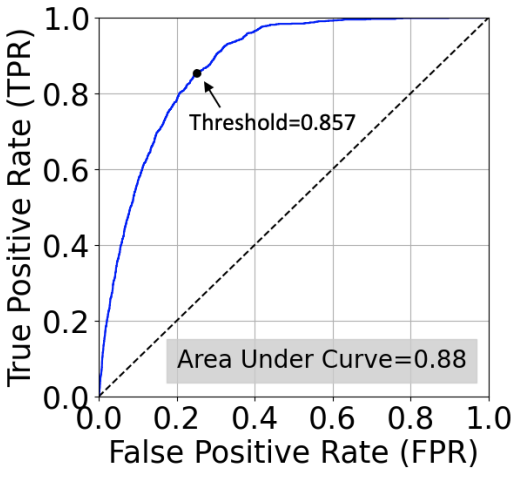
6.5.2 Experiments Performed

The ERE scores and radial errors which were recorded over each dataset's validation set to produce the correlation graphs in Figure 6.13 were used to test the performance of ERE as a classifier of unacceptable predictions. Tolerances of 2mm, 2mm, 4mm, and 10 pixels were chosen for the Cephalometric, Hand, Pelvis, and Infant Hip datasets respectively. The 2mm of tolerance for the Cephalometric and Hand datasets was chosen because existing literature also quote that figure as a 'clinically acceptable' tolerance level [104]. The tolerance levels for Pelvis and Infant Hip datasets were chosen based on clinical common sense. Using these tolerance levels. and the ERE and radial error scores recorded, ROC graphs were plotted for each dataset. The AUC scores were calculated from each of these graphs to provide a measure of how effective each classifier is. Additionally, the ERE threshold which maximised $TPR \times (1 - FPR)$ was calculated from each AUC graph. This threshold was chosen because it is a clearly defined point which made clinical sense. It is worth noting, however, that in certain clinical applications, it would be safer to have a lower threshold which prioritizes the recall of unacceptable predictions to ensure patient safety.

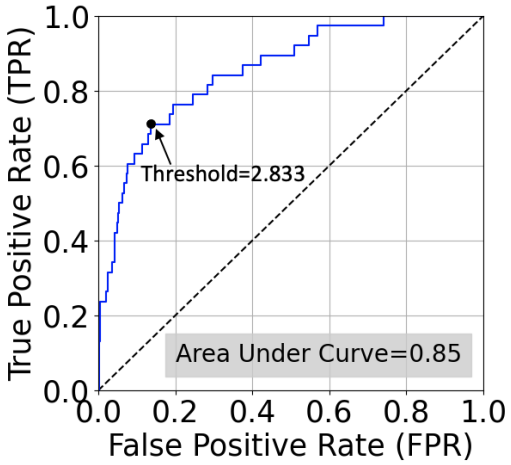
Once the thresholds were calculated, the same models which produced the radial errors and ERE scores over the validation sets were then tested over each dataset's test set. These models are described in Section 6.4.3. The test set for the Cephalometric dataset is Test Set 2. The test set for the Hand dataset is the test set of the first defined training/test set pair given by the authors of the dataset. The test set for the Pelvis and Infant Hip datasets was computed using the KS test method as described in Section 4.4. Each landmark prediction the models made in the test sets was classified based on whether its ERE score was above or below the calculated threshold as either acceptable or unacceptable (flagged). The MRE of landmarks which were flagged and accepted is reported in Section 6.5.4 along with examples of landmarks shown in Figures 6.17 to 6.20.



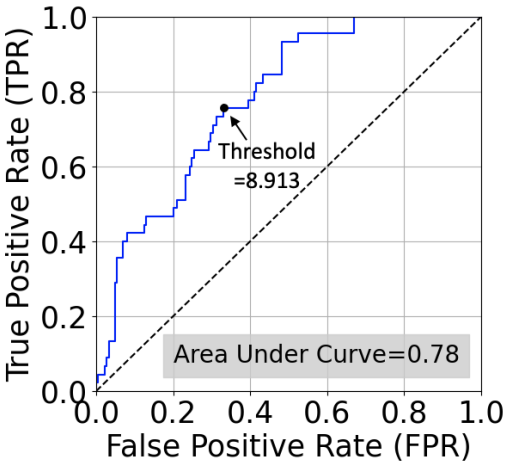
(a) ROC curve for the Cephalometric dataset with a tolerance set to 2mm.



(b) ROC curve for the Hand dataset with a tolerance set to 2mm.



(c) ROC curve for the Pelvis dataset with a tolerance set to 4mm.



(d) ROC curve for the Infant Hip dataset with a tolerance set to 10 pixels.

Figure 6.16: ROC curves for thresholding the ERE score to flag predictions that fall outside tolerance levels specific to each dataset. It can be seen that the Area Under Curve (AUC) score for each dataset was significantly above the 0.5 a random classifier would achieve, represented by the dotted line. ERE can be used to flag erroneous predictions with high accuracy. These curves are computed on the validation set for each dataset. The position of the threshold value that maximises $TPR \times (1 - FPR)$ is shown on each curve.

6.5.3 ROC Results

The ROC graphs for each dataset are shown in Figure 6.16. The AUC scores are 0.83, 0.88, 0.85 and 0.78 for the Cephalometric, Hand, Pelvis and Infant Hip datasets respectively. These results show that all ERE-based classifiers achieved a significantly higher AUC than 0.5 which represents a random classifier. The ERE thresholds that optimized $TPR \times (1 - FPR)$ for these graphs were 1.277, 0.857, 2.833, and 8.913 for Cephalometric, Hand, Pelvis, and Infant Hip datasets respectively.

6.5.4 Quantitative and Qualitative ERE Results

Figures 6.17 to 6.20 show sample output from an image in each dataset's test set with expanded sections showing the ERE scores next to the predictions (yellow dots) and the ground truth (blue dots). Next to each landmark is a tick or a cross showing whether that landmark prediction was accepted (tick) or flagged (cross). The acceptance category for each prediction was decided by whether its ERE score was above the threshold for each dataset. These thresholds are listed in Section 6.5.3.

The majority of the accepted landmark predictions in the figures had radial errors within the clinically accepted range. There is the exception of the two predictions with ERE scores of 1.26 and 1.03 on the chin of the patient in Figure 6.17 (bottom right panel of the figure). However, this is due to the systematic shift in how that landmark is placed between Cephalometric Test Set 1 and Test Set 2 as discussed in Section 3.2. No uncertainty measure can detect labelling discrepancies in a test set because they are unpredictable events caused by human error.

Amongst the landmarks which were flagged, the majority have larger than average localisation errors. This is not the case for all examples though and it is important to make the distinction that a high ERE score is an indicator of a potentially high localisation error rather than a guarantee. This is illustrated by the landmark on the temple of the patient with ERE of 1.70 in Figure 6.17 and the landmarks with EREs of 8.94 and 13.60 in Figure 6.20 on the Ilium and labrum respectively. These landmarks have very low radial errors (accuracy detected) but relatively high ERE scores.

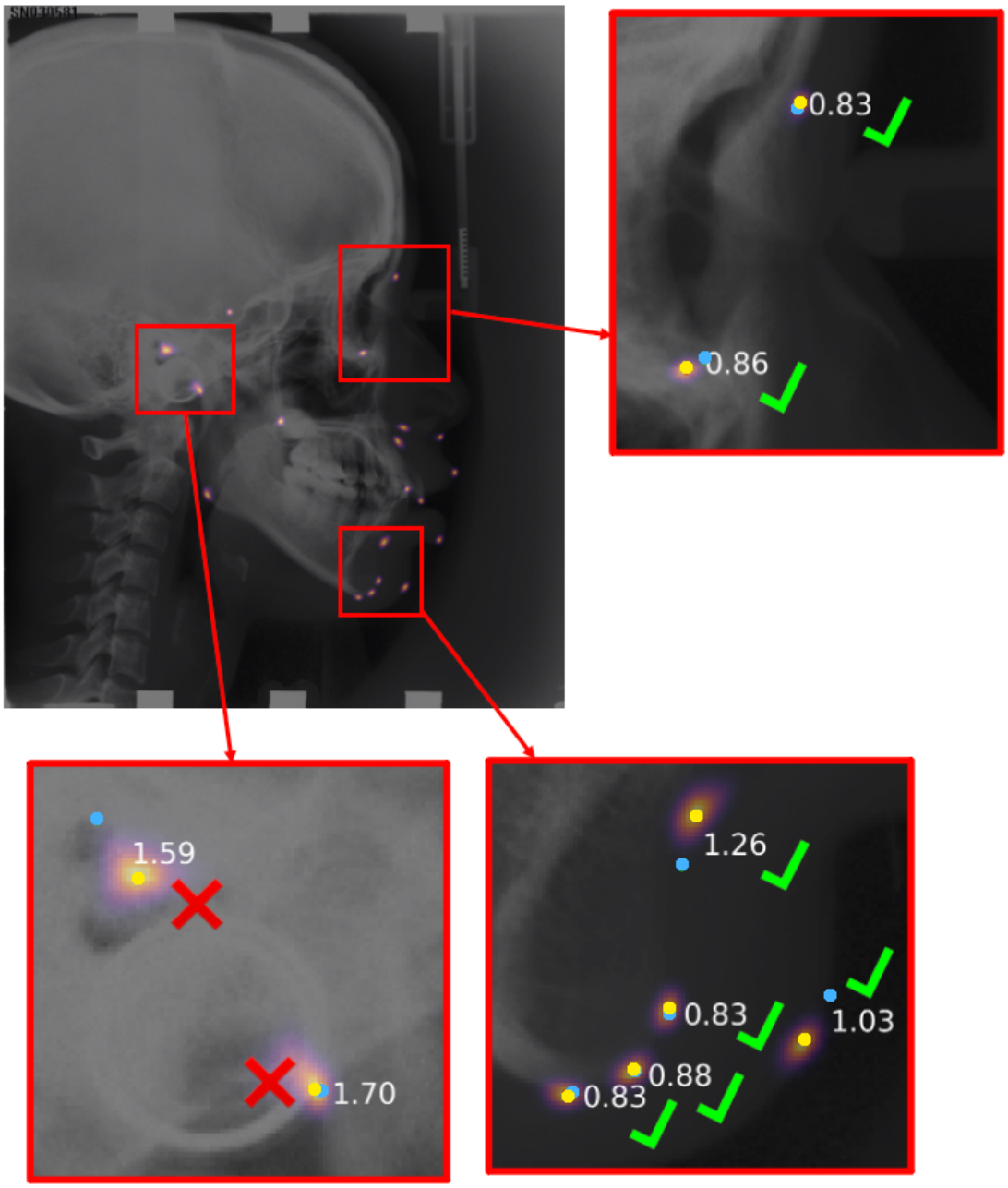


Figure 6.17: In this figure ticks and crosses represent either accept or flag for each prediction and this was decided using the threshold of 1.277 on the ERE score. The ERE score is shown in white text with the ground truth and predicted points shown by the blue and yellow dots respectively. Out of the two landmarks rejected as unacceptable in this figure, one had a large radial error whereas the other was accurately detected. This example shows that ERE flags potentially unacceptable landmarks rather than giving a guarantee that the flagged landmark will be unacceptable. Two of the landmarks accepted in the bottom right region had high radial errors however these were caused by inconsistent ground truth labelling as opposed to a model error.

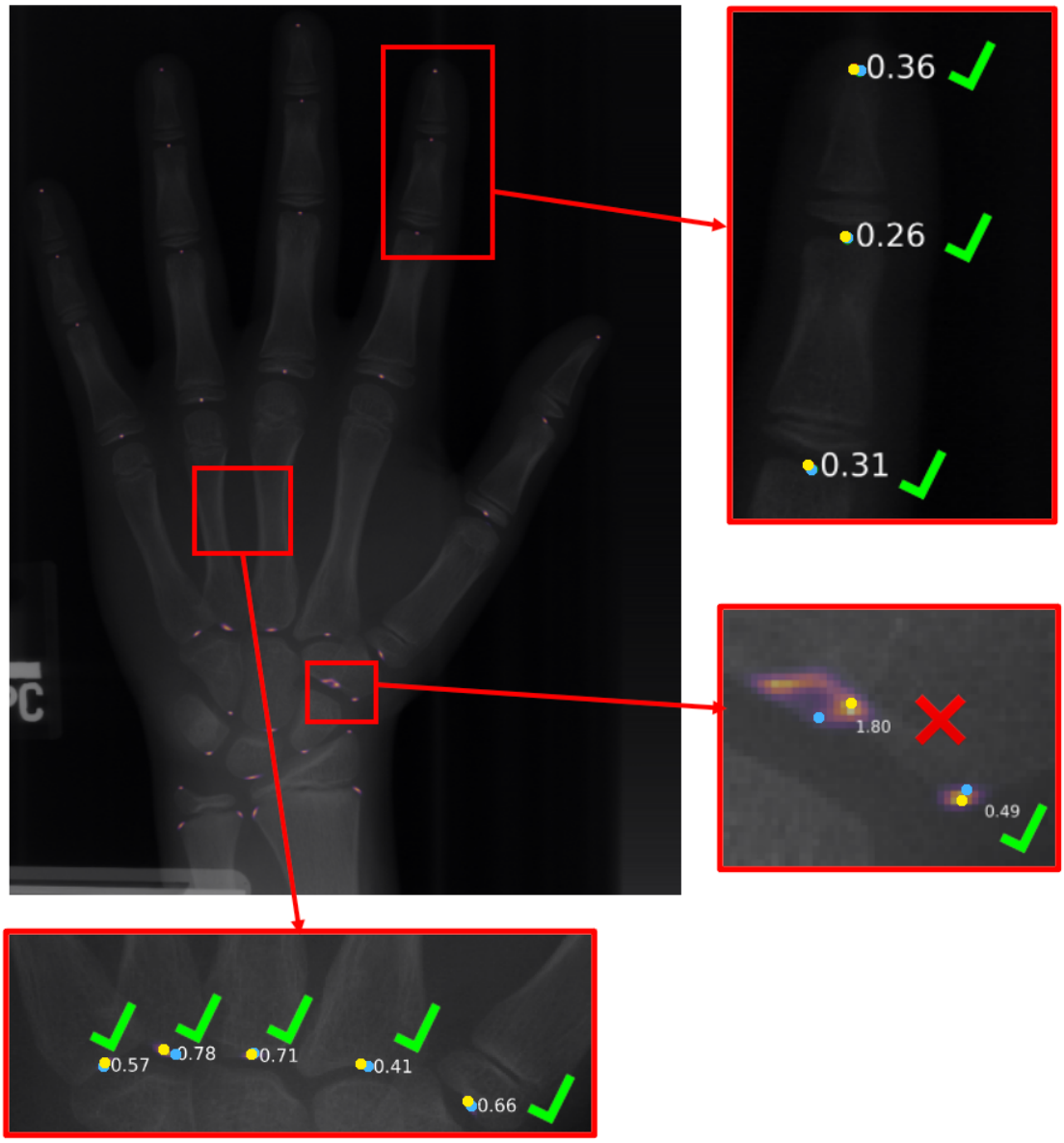


Figure 6.18: In this figure ticks and crosses represent either accept or flag for each prediction and this was decided using the threshold of 0.857 on the ERE score. The ERE score is shown in white text with the ground truth and predicted points shown by the blue and yellow dots respectively. All accurately detected landmarks were accepted. The one landmark prediction flagged as unacceptable had the highest radial error in the image.

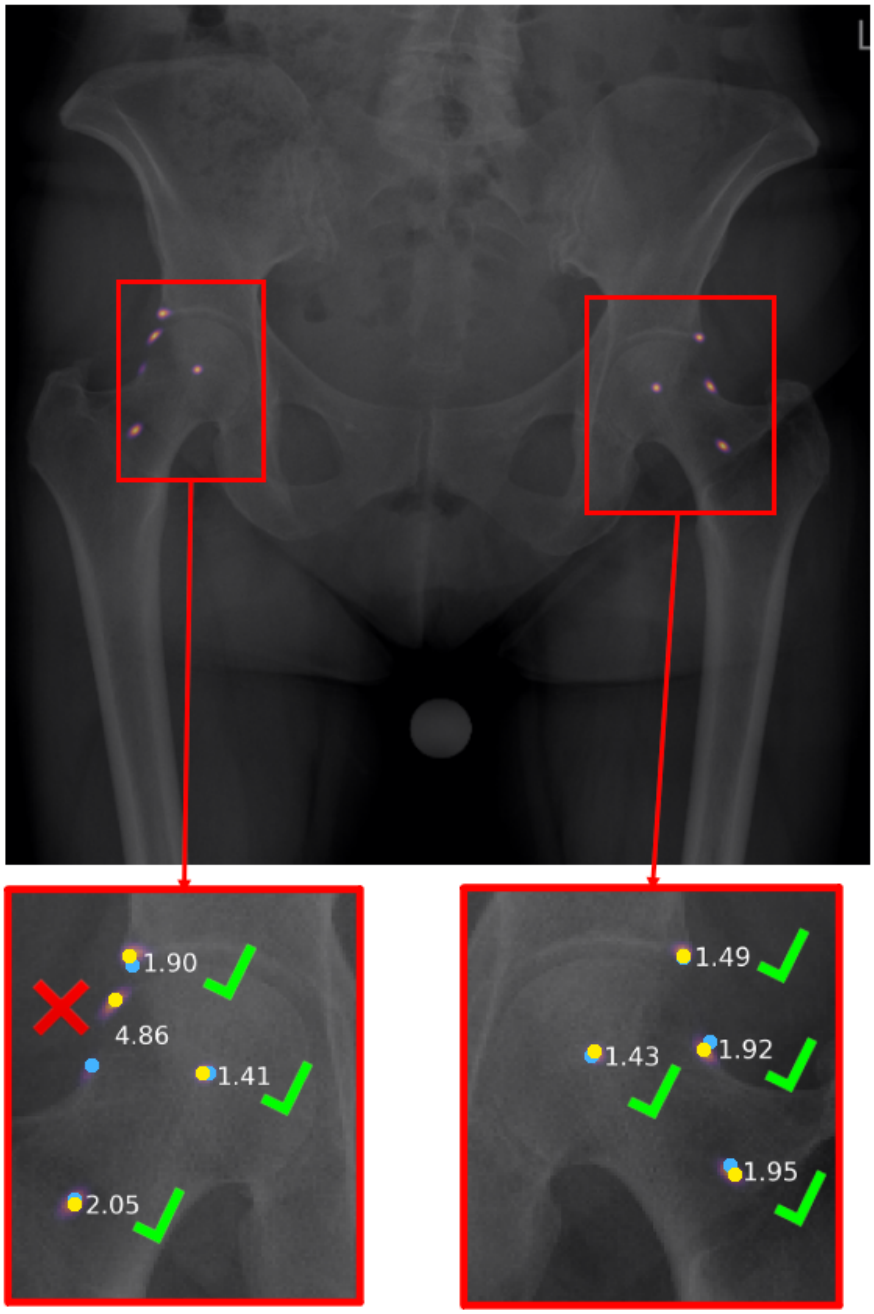


Figure 6.19: In this figure ticks and crosses represent either accept or flag for each prediction and this was decided using the threshold of 2.833 on the ERE score. The ERE score is shown in white text with the ground truth and predicted points shown by the blue and yellow dots respectively. All accurately detected landmarks were accepted. The right lateral cam point landmark (left side of the image) prediction was flagged as unacceptable and had the highest radial error in the image. The heatmap for this landmark was multi-modal with one mode on the predicted landmark point and the other on the ground truth point. This correctly produced a high ERE score and thus this landmark was flagged as potentially unacceptable.

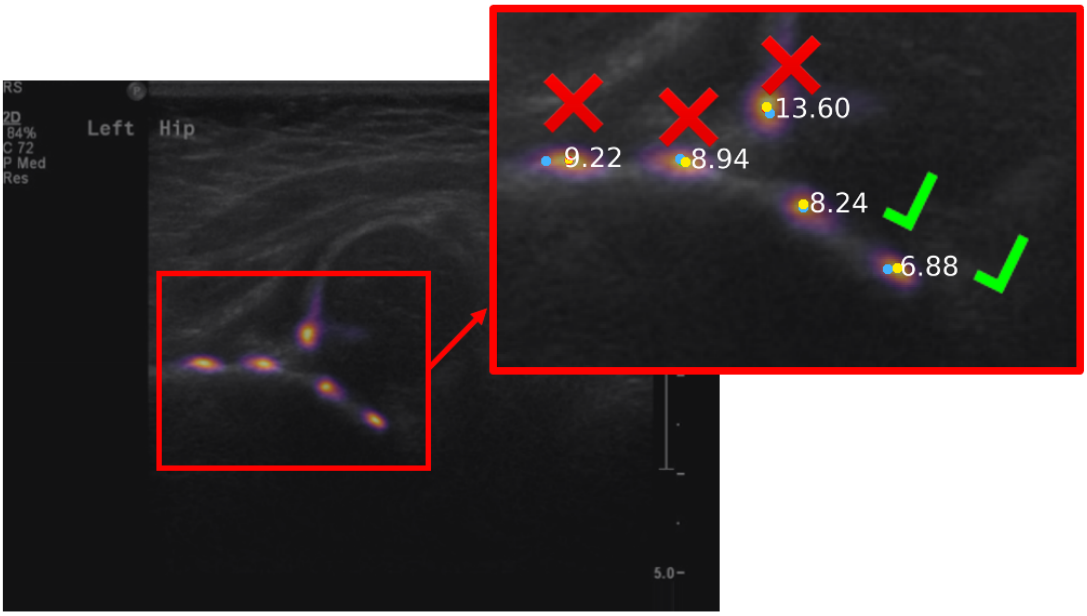


Figure 6.20: In this figure ticks and crosses represent either accept or flag for each prediction and this was decided using the threshold of 8.913 on the ERE score. The ERE score is shown in white text with the ground truth and predicted points shown by the blue and yellow dots respectively. The ilium point 1 had the highest radial error and it was flagged as potentially unacceptable. The turning point had the lowest radial error and it was accepted. The other landmarks had very similar radial errors of which two were flagged as unacceptable.

Tables 6.2 to 6.5 show the MRE and SDR differences between accepted and flagged landmarks for the test set of each dataset. The percentage of landmarks flagged for Cephalometric, Hand, Pelvis and Infant Hip datasets were 12.42%, 13.10%, 9.21% and 24.25% respectively. These flagged landmarks have an average radial error 1.51 times, 2.32 times, 3.97 times and 1.64 times higher respectively than the average radial error for a landmark in their respective test sets. This means that flagged landmarks have a significantly higher radial error than the average landmark and are less accurately predicted. This, in turn, means that this approach was successful at flagging up unacceptable predictions and can be used to alert clinicians of potentially incorrect landmark detections.

Set	# Landmarks	MRE (mm)	SDR (%)			
			2mm	2.5mm	3mm	4mm
Overall	1900	1.40	76.15	83.63	88.47	94.42
Accepted	1664	1.30	77.70	85.40	90.00	95.37
Flagged	236	2.11	65.25	71.19	77.97	87.71

Table 6.2: Statistics for accepted versus flagged landmarks in Test Set 2 for a Cephalometric model trained on $\sigma = 2$ Gaussian heatmaps with a threshold of 1.277 (threshold computed in Section 6.5.3). In total, 12.42% of landmarks were flagged as potentially erroneous and these have an MRE 1.51 times higher than the MRE of the entire test set.

Set	# Landmarks	MRE (mm)	SDR (%)		
			2mm	4mm	10mm
Overall	11026	0.66	95.17	99.24	99.97
Accepted	9582	0.53	98.20	99.84	100.00
Flagged	1444	1.53	75.07	95.22	99.79

Table 6.3: Statistics for accepted versus flagged landmarks in the test set for a Hand model trained on one-hot heatmaps with a threshold of 0.8566 (threshold computed in Section 6.5.3). In total, 9.21% of landmarks were flagged as potentially erroneous and these have an MRE 2.32 times higher than the MRE of the entire test set.

Set	# Landmarks	MRE (mm)	SDR (%)		
			4mm	6mm	10mm
Overall	456	2.63	89.70	95.40	96.05
Accepted	414	1.84	93.72	98.80	99.03
Flagged	42	10.43	50.00	61.90	66.67

Table 6.4: Statistics for accepted versus flagged landmarks in the test set for a Pelvis model trained on $\sigma = 2$ Gaussian heatmaps with a threshold of 2.833 (threshold computed in Section 6.5.3). In total, 13.10% of landmarks were flagged as potentially erroneous and these have an MRE 3.97 times higher than the MRE of the entire test set.

Set	# Landmarks	MRE (pixels)	SDR (%)		
			10 pixels	15 pixels	20 pixels
Overall	235	5.32	86.80	94.89	98.30
Accepted	178	4.22	93.82	97.75	99.44
Flagged	57	8.73	64.91	85.97	94.74

Table 6.5: Statistics for accepted versus flagged landmarks in the test set for an Infant Hip model trained on $\sigma = 1$ Gaussian heatmaps with a threshold of 8.913 (threshold computed in Section 6.5.3). In total, 24.25% of landmarks were flagged as potentially erroneous and these have an MRE 1.64 times the MRE of the entire test set.

6.5.5 Conclusion

ERE can be used to flag unaccepted landmark predictions. The ROC curves in 6.16 show that for all datasets, the ERE score provided a classifier for erroneous predictions which, for all datasets, obtained AUC scores higher than 0.78. In addition, using ROC curves means the end-user can select a threshold based on the TPR:FPR ratio which is appropriate for their particular application. This is especially important in medical imaging where it is often safer for the model to be cautious to ensure patient safety.

Tables 6.2 to 6.5 show that the flagged landmarks were significantly less accurate than accepted landmarks across all datasets. This shows that this method is generic and relevant to many different medical applications where landmarks need to be detected safely.

There is room for improvement with this approach because not all of the inaccurate predictions were filtered out. Additionally, not all flagged landmarks were inaccurate predictions. However, building a system which reliably only flags erroneous examples is difficult because should such a system exist then that information could be built back into the model to improve the original accuracy rather than needing a flagging system at all.

Additionally many of the inaccurate predictions that were not flagged had incorrect ground-truth annotations such as the two Cephalometric landmarks with ERE scores of 1.26 and 1.03 shown in the bottom right panel of Figure 6.17. This provides an additional use case for ERE: instead of using ERE to flag incorrectly predicted landmarks, it could, instead, be used to flag inconsistent ground truth annotations. If the ERE for a landmark is low but the radial error for that landmark is high the ground truth landmark has likely been placed differently to those in the training set. This is an avenue for future work.

6.6 Ensembling

Manually placing landmarks on images to produce landmark detection datasets is time-consuming for expert annotators, and for this reason, large annotated datasets are rare in the medical field. Any viable landmark detection method should therefore work as accurately as possible but, ideally, not require a large-scale training set. A viable solution is to ensemble networks where predictions from multiple models are aggregated to find a final, more accurate prediction. Ensembling has been shown to improve results when limited training data is available in previous literature. However, these approaches worked for different machine learning tasks and there is very little literature on ensembling landmark detection networks. In addition, the existing approaches do not generalise to landmark detection as explained in the 'Few Ensembling Approaches' part of Section 2.5.3.

This section describes how the ERE statistic was used as a weight to perform a weighted averaging of individual landmark prediction models. Using a weight of $\frac{1}{ERE}$ means that the final prediction will be closer to the individual models with the highest confidence (lowest ERE score). As the ERE score correlates with the radial error it is reasonable to expect that this ensemble prediction will be more accurate than the mean average of the models' predictions.

This ensembling method has the advantage that no modification or retraining to individual models is required. Since the majority of existing SOTA landmark detection approaches output heatmaps for their models, the ERE statistic can be calculated on these and then used to combine predictions. This makes this method lightweight and simple whilst being effective.

6.6.1 Formal Definitions

Given m individually trained models each outputting a heatmap $\{H_1...H_m\}$ for each landmark where $H_i \in \mathbb{R}^{w \times h}$ and w and h are the weight and height of input image. Points $\{\mathbf{p}_1...\mathbf{p}_m\}$ denote the hottest points (pixel with the highest value in the output heatmaps) and $\{ERE_1...ERE_m\}$ are the ERE scores for each heatmap. The reciprocal of the ERE score w_i represents the weight as defined below:

$$w_i = \frac{1}{ERE_i} \tag{6.3}$$

The **Confidence-Weighted Ensembling** function gives a final landmark prediction $\mathbf{P}$ which is calculated as a weighted average of each base estimator's predictions:

$$\mathbf{P} = \frac{1}{W} \cdot \sum_{i=1}^{m} (w_i \cdot \mathbf{p}_i) \quad \text{where} \quad W = \sum_{i=1}^{m} w_i \tag{6.4}$$

The final predicted point $\mathbf{P}$ is closer to the more confident predictions. Figure 6.21 shows a comparison between mean average ensembling and Confidence-Weighted ensembling for a prediction. Mean average ensembling is where all individual landmark predictions are averaged with the same weight.

6.6.2 Experiments Performed

The models for which results are reported in this section were trained using the same pipeline as in previous sections. The same hyperparameters were used, such as data augmentation, learning rate, and number of training epochs. The value $\sigma = 3$ was chosen for the Gaussian training heatmaps for all datasets because this value of σ was shown to achieve the most accurate results in Section 5.2.3. An ensemble of size three was used so for each experiment three separate UNet++ models were independently trained on the same data. An ensemble size of three was chosen for practical reasons. To train a model on the full hand datasets took approximately six hours on a GPU machine so an ensemble of three models took a total of 18 hours. Ensembles were therefore chosen to be composed of three models so that the training time did not become infeasible whilst still being enough to show the performance improvement an ensemble gives over a single model.

To test how ensembles perform when limited amounts of training data are available, separate ensembles were trained on different subsets of each dataset's training data. For the Cephalometric dataset ensembles were trained on subsets of 20, 50, 100 and 150 training images. For the Hand, Pelvis and Infant Hip datasets, models were trained on subsets of 20%, 50% and 100% of each training

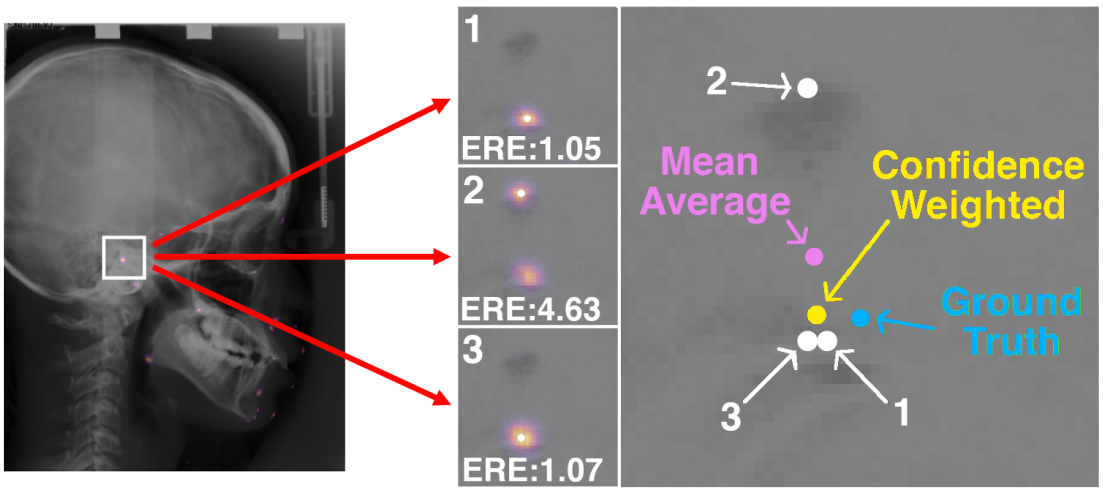


Figure 6.21: This figure shows a comparison between mean average aggregation and Confidence-Weighted aggregation. On the left, a full Cephalometric image is shown with a white box highlighting which landmark the figure on the right pertains to. In the middle of the figure, three individual predictions are shown with their heatmaps and corresponding ERE scores. Prediction two has a higher ERE score because the heatmap has two distinct modes. On the right all three predictions are seen in a single view as well as the ground truth point in blue, the mean average of the individual predictions in pink, and the Confidence-Weighted prediction in yellow. In this example, the Confidence-Weighted prediction is slightly closer to the ground truth than the mean average prediction. It is worth noting that it could be interpreted from this figure that Confidence-Weighted Ensembling relies on inaccurate model predictions to achieve accurate results because model two's prediction moves the Confidence-Weighted prediction upward from model one and three's prediction and closer to the ground truth landmark point. This is however an outlier case and it is shown by results presented in Section 6.6.3 that Confidence-Weighted Ensembling achieves more accurate results the more accurate the individual models are.

set. Once the ensembles had been trained on these different amounts of data they were tested on the corresponding dataset's test set. In this case, the test set for the Cephalometric dataset was selected as Test Set 2. For the Hand dataset the test set in the first training/test set pair proposed by the authors was used. For the Pelvis and Infant Hip datasets the test sets computed by the KS test method in Section 4.4 were used. MRE Results for both Confidence-Weighted Ensembling and Mean Average ensembling over these test sets are reported in the next section. MRE Results are also reported for a single model, as a baseline. This was computed by averaging the MRE of each model in each ensemble.

	Cephalometric Dataset			
	MRE ↓ (mm) for No. Of Training Images			
Method	20	50	100	150
Single Model	1.94	1.50	1.43	1.41
Mean Average	1.82	1.44	1.39	1.38
Confidence-Weighted	**1.61**	**1.42**	**1.38**	**1.36**

Table 6.6: MRE for single models and ensembles trained on 20, 50, 100 and 150 images from the Cephalometric training dataset. The results are the average of three runs. An accuracy improvement of 18% and 11% was obtained by using Confidence-Weighted Ensembling over a single model or mean average ensemble when only 20 training images were available. In addition, a SOTA 1.36mm MRE was obtained on the full dataset using the Confidence-Weighted Ensembling method.

	Hand Dataset		
	MRE ↓ (mm) for % Of Training Images		
Method	20%	50%	100%
Single Model	0.87	0.72	0.68
Mean Average	0.82	0.68	**0.65**
Confidence-Weighted	**0.74**	**0.67**	**0.65**

Table 6.7: MRE for single models and ensembles trained on 20% (119 images), 50% (298 images) and 100% (596 images) of the hand training dataset. The results are the average of three runs. A SOTA 0.65mm MRE was obtained on the full dataset using the Confidence-Weighted Ensembling method. In addition, using a Confidence-Weighted Ensembling on 50% of the training data achieved better accuracy than a single model trained on the entire dataset.

	Pelvis Dataset		
	MRE ↓ (mm) for % Of Training Images		
Method	20%	50%	100%
Single Model	3.94	2.66	2.35
Mean Average	3.50	2.56	**2.26**
Confidence-Weighted	**3.04**	**2.51**	**2.26**

Table 6.8: MRE for single models and ensembles trained on 20% (52 images), 50% (131 images) and 100% (262 images) of the Pelvis training dataset. The results are the average of three runs. An accuracy improvement of 23% and 14% was obtained using Confidence-Weighted Ensembling over a single model or mean average ensemble when only 20% of images were used to train the models.

	Infant Hip Dataset		
	MRE ↓ (mm) for No. Of Training Images		
Method	20%	50%	100%
Single Model	9.69	7.92	7.04
Mean Average	8.60	7.27	6.72
Confidence-Weighted	**8.33**	**7.22**	**6.57**

Table 6.9: MRE for single models and ensembles trained on 20% (23 images), 50% (47 images) and 100% (95 images) of the Infant Hip training dataset. The results are the average of three runs.

6.6.3 Ensembling Results

Results for the Cephalometric, Hand, Pelvis and Infant Hip datasets are given in Tables 6.6 to 6.9 respectively. Confidence-Weighted Ensembling outperformed both an individual model and the mean average aggregation methods on all datasets for all percentages of training images. The results show that the fewer training images the more effective the weighted ensembling method was relative to an individual model or mean average ensemble, for all datasets. For example, the Cephalometric data results in Table 6.6 using Confidence-Weighted Ensembling reduced the MRE from 1.94mm for a single model and 1.82mm for the mean average ensemble to 1.61mm when 20 images are used for training. This is a performance improvement of 18% and 11% respectively, as opposed to when the full training set which produced an improvement of only 2-3%. The reason ensembles outperformed single models, especially for smaller datasets, is that models trained on smaller datasets have higher variance among their predictions and more frequently give erroneous output. Confidence-Weighted Ensembling chooses confident predictions from the ensemble and thus cuts out many erroneous predictions. It is important to note that this does not mean that Confidence-Weighted Ensembling relies on inaccurate predictions to achieve accurate results (as discussed in the caption of Figure 6.21). Although Confidence-Weighted Ensembling achieved less of a percentage improvement over a single model when accurate models, trained on 150 images, were used, it still produced the most accurate result of 1.38mm MRE in this case.

Future work should assess how the number of individual models in an ensemble affects the accuracy of the ensemble. In this section, all ensembles described were made up of three models but this is a hyperparameter which can be changed. The more models an ensemble is composed of the more accurate the ensemble is likely to be. This is because the more models there are in the ensemble the more models there will be that produce highly confident and accurate predictions which make the final Confidence-Weighted prediction more accurate. This should be validated in future work. Nevertheless, the performance increase of up to 8% over a single model (in the case of hand dataset) using an ensemble of three models trained

on 50% of the training images is a breakthrough. Clinicians have limited time to manually label datasets but these results show the same performance can be obtained by using an ensemble trained on only 50% of the data, to a single model trained on all data, which will save hours of annotation time.

6.6.4 Uncertainty in an Ensemble

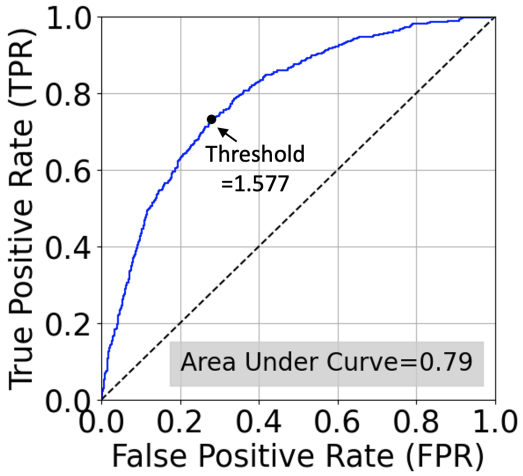
This section details how a final uncertainty measurement for a prediction made by an ensemble can be obtained so that potentially erroneous predictions can be flagged to the user like in Section 6.5. This combines the advantage that ensembles are more accurate than single models with the ability to flag unacceptable predictions. The simplest solution is to average the ERE scores of the individual models to obtain a final uncertainty measurement for each prediction, and this is shown to be effective. This can be formally defined as follows. Each model i in an ensemble of m models produces a prediction $\mathbf{p}_i$ with an associated ERE score ERE_i. The uncertainty score U for each prediction is calculated as:

$$U = \frac{1}{m} \cdot \sum_{i=1}^{m} ERE_i \tag{6.5}$$

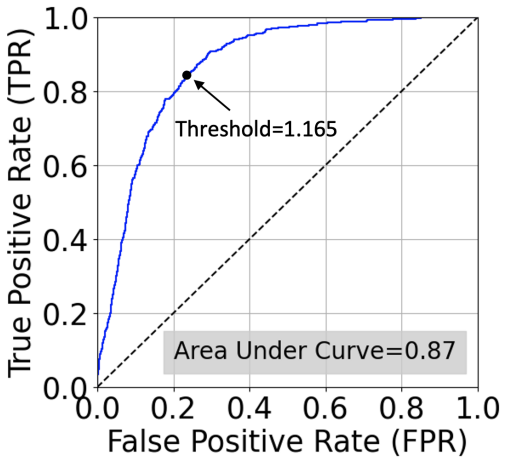
Experiments Performed

Models are less accurate when trained on less data and therefore an alerting system is even more crucial in this context. Therefore results are reported in this section for the ensembles trained on only 20% of the training data for each dataset, or only 20 images in the case of the Cephalometric dataset. These ensembles were described in Section 6.6.2. The Confidence-Weighted Ensembling method was used to aggregate landmark predictions and the uncertainty score U was calculated for each landmark in each validation set. These were the same validation sets detailed in Section 6.2.2. ROC curves were then computed using U and the radial errors. The same clinically acceptable radial error thresholds stated in Section 6.5.2 were used for each dataset. Once ROC curves were computed the ERE threshold which maximised $TPR \times (1 - FPR)$ was calculated in the same way as described in

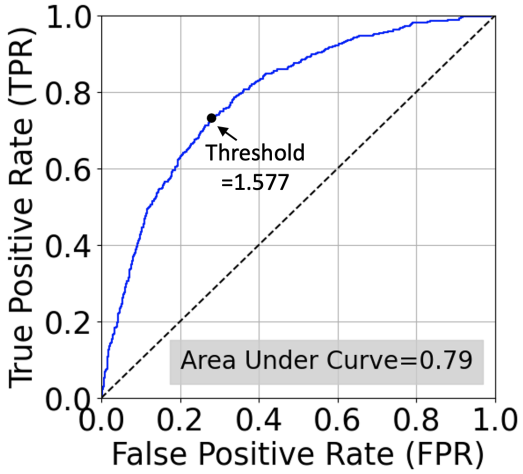
Section 6.5.2. These models were then run over the images in for each dataset's test set. Alongside the radial errors which were reported in Section 6.6.3, the uncertainty score U was also calculated for each prediction. These landmarks were then assigned to either the flagged or accepted category and the MRE for landmarks within each of these sets was calculated and is reported in the following results section.



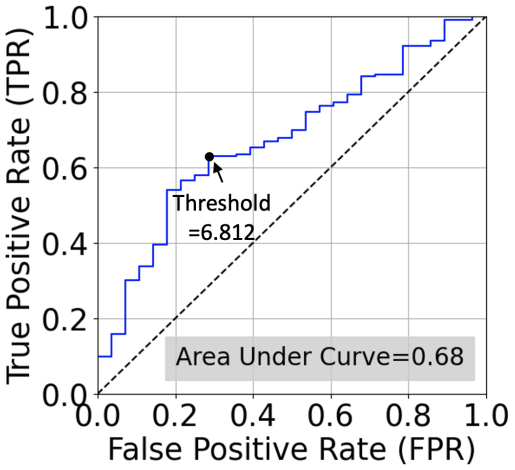
(a) ROC curve for an ensemble with three constituent models trained on 20 images from the Cephalometric dataset with the accepted landmark tolerance set to 2mm.



(b) ROC curve for an ensemble with three constituent models trained on 20% of the hand training set with the accepted landmark tolerance set to 2mm.



(c) ROC curve for an ensemble with three constituent models trained on 20% of the Pelvis training set with the accepted landmark tolerance set to 4mm.



(d) ROC curve for an ensemble with three constituent models trained on 20% of the Infant Hip training set with the accepted landmark tolerance set to 10 pixels.

Figure 6.22: ROC curves show the capability of an acceptable versus unacceptable landmark classifier predicated on the mean average of the ERE scores in the ensemble. The ensemble used in each case is Confidence-Weighted Ensembling. Each ROC curve shows an effective classifier can be constructed. The position of the threshold value that maximises $TPR \times (1 - FPR)$ is shown on each curve. However, the threshold can also be selected based on the clinician's preference for precision or recall.

Set	% Train Data	# Landmarks	MRE (mm)	SDR (%)			
				2mm	2.5mm	3mm	4mm
Overall	20	1900	1.52	75.16	82.32	87.90	94.84
Accepted	20	1593	1.35	79.79	86.19	90.71	96.67
Flagged	20	307	2.42	51.14	62.22	73.29	85.34

Table 6.10: Statistics for accepted versus flagged landmarks for a Cephalometric ensemble of three models trained on 20 images. A threshold of 1.577 was calculated based on the ROC curve. Accepted landmarks have a superior MRE to previous SOTA methods trained on the entire dataset. In total, 16.16% of landmarks were flagged as potentially unacceptable and these have an MRE 1.59 times less accurate than the MRE of the entire test set.

Set	% Train Data	# Landmarks	MRE (mm)	SDR (%)		
				2mm	4mm	10mm
Overall	20	11026	0.72	94.57	99.18	99.86
Accepted	20	9921	0.62	96.46	99.62	99.98
Flagged	20	1105	1.65	77.56	95.29	98.82

Table 6.11: Statistics for accepted versus flagged landmarks for a hand ensemble of three models trained on 20% (119 images) of the training set. A threshold of 1.165 was calculated from the ROC curve. Accepted landmarks have a superior MRE to previous SOTA methods trained on the entire dataset. In total, 10.00% of landmarks were flagged as potentially unacceptable and these have an MRE 2.29 times less accurate than the MRE of the entire test set.

Set	% Train Data	# Landmarks	MRE (mm)	SDR (%)		
				4mm	6mm	10mm
Overall	20	448	2.97	80.58	89.51	95.69
Accepted	20	303	2.13	89.11	95.38	98.68
Flagged	20	145	4.74	62.76	77.241	89.66

Table 6.12: Statistics for accepted versus flagged landmarks for an ensemble of three models trained on 20% (52 images) of the Pelvis training set. A threshold of 2.271 was calculated from the ROC curve. Accepted landmarks have a superior MRE to all landmarks predicted by single models and ensembles trained on the full data as shown in 6.8. In total, 32.36% of landmarks were flagged as potentially unacceptable and these have an MRE 1.60 times less accurate than the MRE of the entire test set.

Set	% Train Data	# Landmarks	MRE (pixels)	SDR (%)		
				10 pixels	15 pixels	20 pixels
Overall	20	235	8.72	71.06	87.23	93.62
Accepted	20	98	5.60	85.71	94.90	98.98
Flagged	20	137	10.95	60.58	81.75	89.78

Table 6.13: Statistics for accepted versus flagged landmarks for an ensemble of three models trained on 20% (19 images) of the Infant Hip training set. A threshold of 6.812 was calculated from the ROC curve. Accepted landmarks have a superior MRE to all landmarks predicted by single models and ensembles trained on the full data as shown in 6.9. In total, 58.30% of landmarks were flagged as potentially unacceptable and these have an MRE 1.26 times less accurate than the MRE of the entire test set.

Ensembling Results Using Uncertainty

The ROC curves for the ensembles are shown in Figure 6.22 and have AUC scores within 10% of the AUC scores obtained using single models trained on the whole dataset, except Infant Hip which has an AUC score of 0.68 compared to 0.78. The reason for these reduced AUC scores is that the models were only trained on 20% of data which for ultrasound was only 23 images. Training on a dataset of this size not only negatively affects the landmark prediction accuracy but also the correlation between ERE score and landmark accuracy. Nevertheless, all ROC curves represent classifiers significantly better than random chance and as such they were used to flag potentially erroneous predictions. From these ROC curves, ERE thresholds of 1.577, 1.165, 2.271 and 6.812 were calculated to maximise $TPR \times (1 - FPR)$. These ERE thresholds were then used to flag or accept landmarks in the test sets. Tables 6.10 to 6.13, show the accuracy of the accepted versus flagged predictions in the test set for an ensemble trained on 20% of the training set. Table 6.10 shows that the accepted predictions were more accurate than all previous models trained on the full Cephalometric dataset. This shows that even if very limited data is available a landmark detection system can be created where a model makes predictions that are either accepted, and as accurate as a SOTA approach trained on all data, or flagged as needing manual adjustment. The same conclusion can be drawn from the other tables in this section. The accepted predictions for the hand dataset on 20% of the data were more accurate than all past methods in previous literature. Tables 6.12 and 6.13 additionally show that the localisation accuracy of the accepted landmark set for the Pelvis and Infant Hip dataset was superior to the overall localisation accuracy of single and ensemble models displayed in Tables 6.8 and 6.9. This is particularly remarkable because these ensembles were only trained on 52 and 19 images respectively. These results show that, despite these ensembles being trained on limited data, they do not overfit because they obtain good results on the unseen images in the test set. This shows the potential of using the ensemble method combined with an alerting system to produce a landmark detection system with very limited training data.

6.6.5 Ensembling Conclusions

Ensembling individual landmark detection models is a fast and effective way of increasing overall landmark accuracy especially when limited training data is available. The main advantage of this approach over previous literature is that no new architectures are required or need to be retrained to achieve an accuracy improvement. The Confidence-Weighted Ensembling method using the ERE statistic outperforms mean average ensembling and is the SOTA in the landmark ensembling space. The success of this method proves the hypothesis that, if the aggregated final prediction of an ensemble is closer to the individual predictions that are more confident, then the final prediction will be more accurate than the average of the individual predictions. In addition to serviceable results on limited training data, the Confidence-Weighted Ensembling achieves the most accurate localisation results seen in the literature on the Cephalometric and Hand datasets. This provides future validation of the ERE statistic.

More interestingly, an additional element of novelty is proposed where uncertainty can be passed through the ensemble by averaging the ERE scores of the individual models. This is especially powerful because the accurate results the ensembles achieve on limited training data can be combined with the ability to flag to the user potentially incorrect predictions. Results show that when models are trained on only 20% of the training data the accepted landmarks in the test set for each dataset are as accurately located as the average landmark in the test set when all training data is used.

In conclusion, all landmark detection systems would benefit from using an ensemble of models, to obtain the maximum accuracy, with an uncertainty reading to ensure the safety of patients and build trust in AI. The Confidence-Weighted Ensembling method makes landmark detection systems viable even with very limited training data.

6.7 Hip Studio

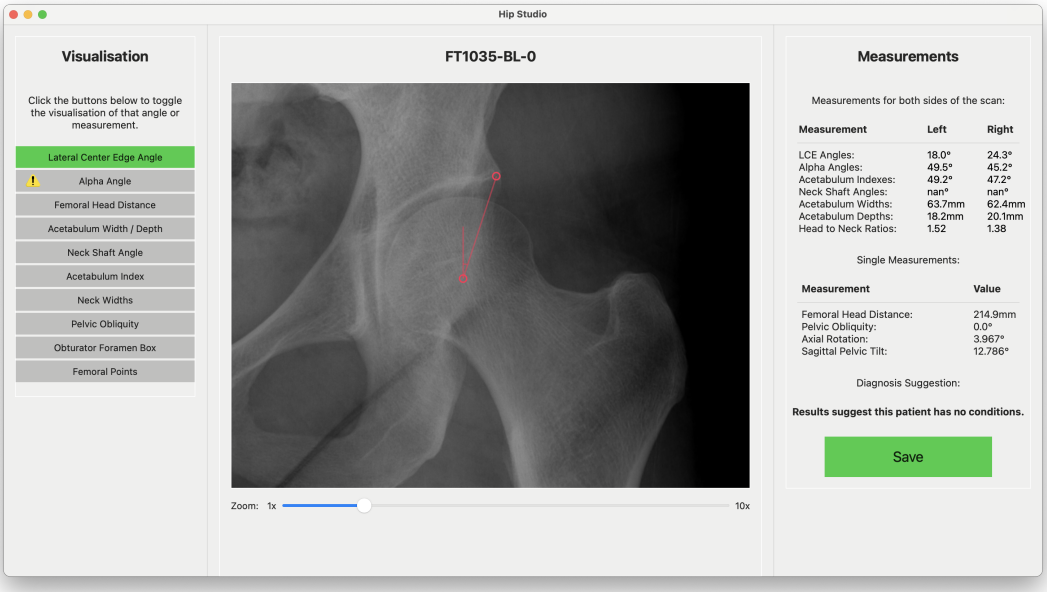


Figure 6.23: A screenshot from Hip Studio showing an automatically placed LCE-angle. The user is then free to adjust these landmarks and alter the proposed LCE-angle. The angle readings are shown on the right panel under 'Measurements'.

Alongside the research work in my DPhil, I also developed software (called Hip Studio) which implements the research proposed in this chapter. This program, shown in Figures 6.23 to 6.26, has a graphical user interface that clinicians can use. The user can open a scan of a Pelvis or Infant Hip inside the software and the relevant clinical landmarks are automatically placed on the image. It automatically joins these landmarks to visualise the line segments and angles. For Pelvis, LCE-angle and α-angle are automatically calculated and for Infant Hip, α-angle and β-angle are automatically calculated. The angles in degrees are shown to the user in a separate panel and can be exported from the software.

Hip Studio has three key features which, as highlighted in previous conclusion sections, are required for automatic landmark detection to be adopted in the clinic.

1. In Hip Studio clinicians have the ability to adjust landmarks once the model has predicted their locations (as discussed in Section 5.6). This feature means the clinician is in the decision loop and can quickly adjust any incorrect predictions the AI makes.

2. Hip Studio has an alerting system as described in Section 6.4. This system uses the ERE score and the thresholding system proposed to alert the user when a landmark is likely to be placed incorrectly. The ERE scores for the alerting system are hidden from the user so that the user simply sees a button to approve landmarks which the model has flagged as potentially incorrect. This is shown in Figure 6.25. This system enforces that the user checks these landmarks by drawing their attention to them and makes them confirm their placement with an approve button.

3. Hip Studio automatically suggests a diagnosis to the clinician if the angles predicted are above a certain threshold. This was discussed and results were presented in Section 5.4 and Section 5.5. In the case of automatically diagnosing FAI thresholds of 40° and 65° are chosen for LCE-angle and α-angle respectively. For diagnosing DDH from ultrasound a threshold of 50° is applied to the α-angle like in Section 5.4. If one of the angles is above the relevant threshold, red text alerts the clinician of a possible diagnosis as shown on the right of Figure 6.26. The α-angle threshold of 50° has since been superseded but this threshold was used based on the medical advice at the time.

Hip Studio takes approximately five seconds to make predictions, calculate which landmarks are potentially unacceptable and should be flagged to the user, and suggest a diagnosis for each image. As a result, if it were integrated into PACS software in hospitals, this technology has the potential to speed up the process of taking these measurements.

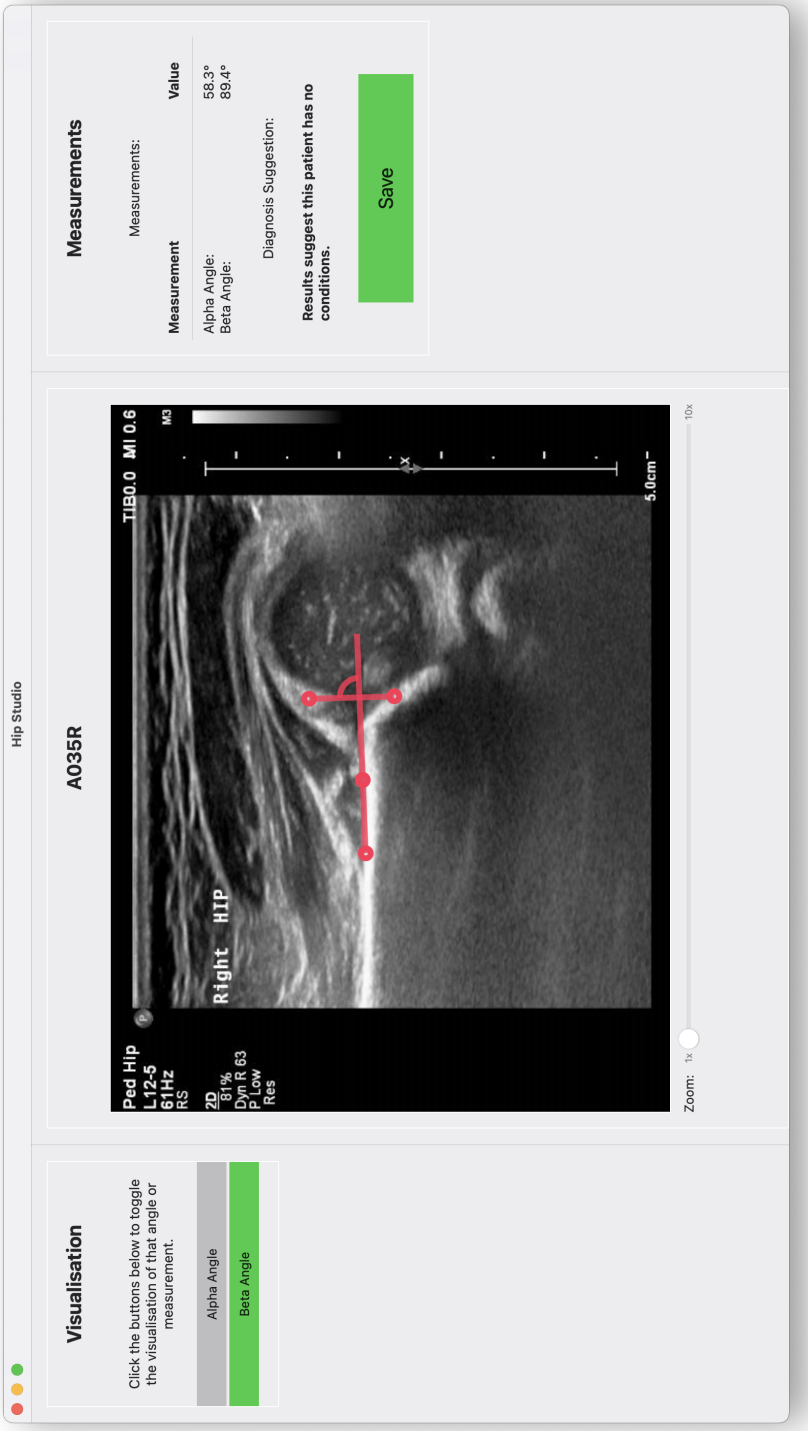


Figure 6.24: A screenshot from Hip Studio showing an automatically placed β-angle on an Infant Hip scan. α-angle and β-angle measurements are shown for this patient in the right panel. This figure shows that the Hip Studio software works for ultrasound images as well as X-ray images shown in Figure 6.23.



Figure 6.25: A screenshot from Hip Studio showing an automatically placed α-angle. Behind the interface of Hip Studio ERE scores are calculated for each landmark. The ERE score for the lateral cam point is above the chosen threshold and so an approve button is shown to alert the user that the software is unsure about this landmark.

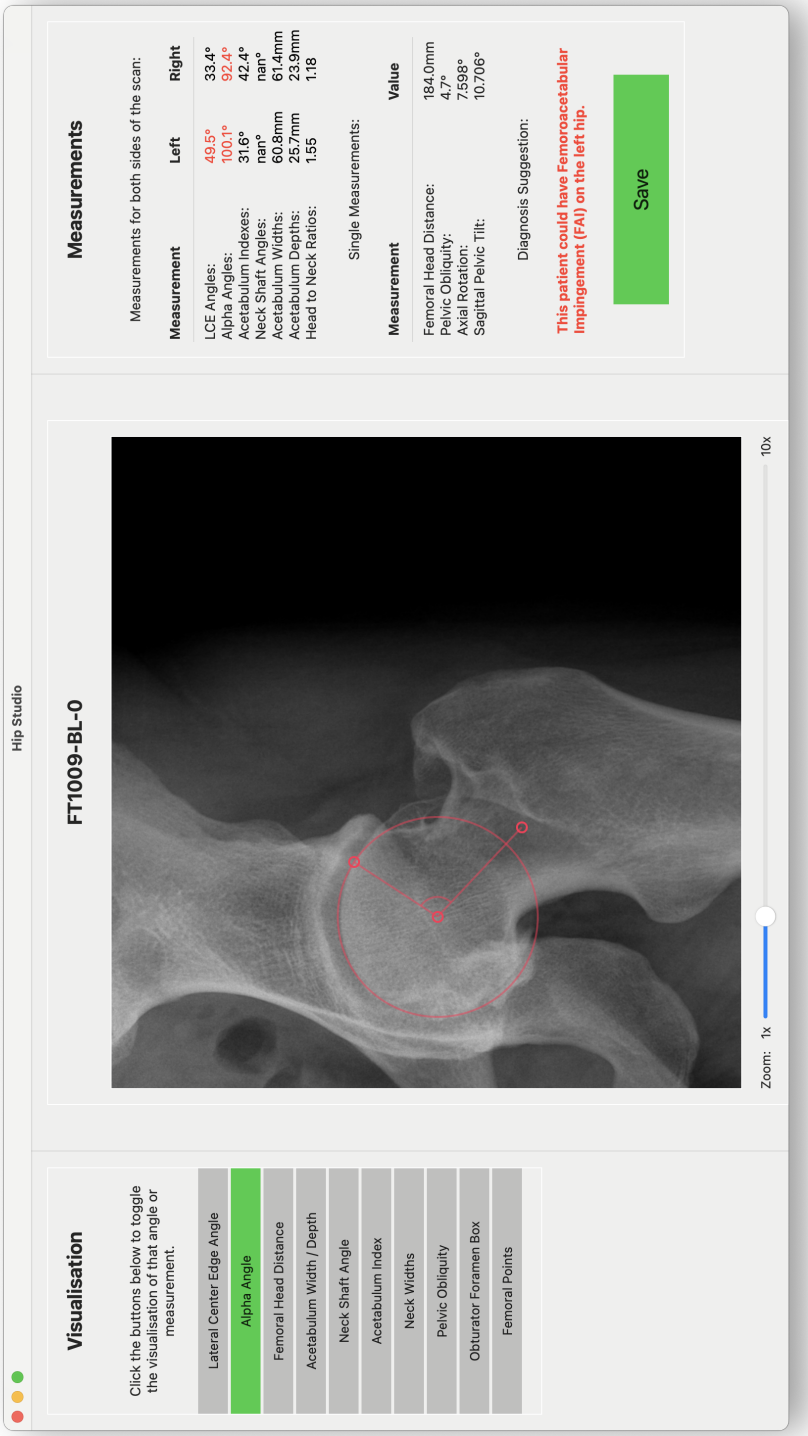


Figure 6.26: This screenshot shows a patient with a high α-angle on their left hip. The red text on the right-hand-side panel suggests that this patient has FAI on their left hip because the AI predicted an α-angle over 65 degrees.

7 Conclusion

Contents

7.1 Summary 191
7.2 Filling the Gaps in the Literature 194
7.3 Future Work 197
7.3.1 Obtaining Data 197
7.3.2 Experimental Setup for Diagnostic Tasks 198
7.3.3 Applying Uncertainty to Angle Predictions 198
7.3.4 Using Uncertainty to Detect Incorrect Labels 199
7.3.5 Improving Hip Studio 199
7.3.6 Translating Software into Clinic 200

7.1 Summary

Instead of following the norm in the medical imaging landmark detection area by proposing another network architecture which achieves slightly improved results on a publicly available dataset, I have taken the emerging field of medical imaging landmark detection in three distinctly different directions.

Firstly, there is a greater emphasis on the importance of the images allocated to the training, validation and test set in this work when compared to previous work. In Chapter 4 a new Kolmogorov-Smirnov method is proposed for allocating images to these sets to ensure that they are balanced with respect to chosen

parameters, for example, the image brightness and patient information such as their condition severity, their age or their ethnicity. Balancing the training, validation and test set ensures that the model will learn from a representative subset of the population and will not be biased towards a specific group. If future models are to be deployed for clinical use, they must be trained on representative and varied data to improve robustness and confidence. The work presented in Chapter 4 provides a viable solution.

The second, new direction presented in this work is how landmark detection can be used to diagnose patient conditions in orthopaedics. Previous methods in landmark detection for medical imaging were validated by finding the distance between the predicted landmarks and the ground truth landmarks. Whilst that analysis is performed, in this work I take it a step further and calculate angles and then a diagnosis from these landmarks. In Chapter 5 it is described how Femoroacetabular Impingement and Developmental Dysplasia of the Hip diagnosis were calculated automatically using landmark detection models on images of the pelvis and hip. Additionally, I show that SOTA results can be obtained by the established and easily reproducible UNet++ model and validate this over all four datasets which were experimented with. The summation of these findings is that landmark detection models can diagnose conditions on a broad range of body parts and image modalities and are easy to implement for other researchers.

The final, major direction of the work presented in this thesis is the addition of an uncertainty measurement with landmark predictions. The majority of previous work focused on obtaining maximum landmark detection accuracy but with no uncertainty measurement attached. This is dangerous because all models will make mistakes and it is important that these mistakes can be flagged in a clinical pipeline so they can be fixed by a clinician before the errors are passed onto the patient in the form of an incorrect diagnosis or treatment. This is not unique to landmark detection but all areas where AI is used in the clinical pipeline. Uncertainty is becoming a necessary component of all systems that want to be used in practice. In Chapter 6 a method for obtaining an uncertainty measurement, named Expected Radial Error,

was proposed. The results presented in this chapter showed it correlated highly with the accuracy of the landmark predictions. Crucially no new architecture or model had to be trained to obtain uncertainty measurements using this approach. Later in the chapter, a method was described for using this uncertainty measurement to flag inaccurate predictions and results show this method to be effective. Finally, a piece of software I developed during my DPhil called Hip Studio was presented which shows how this technology can be built into a product to improve the efficiency and accuracy of measurements taken in the clinic whilst providing clinicians with the ability to edit landmark predictions to ensure they are correct.

7.2 Filling the Gaps in the Literature

The current gaps in landmark detection for medical imaging literature, and landmark detection uncertainty literature, were listed in Section 2.5.3 and Section 2.6.4 respectively. The gaps in landmark detection for medical imaging literature have been filled by work presented in this thesis in the ways listed below. The headings in bold correspond with the headings used in Section 2.5.3.

1. **Reproducibility** - The UNet++ architecture is used for all experiments described in this thesis after it was shown to outperform the U-Net architecture in Section 5.2.4 and obtain SOTA results in Section 5.3. The UNet++ architecture is easy to reproduce because it is implemented in the popular machine learning library `pytorch segmentation models`. The code used for all experiments in this thesis is available online and provided in Appendix E. Additionally, the Cephalometric and Hand datasets experimented with are publicly available and are therefore available to other researchers.

2. **A Lack Of Images Taken During Routine Practice** - The images in the Infant Hip dataset were taken during routine practice and this dataset is used for all experiments described in this thesis. The results presented in this thesis are an incentive for clinicians to share a larger number of such images and explore the clinical applicability of these methods.

3. **Experimental Robustness** - The Kolmogorov-Smirnov method for computing a single representative training, validation and test set was proposed in Chapter 4. This method was used to create representative training, validation and test sets for the Pelvis and Infant Hip datasets which were used for all experiments described in this thesis. The training, validation and test sets used for the Cephalometric and Hand datasets were the same as those defined by the original authors of those datasets. Hyperparameter tuning experiments described in Section 5.2 were performed on validation sets for each dataset and as such there was no data leakage for any experiment performed in this thesis.

4. **Uncertainty** - A new method for quantifying the uncertainty of landmark detection models which train on heatmaps was proposed in Chapter 6. This method calculates a value called the Expected Radial Error (ERE) from the model's output heatmaps and the results given in Section 6.4.3 show that this value is correlated with the radial error of landmark predictions. Uncertainty as a general concept is becoming an increasingly popular concept in the machine learning community.

5. **Requires Training Data** and **Few Ensembling Approaches** - An ensemble approach called Confidence-Weighted Ensembling is proposed in Section 6.6 based on the ERE statistic. Section 6.6.3 shows that this approach is effective at improving the landmark detection results when limited training data is available.

The gaps specifically identified in uncertainty landmark detection literature have been filled by work presented in this thesis in the ways listed below. The headings in bold correspond with the headings used in Section 2.5.3.

1. **Retraining Required** - The method for quantifying uncertainty in this thesis calculates the ERE statistic from the output heatmaps of a trained landmark detection model. As long as a landmark detection model has been trained on Gaussian training heatmaps then no new network architecture needs to be implemented or trained for another researcher to obtain uncertainty measurements using this method.

2. **Assuming Gaussian Distribution** - As demonstrated in Section 6.1, selecting the final landmark position as the mean of an output heatmap will not produce accurate predictions when the true landmark position is likely to have been one of two distinct locations. In other words, the output heatmap is multi-modal. The work presented in Chapter 6 makes no such assumption and instead uses the mode of the output heatmap as the final landmark prediction position.

3. **Limited Heatmap Analysis** - Reliability diagrams in Section 6.2 were used to analyse whether the values in a model's output heatmaps were representative of the true probabilities of the true landmark position. Although the accuracy of this approach is limited for small datasets, it was shown on the Cephalometric and Hand datasets that different training heatmap generation methods result in more or less calibrated landmark detection models.

4. **Lack of Practical Analysis** - In Section 6.5 it was described how the ERE statistic can be used to flag potentially incorrect landmark detection predictions. Accuracy results are given for flagged landmarks and accepted landmarks for each dataset. The results shown in Section 6.5.4 show that this method consistently flags predictions that are more likely to be inaccurate across all datasets.

7.3 Future Work

7.3.1 Obtaining Data

Thousands more pelvis X-ray images should be obtained from a variety of hospitals and acquired using different scanners. The research group to which I belong has recently obtained more than ten thousand Infant Hip images although they were not available to me during my DPhil. It is important in both the Pelvis and Infant Hip cases that images are collected during routine clinical practice to ensure that any model subsequently trained on this data will be able to perform when it is deployed in a clinic. Moreover, the images in the Pelvis and Infant Hip datasets used in this thesis belong to patients who either presented in the clinic with evidence of a hip condition or who had a family history of a condition. In future, images from control patients should also form part of the datasets. Metadata such as age and existing hip conditions were provided with each image in the Pelvis dataset used in this thesis and that metadata should also be gathered for all future images acquired. This is so that the Kolmogorov-Smirnov method described in Chapter 4 can be used to distribute patients into training, validation and test sets proportionality and avoid any subsequent model bias. Additionally, multiple clinicians should annotate landmarks on each image multiple times. These additional measurements will provide the intra-clinician, and inter-clinician, variability statistics for each landmark and for subsequent angle measurements and diagnosis which can be compared with a trained model. This way it can be deduced whether the model is producing predictions within the range expected from clinicians. This is especially important for diagnosing FAI from pelvis images where there can be disagreement between clinicians regarding the existence of small bumps on the femoral head as shown by the high inter-rater variability scores given in previous literature [4].

7.3.2 Experimental Setup for Diagnostic Tasks

As well as future work to obtain more data, how this data should be allocated to training, validation and test sets when the objective is to diagnose conditions from the images should be explored. The Kolmogorov-Smirnov method described in Chapter 4 ensured that the images in the Pelvis dataset were evenly distributed into training, validation and test sets used for experimentation with respect to their α-angle measurements. However in Section 5.6.3 it is stated that the model trained from this data tended to underestimate the α-angle for patients' hip and, as a result, would have missed cases of FAI leading to a knock-on effect on patients if the system were deployed in clinic. Likely, one reason the model tended to underestimate the α-angle measurement was that there were significantly more patients in the dataset that had hips with low α-angle than high α-angle and as such the model became biased towards giving low α-angles. Future work should investigate a solution to this. It is possible that simply gathering more data could fix this problem because the model would learn from more examples of hips with high α-angles. Alternatively, a technical solution could be to oversample hips with a high α-angle during the training process to artificially ensure that the model is not biased towards a specific diagnosis.

7.3.3 Applying Uncertainty to Angle Predictions

In future work, the heatmaps output from landmark detection models can be interpreted as distributions and propagated to downstream angle measurement calculations. This would mean that instead of outputting a single number for an angle prediction, such as α-angle, a distribution or likely range for that angle could be given. A range would provide the clinician with an idea of the uncertainty of predicted angles. Whilst this thesis has presented a method to quantify landmark uncertainty, providing angle uncertainty to the clinician would be additionally beneficial. It would be especially helpful in cases such as α-angle prediction on Infant Hip images, because, as shown in Section 5.5, inaccurate landmark predictions do not necessarily lead to inaccurate α-angle predictions. Moreover, the clinician

would be able to observe which angle predictions have a range overlapping with the diagnostic threshold for that angle. The clinician would then be able to view the distribution of the predicted angle to make it easier for them to deduce which side of the threshold a patient lies. This would lead to better patient treatment because, as described in Section 5.4 and 5.5, several of the inaccurate FAI and DDH diagnosis predictions produced by the models were on patients who had angles very close to the diagnostic thresholds.

7.3.4 Using Uncertainty to Detect Incorrect Labels

In Section 6.5.5 it is stated that several of the landmark predictions flagged as potentially incorrect using the ERE-based method proposed in Section 6.5 are labelling errors. This was deduced by observating that their placement differed from the definition for that landmark and from the ground truth position of the same landmarks on other images. A promising avenue for future work is to validate this on a statistically significant dataset. If a larger dataset of pelvis images could be obtained then a trained model could be run over all these images and the flagged landmarks could then be sent back to the clinicians who labelled the data. These clinicians could then either confirm that the original ground truth is correct and the model has made a mistake or that the original ground truth is incorrect and should be adjusted. The model can then be retrained on the adjusted landmark positions. Additionally, if the majority of the flagged landmarks are labelling errors validated by clinicians then it could be argued that the model is more reliable than a clinician.

7.3.5 Improving Hip Studio

A current limitation of the Hip Studio software presented in Section 6.7 is that it requires the patient to be in the correct alignment and the scan to be taken in precisely the right direction. This is a limitation because not all images taken in the clinic are consistently taken or of sufficient quality. However, currently Hip Studio will not alert the clinician to this and will make, potentially incorrect, predictions regardless. For example, in the case of pelvis X-rays if the patient is standing

with their feet facing outwards then their hips will be rotated and the landmark predictions will not be accurate. Alternatively, in the case of the Infant Hip images the sonographer takes several images before they decide which one is showing the most appropriate view. For example, the ischium must be visible in the image. Future work should explore whether Hip Studio could be built to detect whether an image is suitable in both the Pelvis and Infant Hip case by detecting more landmarks (or the absense of landmarks) and notifying the clinicians of this.

This idea could even be taken a step further in the case of Infant Hip ultrasound data by running Hip Studio in real-time as the sonographer is moving the probe over the patient. Hip Studio could proactively guide the sonographer on where to place the probe to get the best possible view of the patient's hip.

7.3.6 Translating Software into Clinic

As well as technological research and development there are several other tasks which need to be completed to translate this technology into clinics. The Medicines and Healthcare Products Regulatory Agency (MHRA) sets out guidelines that medical devices must adhere to in the UK. In addition, a UKCA marking (the UK equivalent of the European CE mark) must be obtained.[1] A clinical trial may have to be conducted where the effect of this technology on patients and clinicians is assessed and validated robustly. To carry out this translational work experts in regulation, clinical trials and preparing software for auditing would have to be consulted. In addition, the university's translational office (Oxford University Innovation (OUI) in the case of the University of Oxford) would need to be involved to bring in these experts and to facilitate the commercialisation and licensing of the technology.[2]

[1] `https://www.gov.uk/guidance/regulating-medical-devices-in-the-uk`
[2] `https://innovation.ox.ac.uk/`

Appendices

A
Dataset Images

This appendix shows example images from the Cephalometric, Hand, Pelvis and Infant Hip datasets in Figure A.1 to Figure A.4. Each figure shows a diverse selection of images to provide a comprehensive overview of the different image characteristics that occur in each dataset.

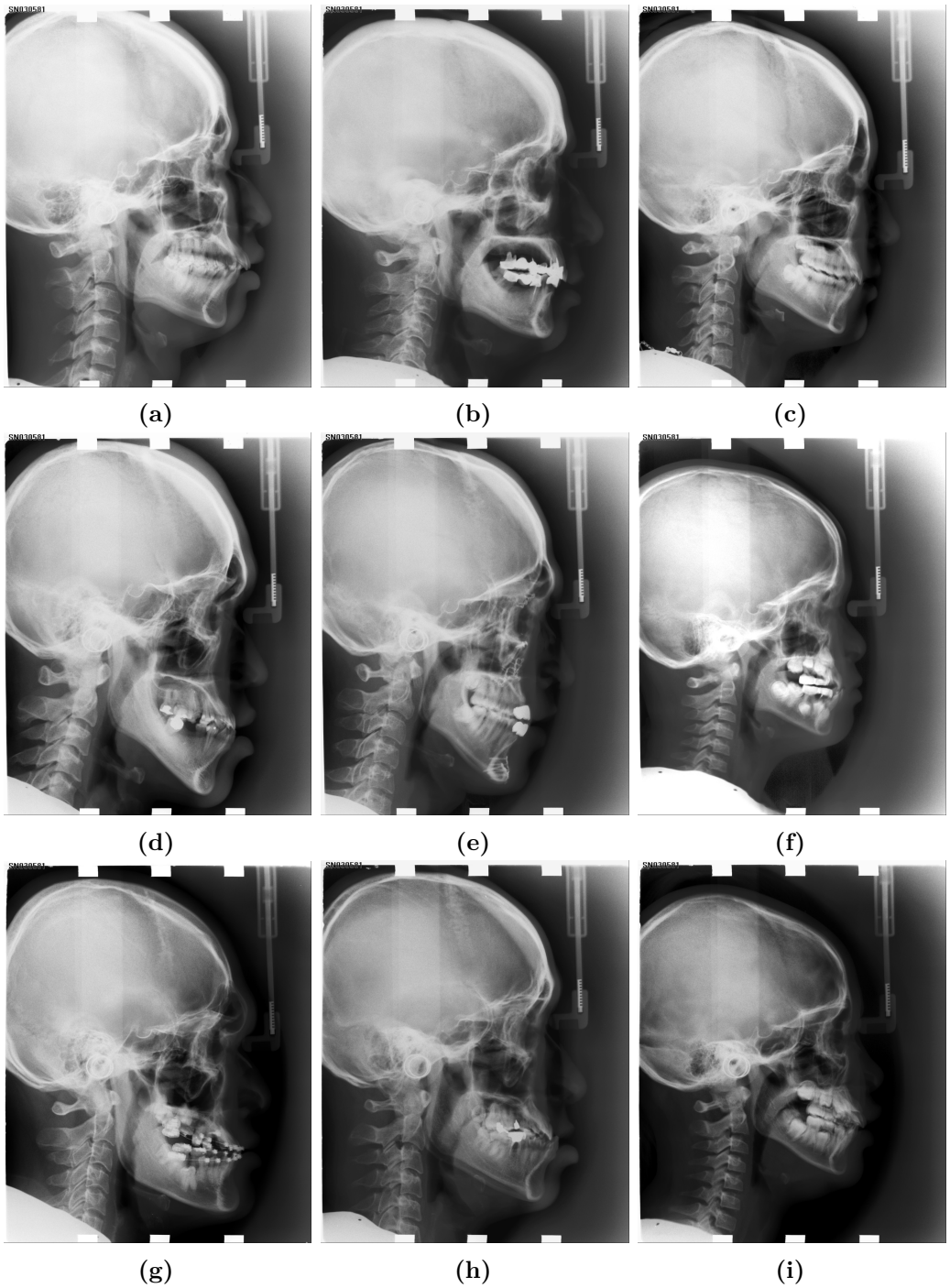


Figure A.1: These figures show the variation of images within the Cephalometric dataset. The images vary slightly in brightness which can be seen by comparing A.1a to A.1i. The images reveal variations in patients' head sizes, for example A.1g, captures a large head where the patient's chin is close to the bottom of the image. In contrast, the head of the patient captured in A.1f is small. Additionally, some images capture patients who have implants as shown in A.1b, A.1f and A.1g. All images capture a patient's head in a similar position with the patient looking straight ahead.

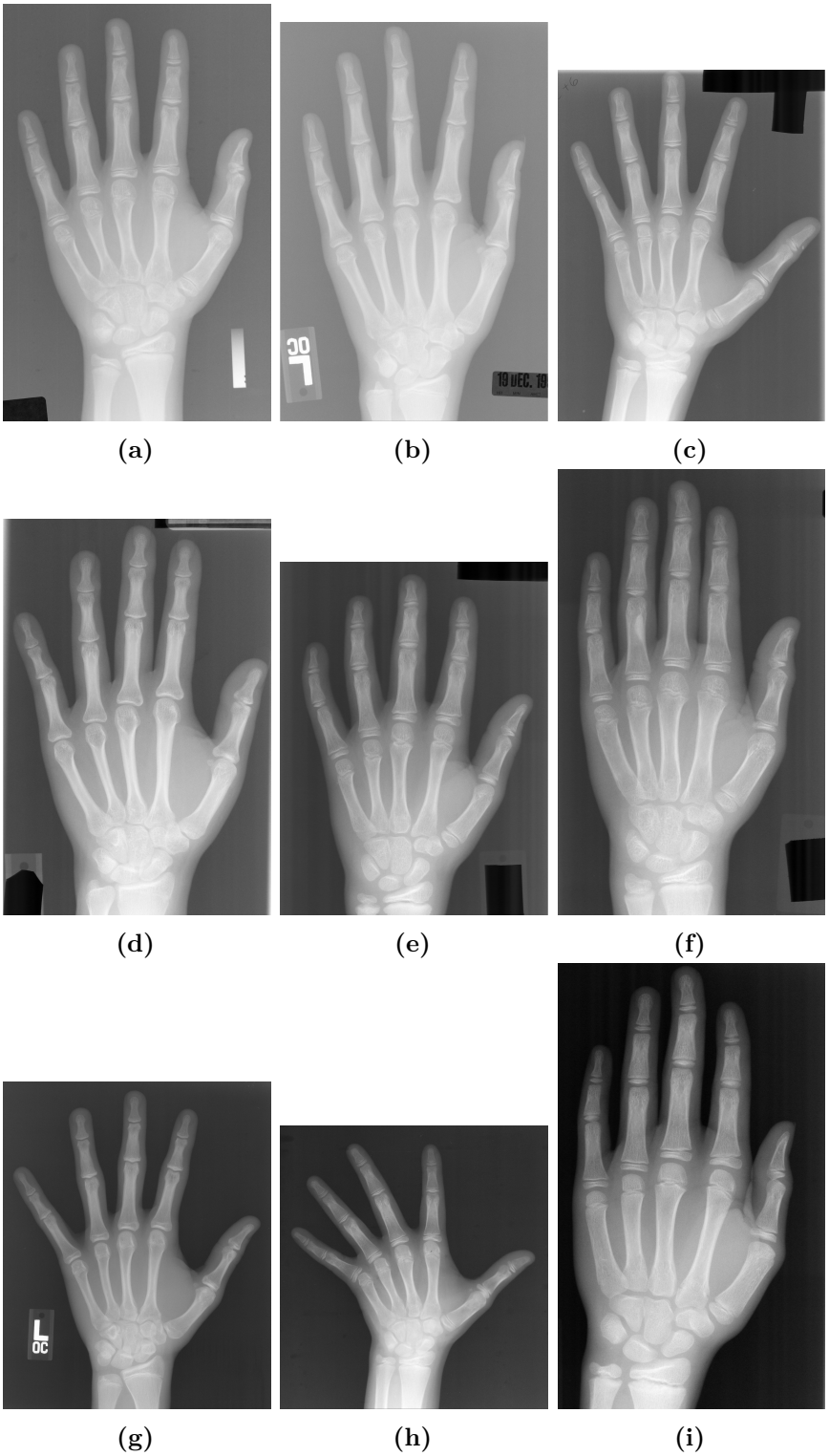


Figure A.2: These figures show the variation of images within the Hand dataset. The brightness of the background of the images varies significantly which can be seen by comparing A.2a to A.2i. In some images, such as Figure A.2f, the fingers of the person captured are together whereas in others, such as A.2h, the fingers are splayed. In addition, some images either have black shapes in the background to mask personal information, such as A.2e, or the letter 'L', such as in A.2g.

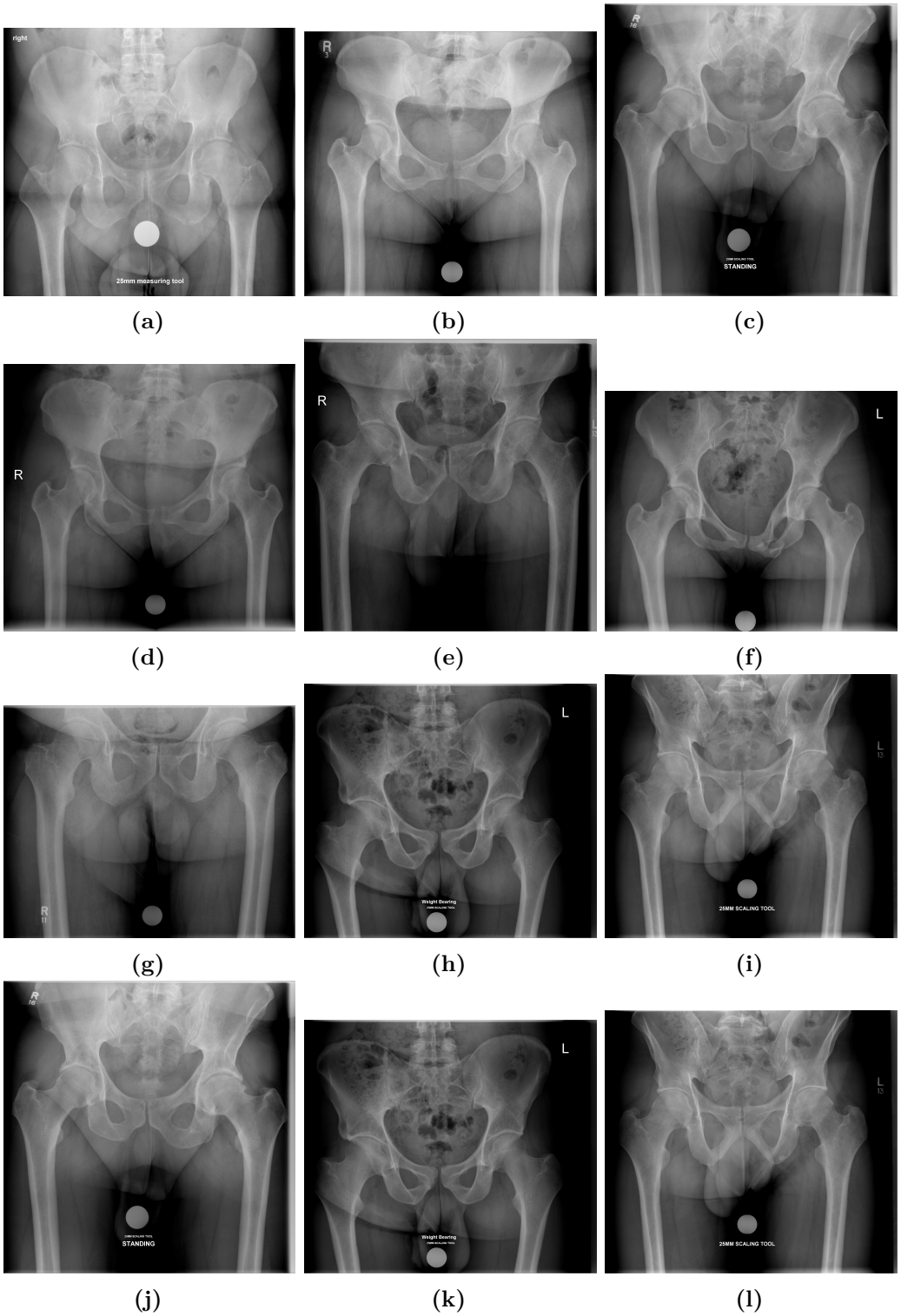
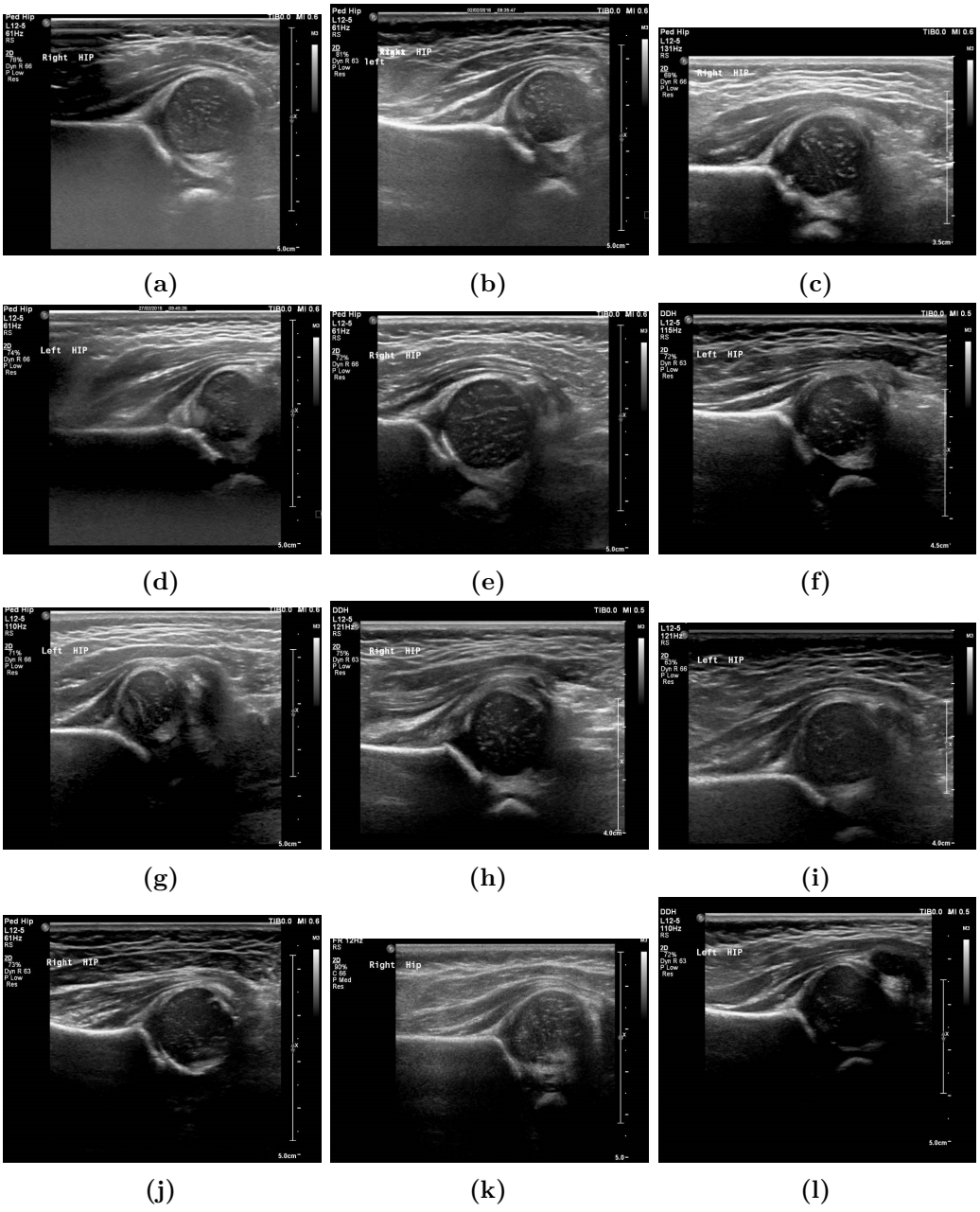


Figure A.3: These figures show the variation of images within the Pelvis dataset. The brightness of the images varies slightly, with a bright image shown in A.3a. Both male and female pelvises are featured in this dataset. In most images, a scale calibration ball is present toward the bottom of the image. How high the patient's hips are in the image varies from being near the top of the image, as seen in A.3g, to below the middle of the image, as seen in A.3k.



(a) **(b)** **(c)**

(d) **(e)** **(f)**

(g) **(h)** **(i)**

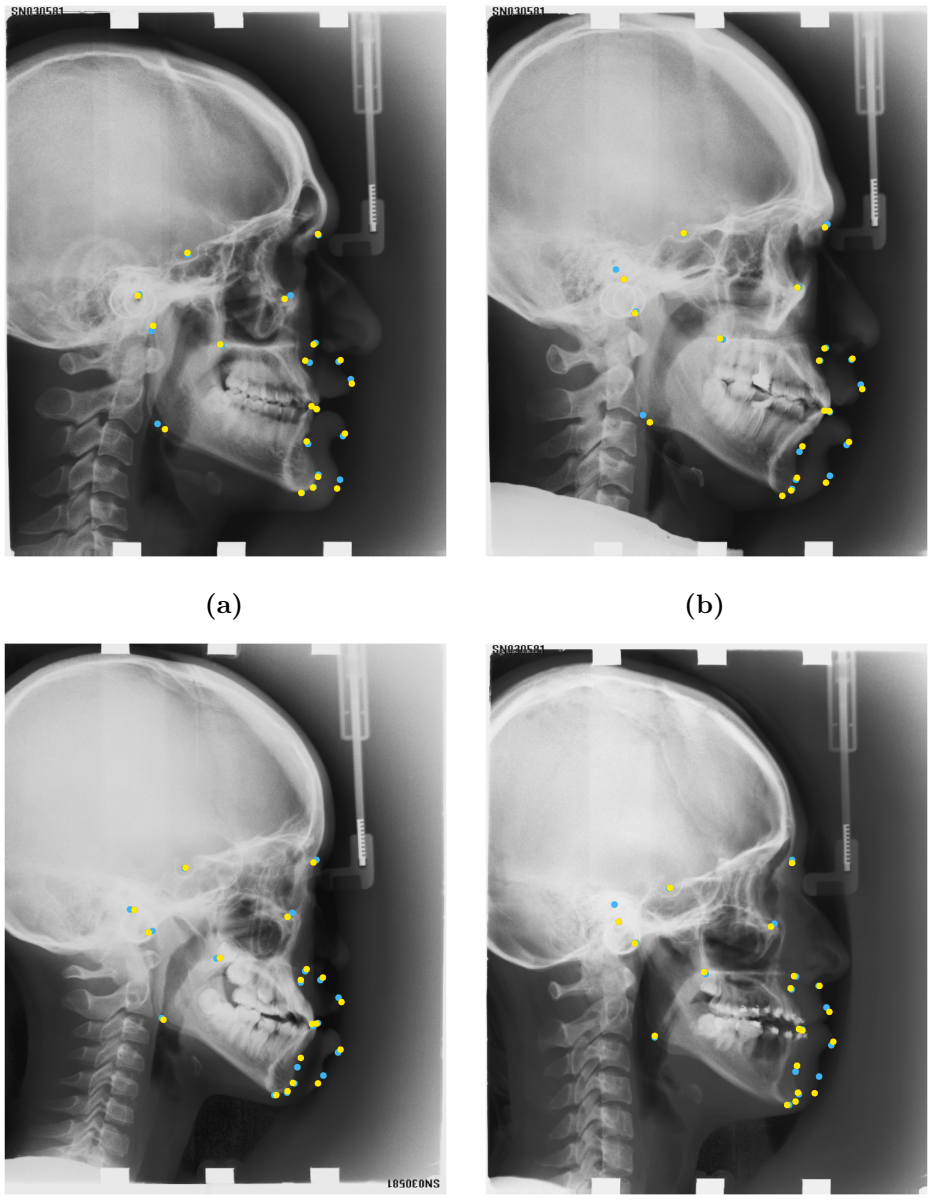
(j) **(k)** **(l)**

Figure A.4: These figures show the variation of images within the Infant Hip dataset. The brightness of the images varies slightly, which can be seen by comparing A.4a to A.4l. All images feature text in the top left corner and a colour gradient on the right side outside the ultrasound part of the image. Although the position of the hip in each image varies slightly, in the majority of images it is close to the centre.

B

Landmark Detection Examples

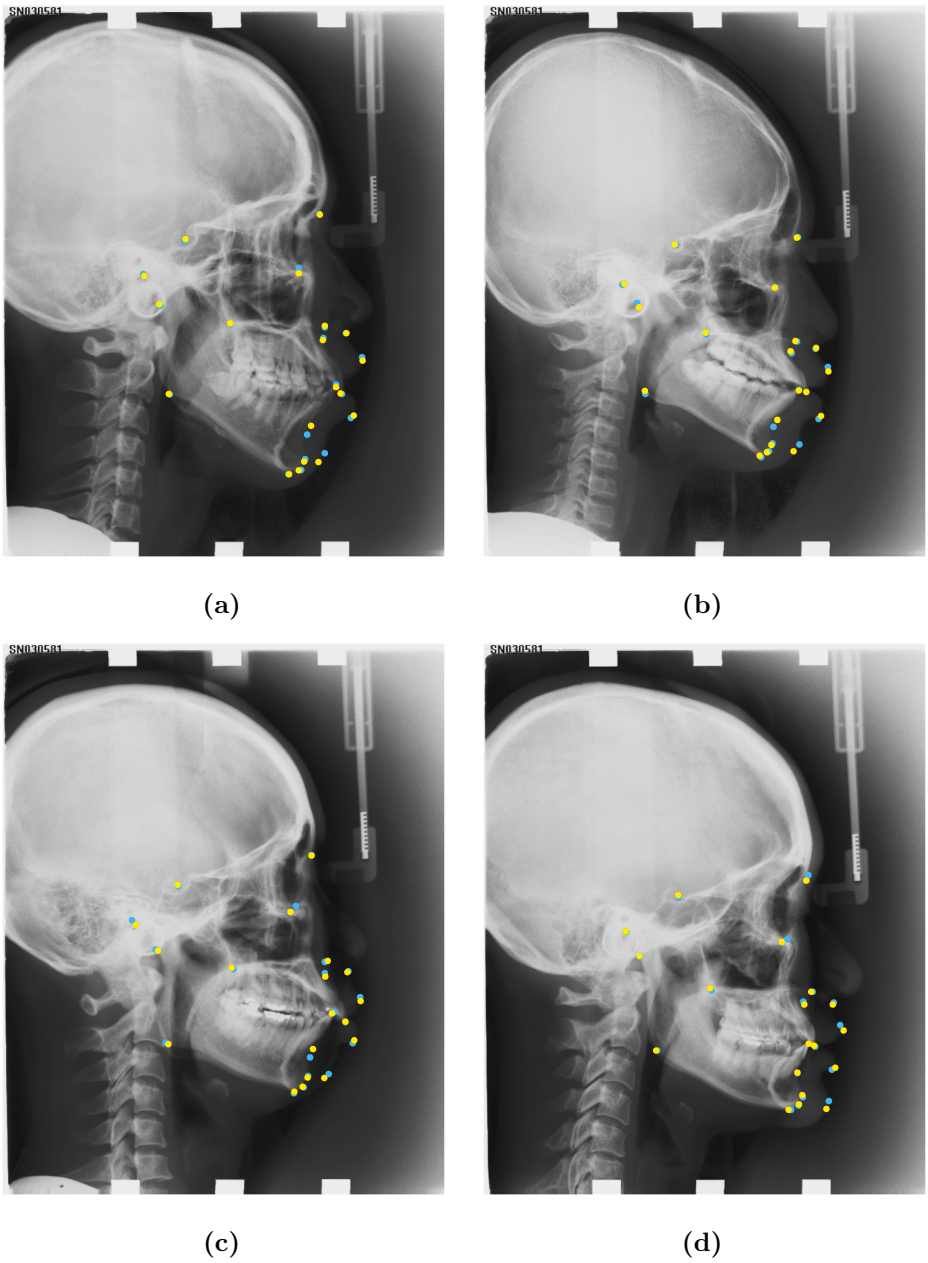
This appendix shows example landmark predictions on images from the Cephalometric, Hand, Pelvis and Infant Hip datasets made by the models described in Section 5.3, Section 5.4 and Section 5.5 . Each figure shows a random selection of images from each dataset's test set. These figures give a further sense of the accuracy of the models beyond the radial error statistics given in the aforementioned sections.



(a) **(b)**

(c) **(d)**

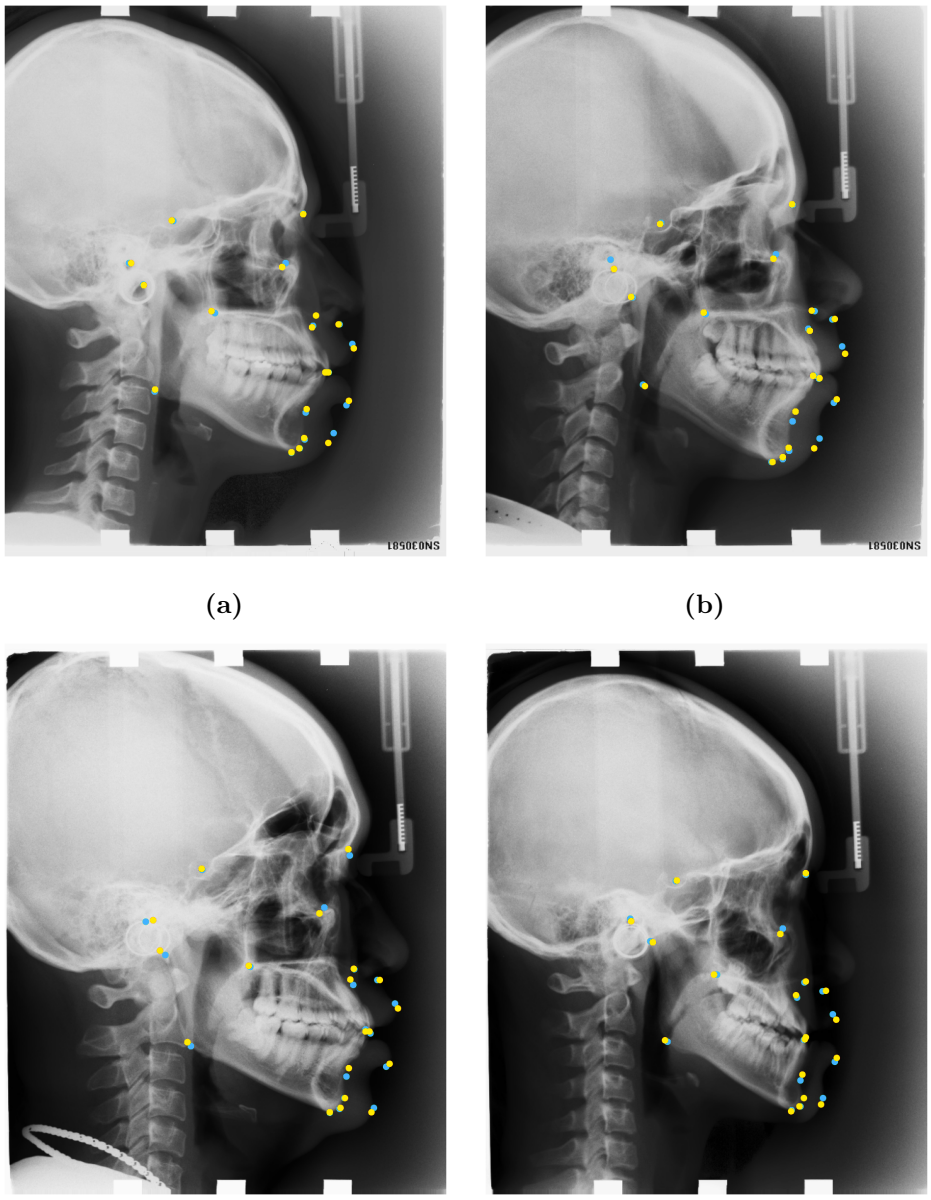
Figure B.1: This figure shows the landmark detection results for four images in Test Set 2 of the Cephalometric dataset made by the model described in Section 5.3.1. The yellow dots represent the predicted landmark positions and the blue dots represent the ground truth landmark positions.



(a) **(b)**

(c) **(d)**

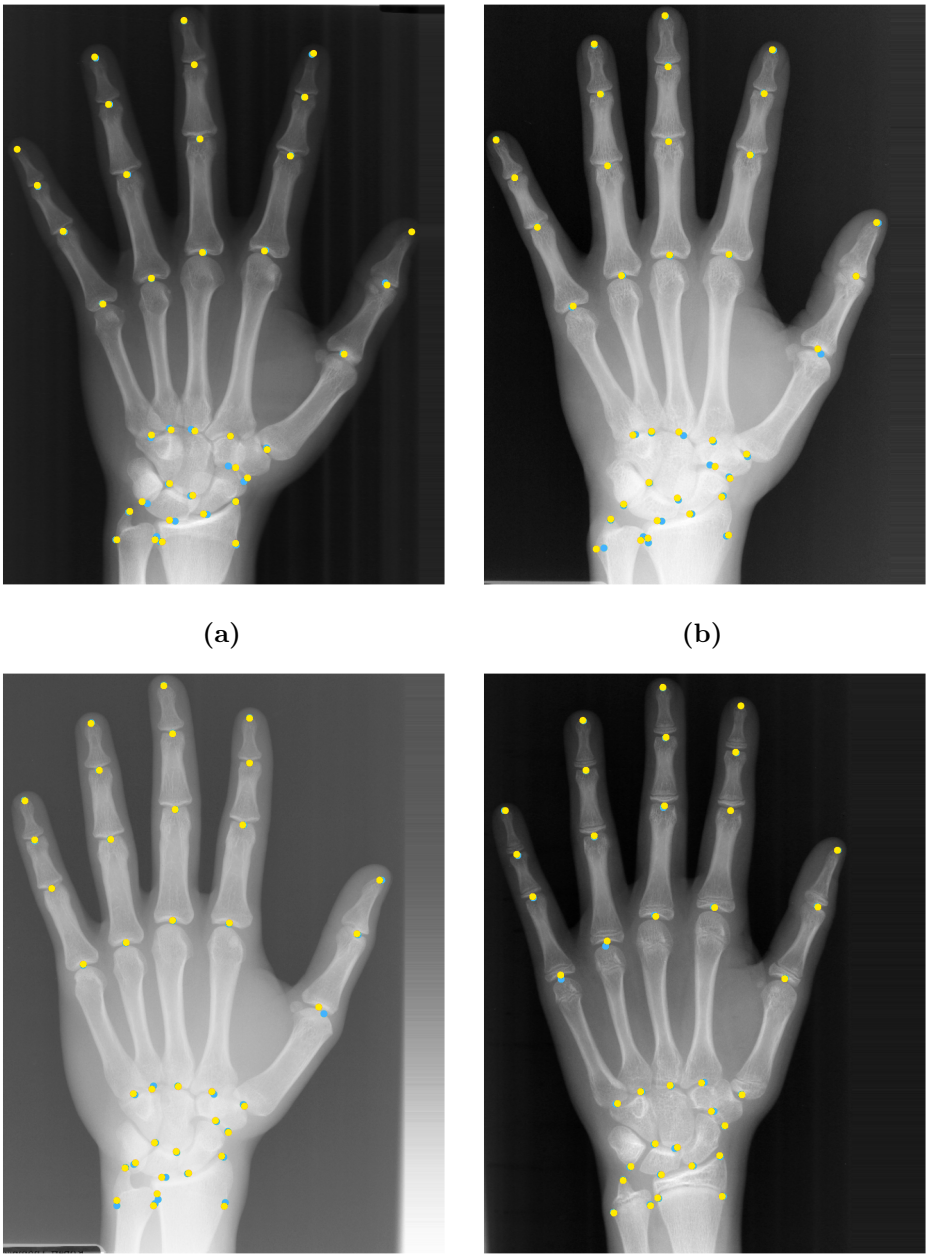
Figure B.2: This figure shows the landmark detection results for four images in Test Set 2 of the Cephalometric dataset made by the model described in Section 5.3.1. The yellow dots represent the predicted landmark positions and the blue dots represent the ground truth landmark positions.



(a) **(b)**

(c) **(d)**

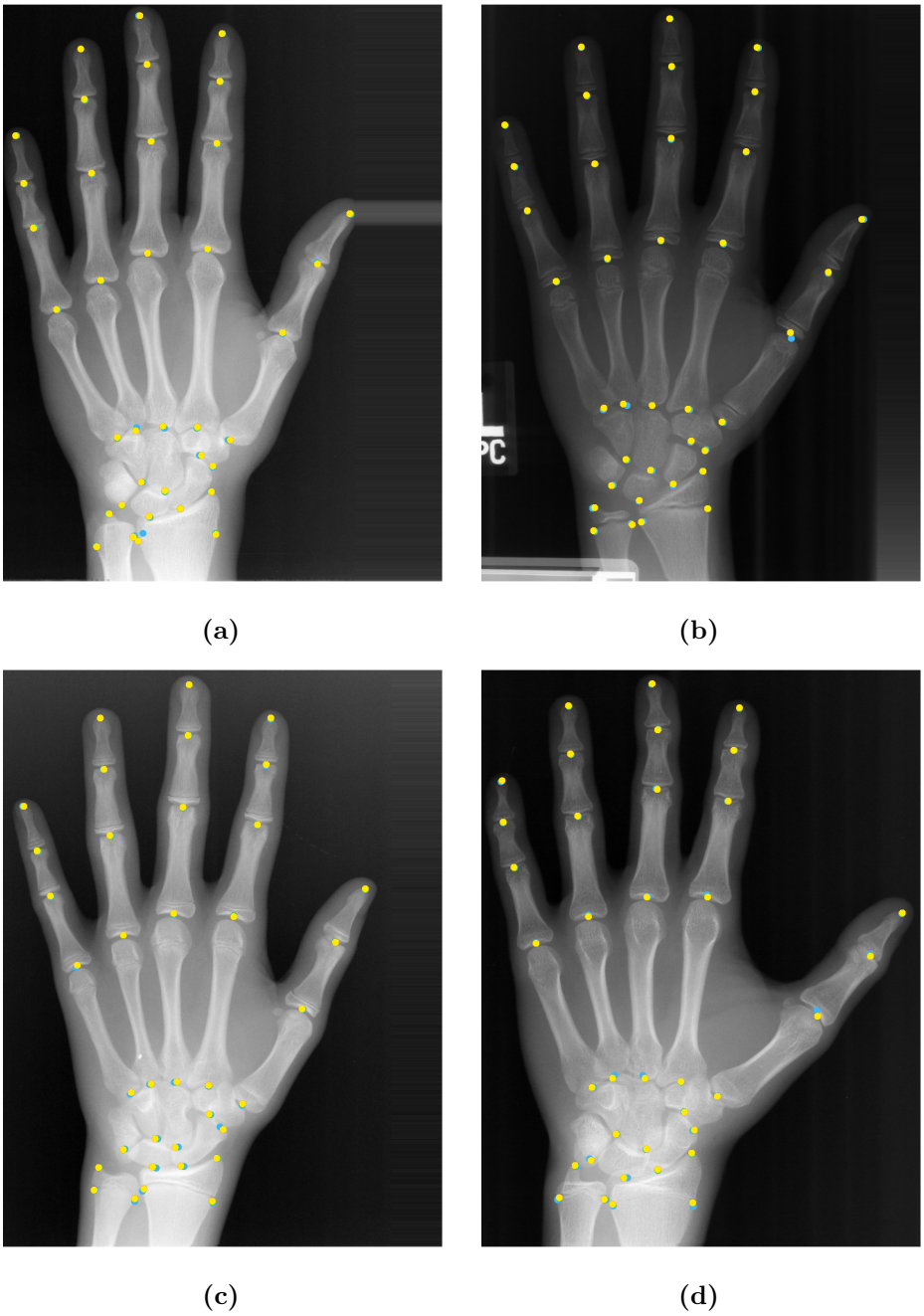
Figure B.3: This figure shows the landmark detection results for four images in Test Set 2 of the Cephalometric dataset made by the model described in Section 5.3.1. The yellow dots represent the predicted landmark positions and the blue dots represent the ground truth landmark positions.



(a) **(b)**

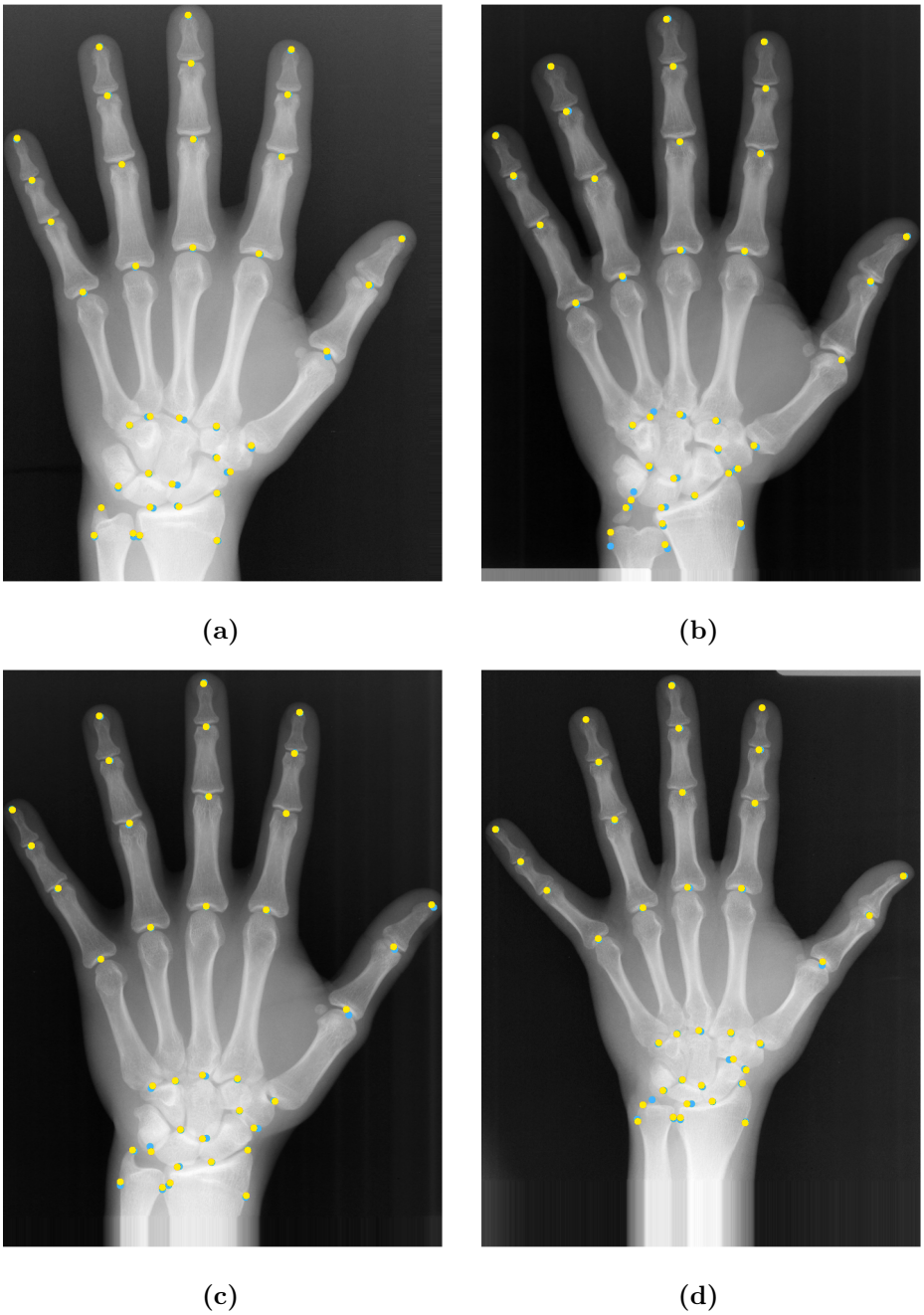
(c) **(d)**

Figure B.4: This figure shows the landmark detection results for four images in the test set of the Hand dataset made by the model described in Section 5.3.2. The yellow dots represent the predicted landmark positions and the blue dots represent the ground truth landmark positions. The landmarks are relatively accurately detected for images in this dataset and therefore the blue ground truth dots are often obscured by yellow predicted dots.



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

Figure B.5: This figure shows the landmark detection results for four images in the test set of the Hand dataset made by the model described in Section 5.3.2. The yellow dots represent the predicted landmark positions and the blue dots represent the ground truth landmark positions. The landmarks are relatively accurately detected for images in this dataset and therefore the blue ground truth dots are often obscured by yellow predicted dots.



(a) **(b)**

(c) **(d)**

Figure B.6: This figure shows the landmark detection results for four images in the test set of the Hand dataset made by the model described in Section 5.3.2. The yellow dots represent the predicted landmark positions and the blue dots represent the ground truth landmark positions. The landmarks are relatively accurately detected for images in this dataset and therefore the blue ground truth dots are often obscured by yellow predicted dots.

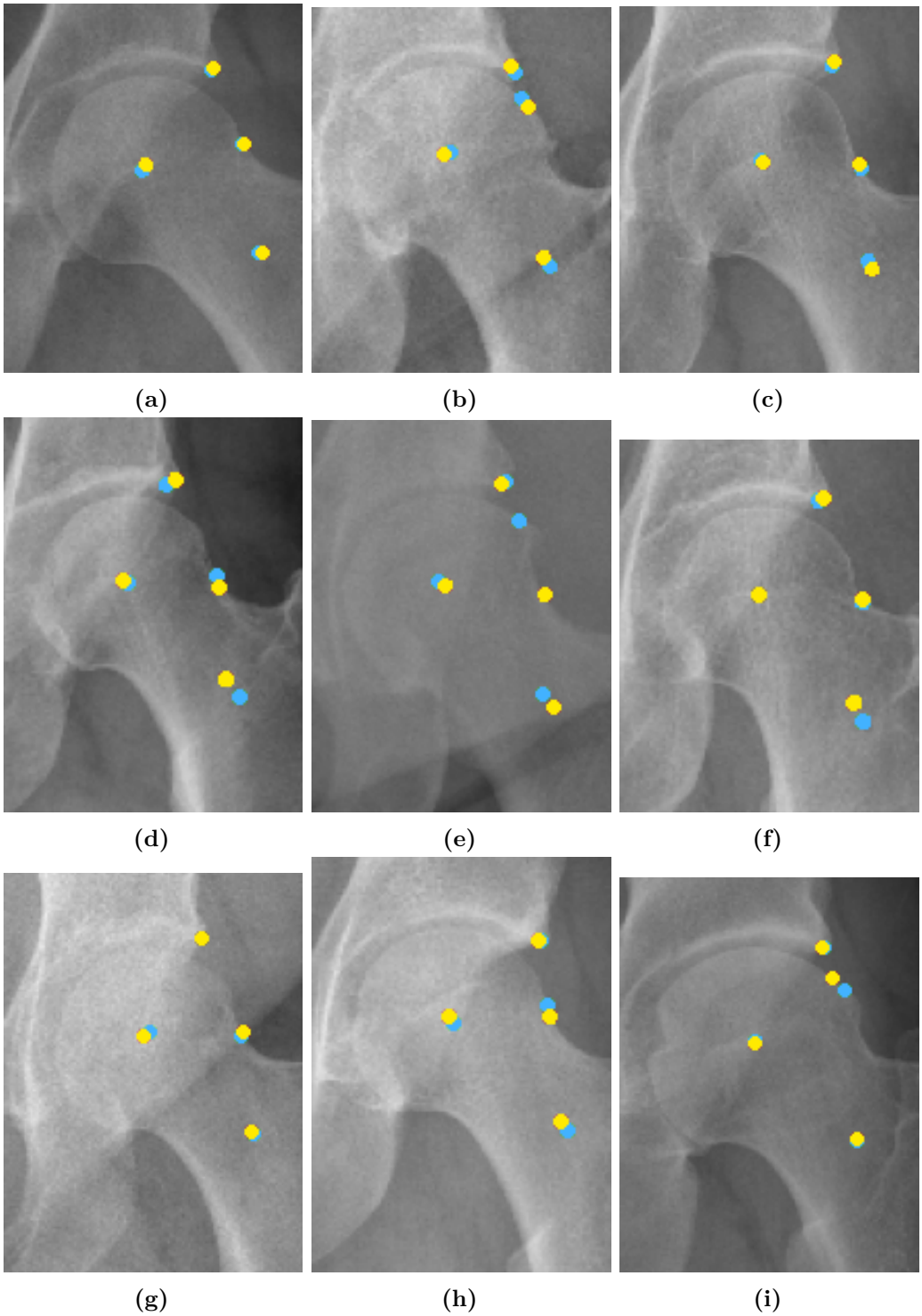


Figure B.7: This figure shows the landmark detection results for nine left hips cropped from images in the test set of the Pelvis dataset made by the model described in Section 5.4. The yellow dots represent the predicted landmark positions and the blue dots represent the ground truth landmark positions.

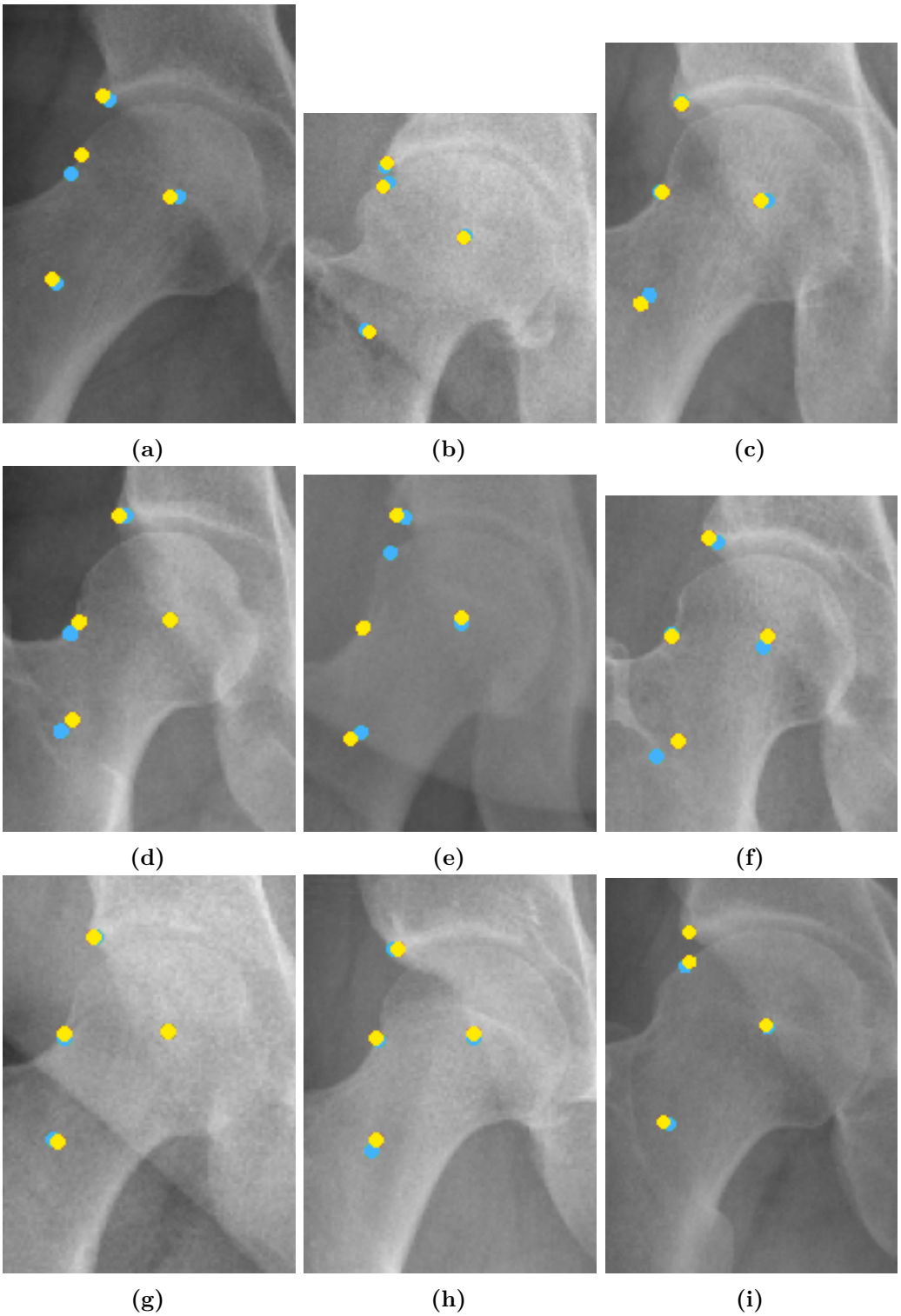
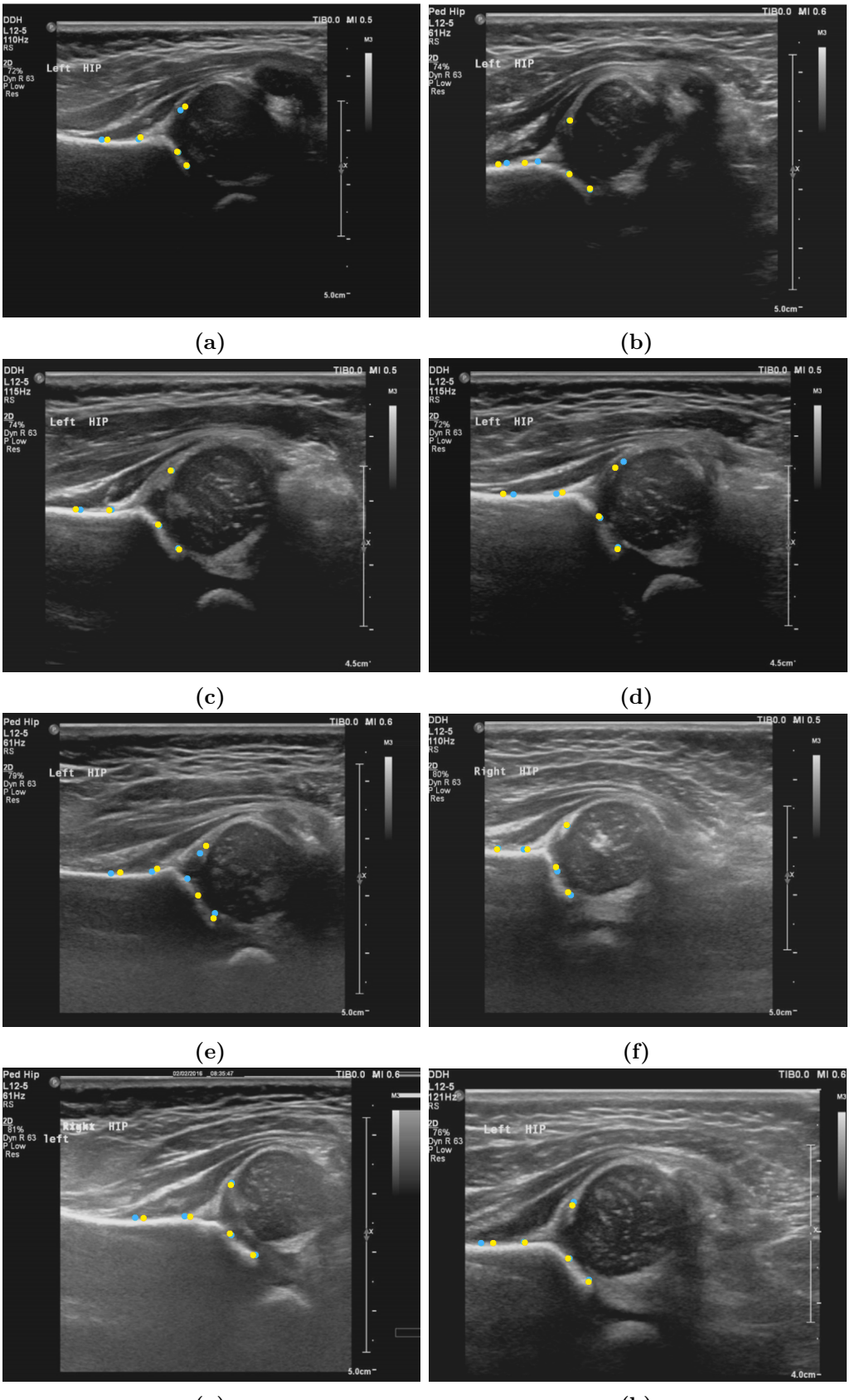


Figure B.8: This figure shows the landmark detection results for nine right hips cropped from images in the test set of the Pelvis dataset made by the model described in Section 5.4. The yellow dots represent the predicted landmark positions and the blue dots represent the ground truth landmark positions.



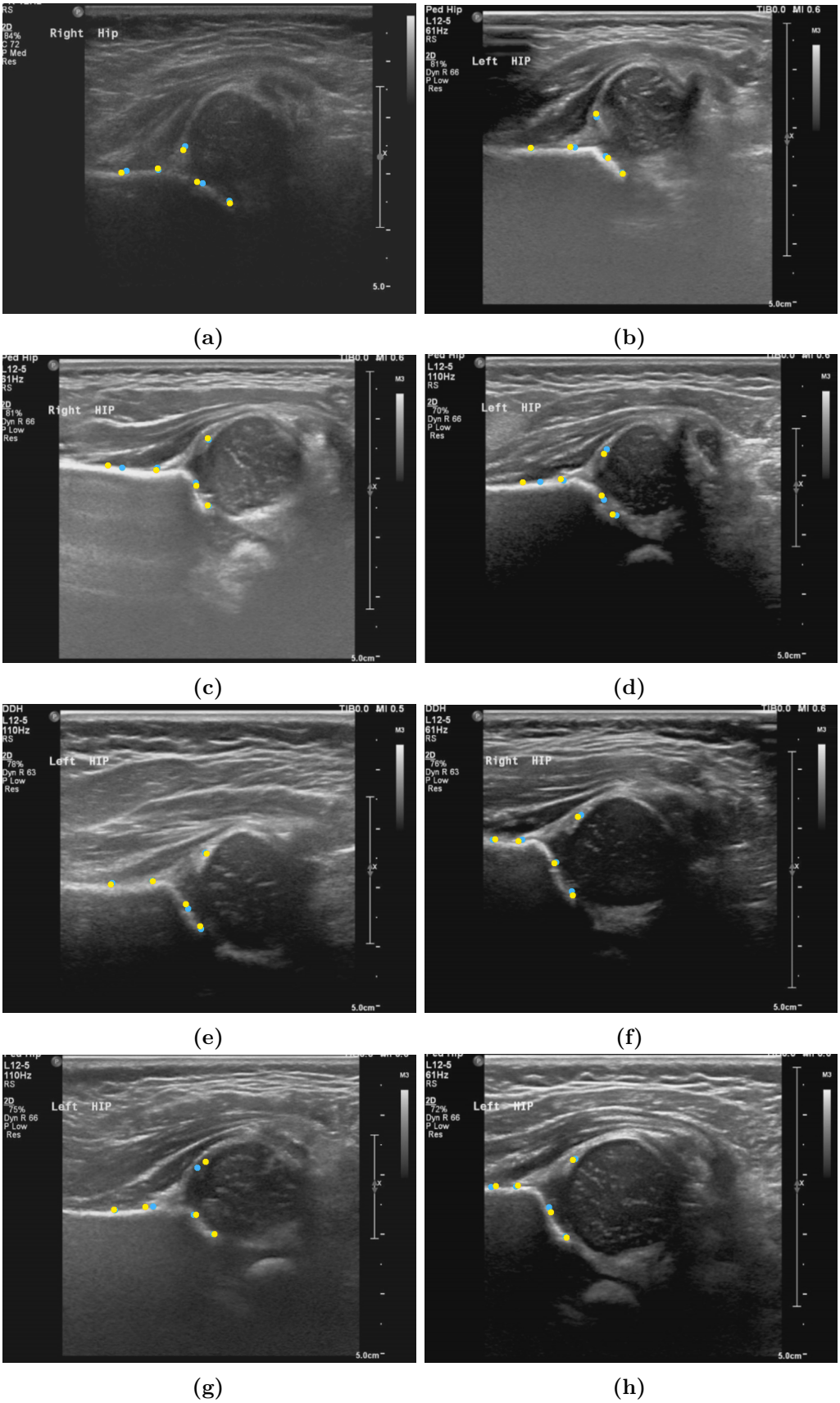
(a) **(b)**

(c) **(d)**

(e) **(f)**

(g) **(h)**

Figure B.9: This figure shows the landmark detection results for eight images in the test set of the Infant Hip dataset made by the model described in Section 5.5. The yellow dots represent the predicted landmark positions and the blue dots represent the ground truth landmark positions.



(a) **(b)** **(c)** **(d)** **(e)** **(f)** **(g)** **(h)**

Figure B.10: This figure shows the landmark detection results for eight images in the test set of the Infant Hip dataset made by the model described in Section 5.5. The yellow dots represent the predicted landmark positions and the blue dots represent the ground truth landmark positions.

C
Pelvis Angle Predictions

Example LCE-angle and α-angle predictions produced by the model described in Section 5.4 on X-ray pelvis images are given in this appendix. Each figure contains sub-figures displaying a croppred region of a patient's hip with both the ground truth angle and the model's predicted angle superimposed. These figures give a further sense of the accuracy of the model beyond the statistics given in Section 5.4.

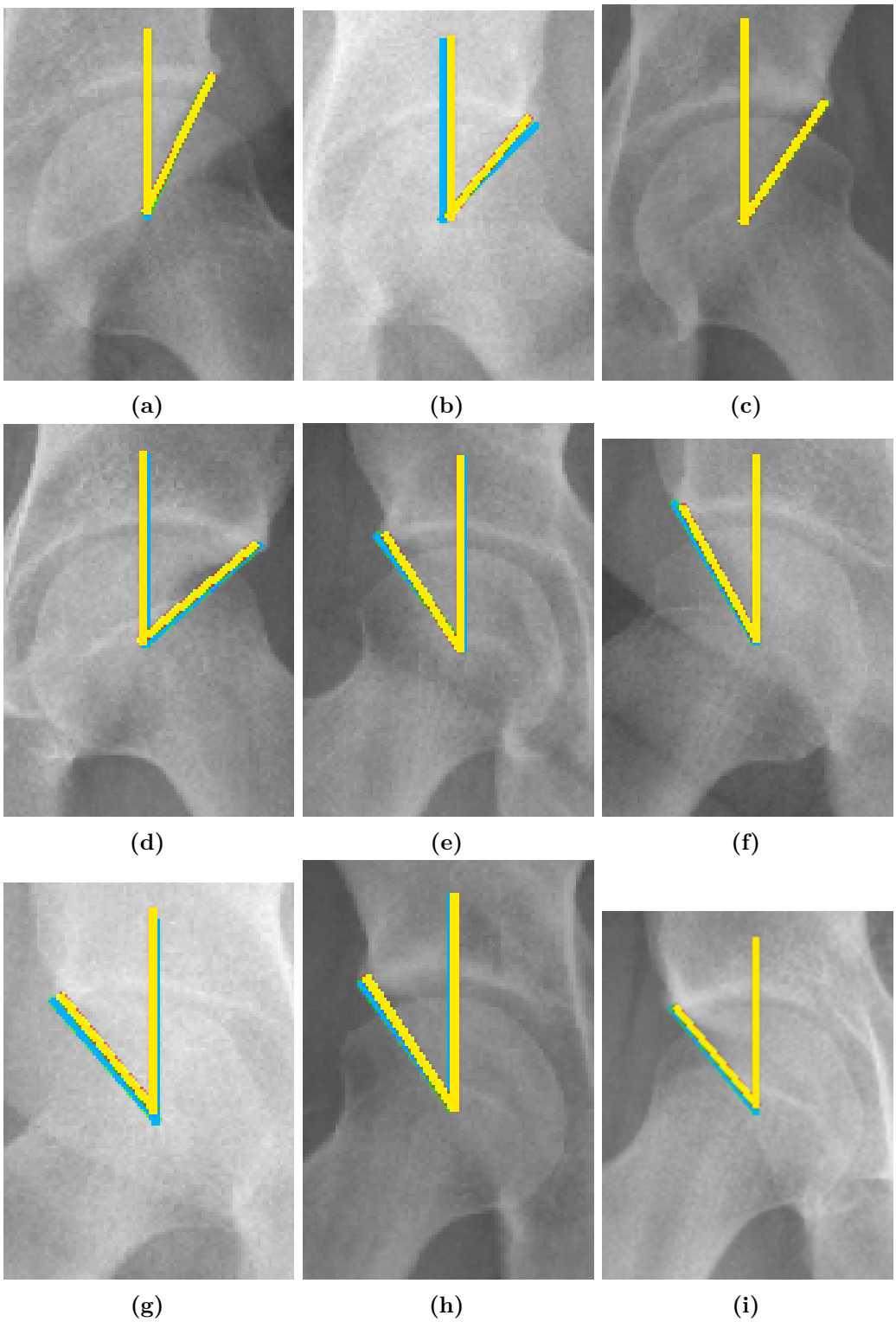


Figure C.1: This figure shows LCE-angle predictions made by the model described in 5.4 on nine X-ray pelvis images. The first four sub-figures capture right hips and the last five sub-figures capture left hips. The blue lines are the ground truth and the yellow lines are the model's predictions. All predicted LCE-angles are detected within five degrees.

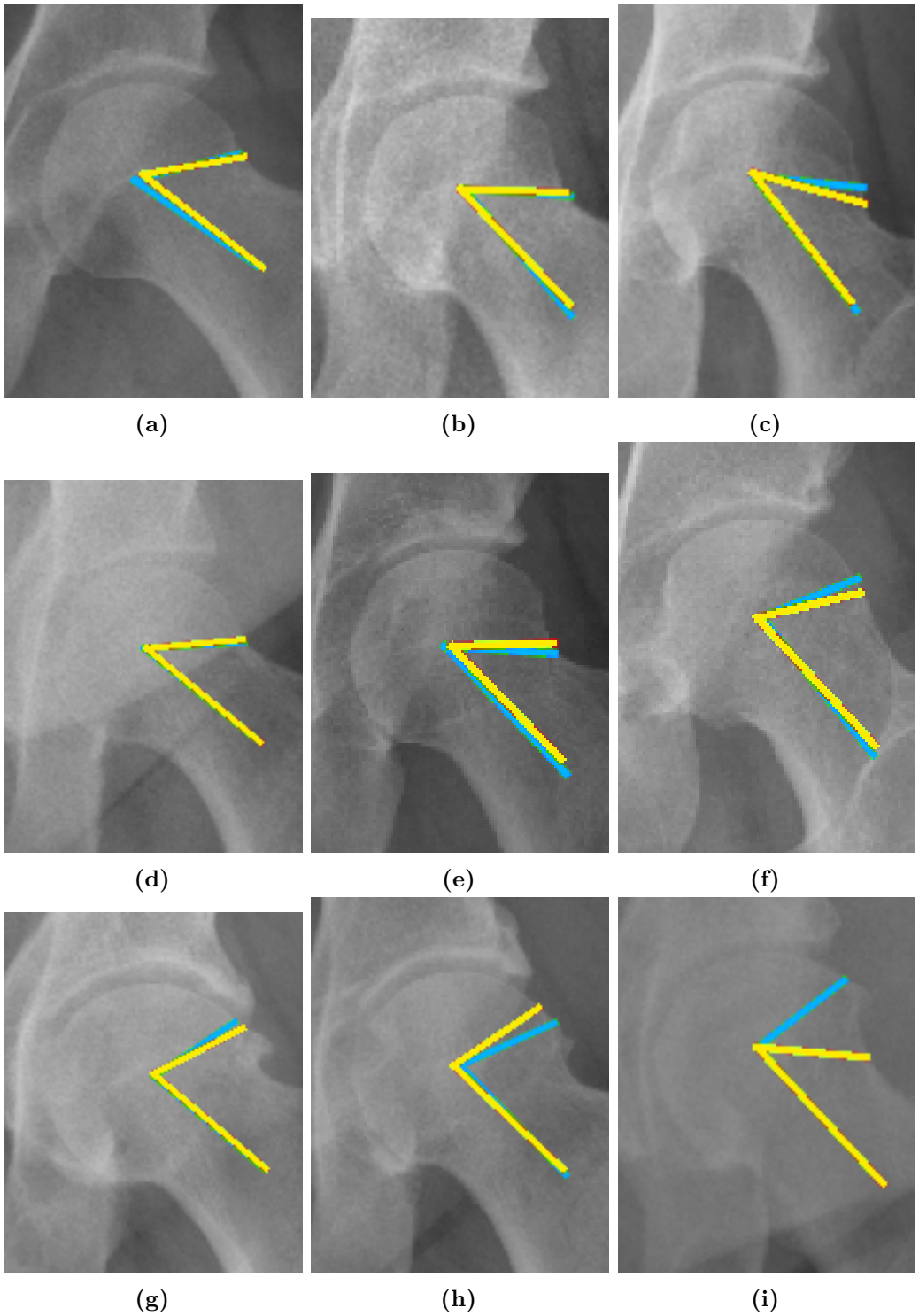


Figure C.2: This figure shows α-angle predictions made on the left hip of patients by the model described in 5.4 on nine X-ray pelvis images. The blue lines are the ground truth and the yellow lines are the model's predictions. Figure C.2a to Figure C.2e show accurate α-angle predictions for hips with α-angles below 65°. Figure C.2f to Figure C.2h show α-angle predictions within ten degrees of the ground truth for hips with α-angles above 65°. Figure C.2i shows an example where the model has not detected a small bump on the femoral head causing it to significantly underestimate the α-angle of this hip.

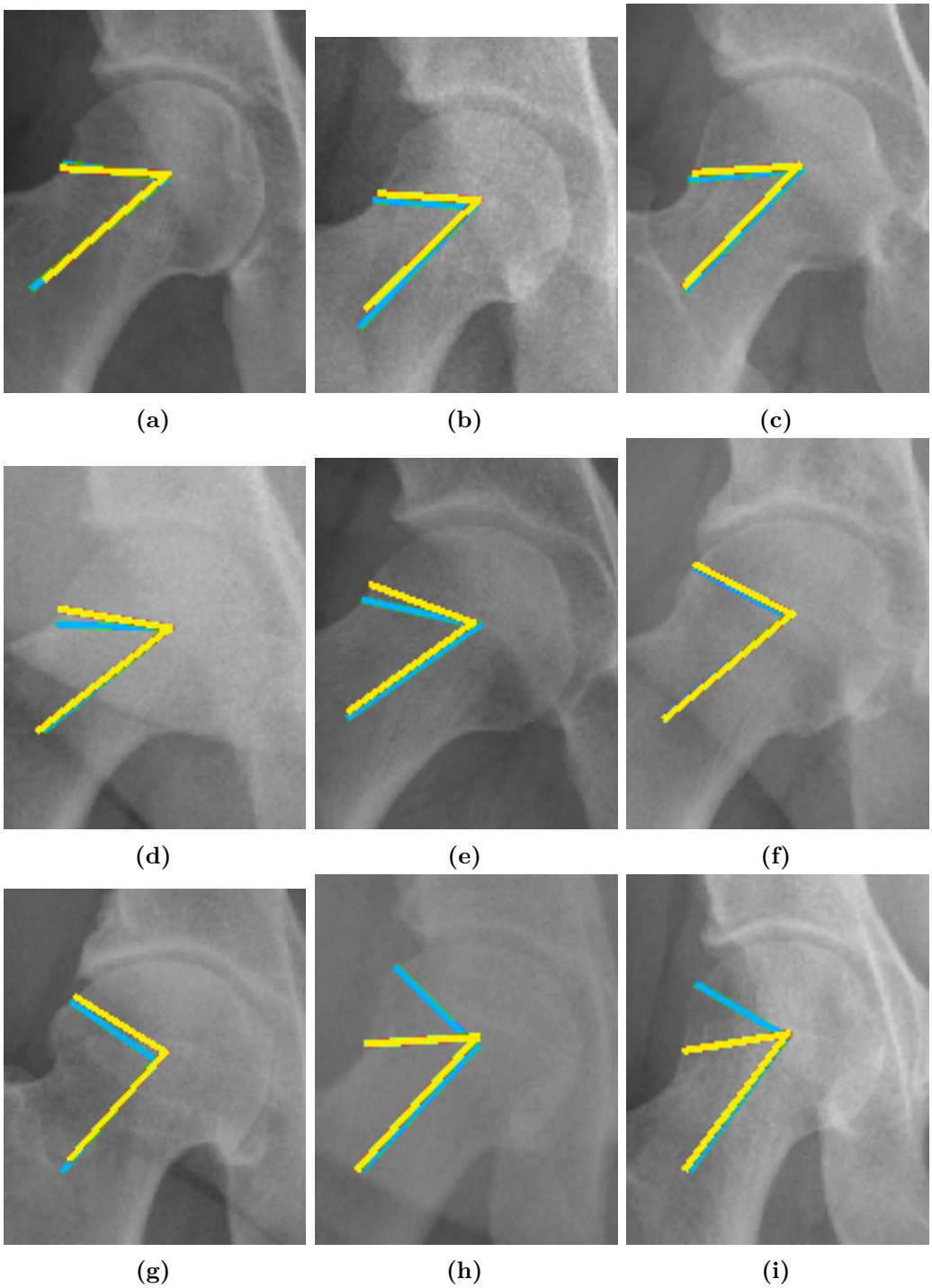
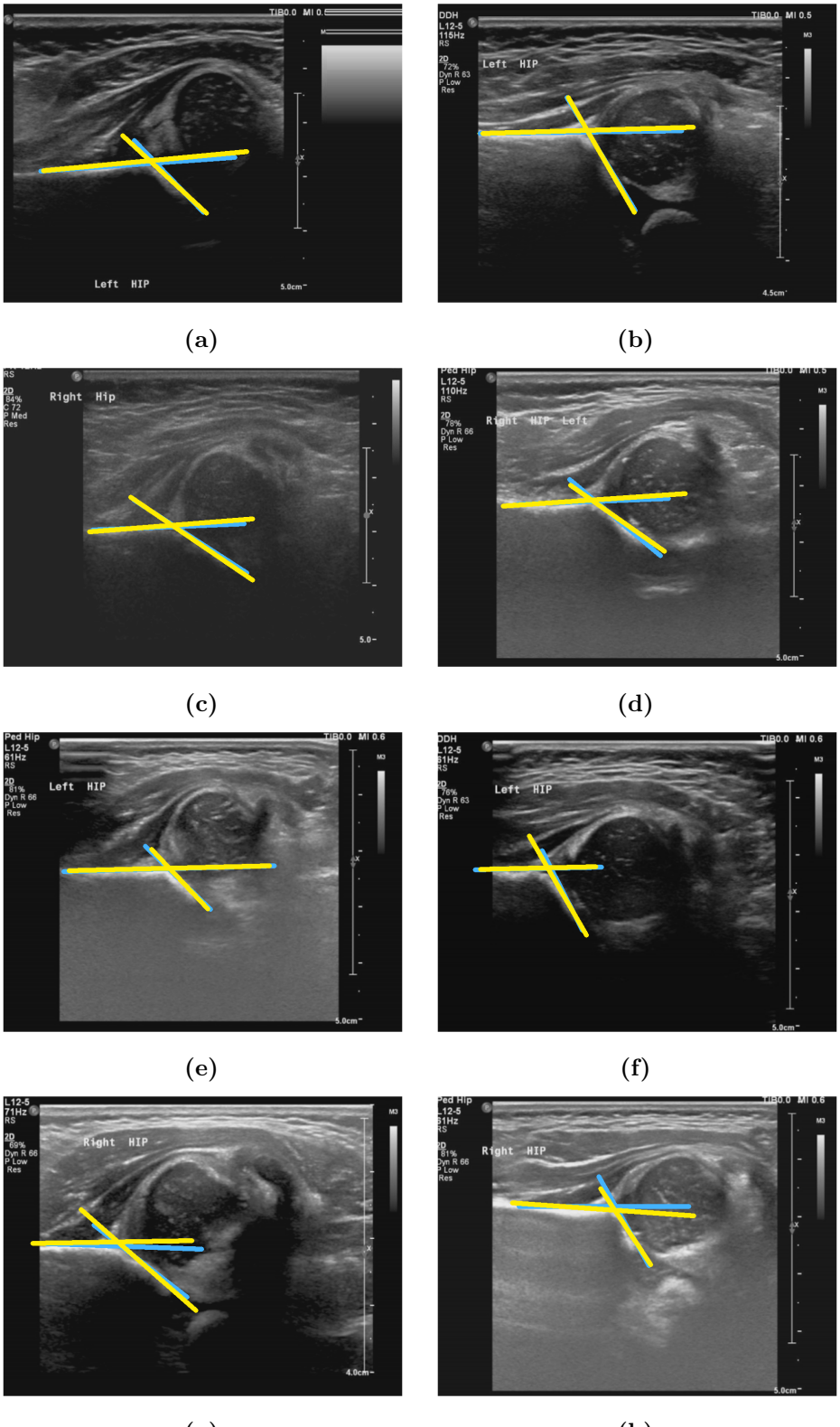


Figure C.3: This figure shows α-angle predictions made on the right hip of patients by the model described in 5.4 on nine X-ray pelvis images. The blue lines are the ground truth and the yellow lines are the model's predictions. Figure C.3a to Figure C.3e show accurate α-angle predictions for hips with α-angles below 65°. Figure C.2f and Figure C.2g show α-angle predictions within ten degrees of the ground truth for hips with α-angles above 65°. Figure C.2h and Figure C.2i show examples where the model has not detected a small bump on the femoral head causing it to significantly underestimate the α-angle of these hips.

D

Infant Hip Angle Predictions

Example α-angle and β-angle predictions produced by the model described in Section 5.5 on ultrasound images of infant hips are given in this appendix. Each figure contains sub-figures capturing an ultrasound image with both the ground truth angle and the model's predicted angle superimposed. These figures give a further sense of the accuracy of the model beyond the statistics given in Section 5.5.



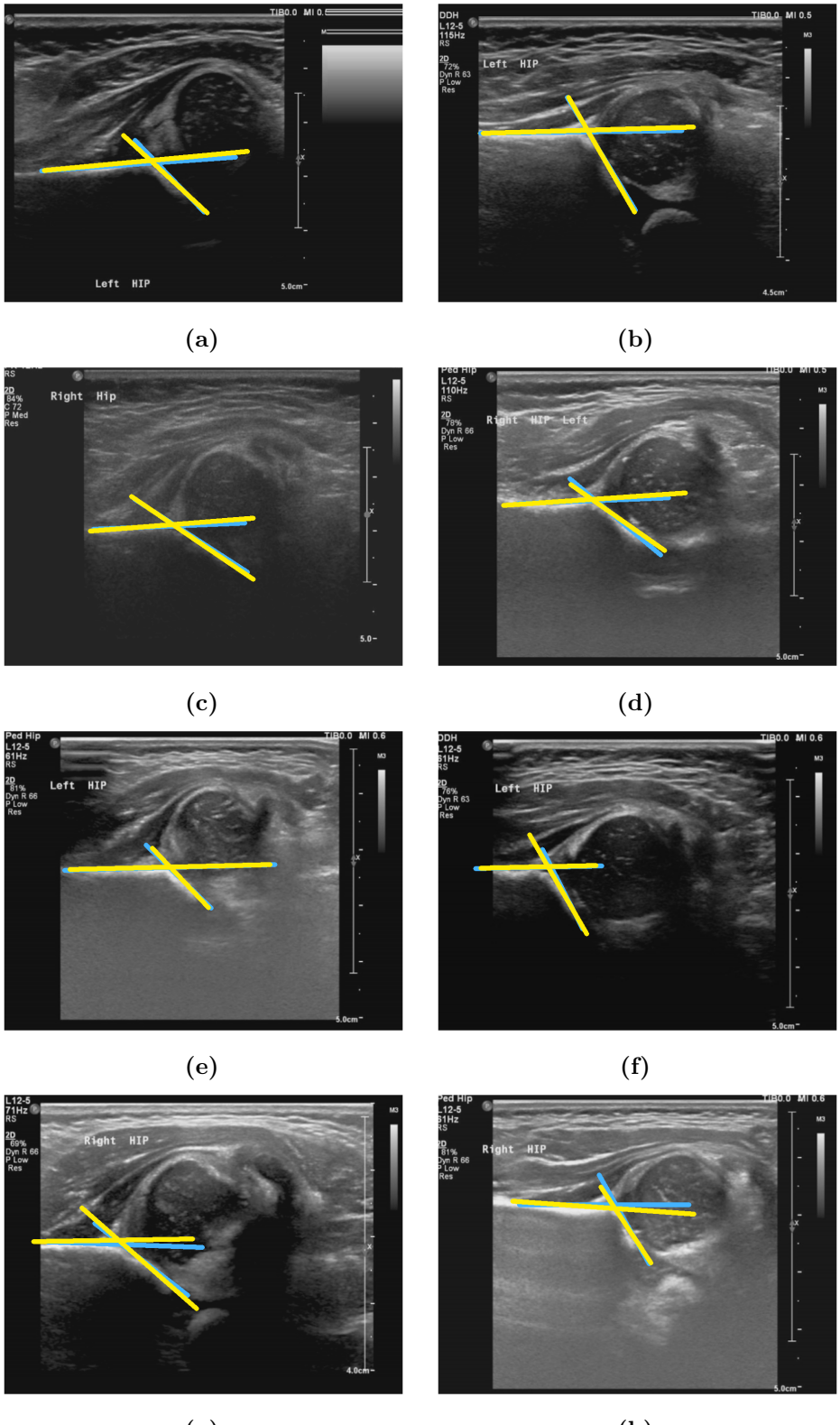
(a) (b)

(c) (d)

(e) (f)

(g) (h)

Figure D.1: This figure shows α-angle predictions made by the model described in 5.5 on eight Infant Hip ultrasound images. The blue lines are the ground truth and the yellow lines are the model's predictions. The first six sub-figures show predicted α-angles within five degrees of the ground truth. Figure D.1g and Figure D.1h show predicted α-angles within ten degrees of the ground truth.



(a) **(b)**

(c) **(d)**

(e) **(f)**

(g) **(h)**

Figure D.2: This figure shows β-angle predictions made by the model described in 5.5 on eight Infant Hip ultrasound images. The blue lines are the ground truth and the yellow lines are the model's predictions. The first six sub-figures show predicted β-angles within five degrees of the ground truth. Figure D.1g and Figure D.1h show predicted β-angles within ten degrees of the ground truth.

E
Experimental Code

The code for all experiments presented in this thesis is publicily available at the DOI resource location - `https://zenodo.org/doi/10.5281/zenodo.10822726`. The python packages and their versions used to run the code in the repository are:

- absl-py==0.11.0
- cachetools==4.1.1
- certifi==2021.5.30
- chardet==3.0.4
- click==8.0.4
- cloudpickle==2.0.0
- cycler==0.10.0
- cytoolz==0.11.0
- dask==2021.3.0
- dataclasses==0.8
- decorator==5.1.0
- docker-pycreds==0.4.0
- efficientnet-pytorch==0.6.3
- future==0.18.2
- gitdb==4.0.9

- gitpython==3.1.18
- google-auth-oauthlib==0.4.2
- google-auth==1.23.0
- grpcio==1.33.2
- idna==2.10
- imagecodecs==2020.5.30
- imageio==2.9.0
- imgaug==0.4.0
- importlib-metadata==2.0.0
- joblib==1.0.1
- kiwisolver==1.3.1
- markdown==3.3.3
- matplotlib==3.3.4
- mkl-fft==1.2.0
- mkl-random==1.1.1
- mkl-service==2.3.0
- monai==0.3.0+68.gde65c48
- munch==2.5.0
- networkx==2.5
- nibabel==3.2.0
- numpy==1.19.2
- oauthlib==3.1.0
- olefile==0.46
- opencv-python==4.5.3.56
- packaging==20.4
- pandas==1.1.0
- pathtools==0.1.2
- pillow==8.0.1
- pip==20.2.2
- pretrainedmodels==0.7.4

- promise==2.3
- protobuf==3.13.0
- psutil==5.9.4
- pyasn1-modules==0.2.8
- pyasn1==0.4.8
- pyparsing==2.4.7
- pypng==0.0.20
- python-dateutil==2.8.2
- pytorch-ignite==0.4.2
- pytz==2020.1
- pywavelets==1.1.1
- pyyaml==5.4.1
- requests-oauthlib==1.3.0
- requests==2.25.0
- rsa==4.6
- scikit-image==0.17.2
- scikit-learn==0.24.2
- scipy==1.5.2
- segmentation-models-pytorch==0.2.1
- sentry-sdk==1.12.1
- setproctitle==1.2.3
- setuptools==49.6.0.post20200814
- shapely==1.7.1
- shortuuid==1.0.11
- six==1.15.0
- smmap==5.0.0
- tensorboard-plugin-wit==1.7.0
- tensorboard==2.4.0
- tensorboardx==2.1
- threadpoolctl==2.2.0

- tifffile==2021.3.17
- timm==0.4.12
- toolz==0.11.1
- torch==1.10.2
- torchaudio==0.10.2
- torchsummary==1.5.1
- torchvision==0.11.3
- tornado==6.1
- tqdm==4.48.2
- typing-extensions==3.10.0.2
- urllib3==1.26.13
- wandb==0.13.7
- werkzeug==1.0.1
- wheel==0.34.2
- yacs==0.1.8
- zipp==3.4.0

CUDA and cuDNN versions are:

- CUDA/11.1.1-GCC-10.2.0
- cuDNN/8.0.4.30-CUDA-11.1.1

F

Failed Approaches

Whilst I have discussed the weaknesses and limitations of the methods presented throughout my thesis I have also omitted some methods I experimented with during my DPhil which did not obtain 'good' results. I did this because I did not gather sufficiently robust or rigorous data using these methods to report once it became clear the method would not produce accurate results. This appendix serves as a brief description of these methods so that other researchers do not also spend time experimenting with them.

- **Deterministic Segmentation Methods** - As mentioned in Section 1.1, the first project I undertook in my DPhil was to automatically diagnose Femoroacetabular Impingement (FAI) from X-ray images of the pelvis. I initially experimented with deterministic algorithms for segmenting the femoral head in X-rays, as an intermediate task, from which to calculate the LCE-angle and α-angle. I experimented with a variety of deterministic segmentation algorithms implemented by the **`scikit-image`** library. These included Canny-Edge detection, Flood-fill and Watershed, as well as other more custom deterministic algorithms.[1] However, the femoral head does not always have a clear visual boundary in the X-ray and, in addition, different X-rays have

[1] `https://scikit-image.org/docs/stable/api/skimage.segmentation.html`

different lighting characteristics. This meant that no deterministic method experimented with could segment the femoral head consistently across a test set of images, with many of the output segmentations having very large errors.

- **Computing Angle Measurements from Segmentations** - After deterministic segmentation algorithms failed I trained a U-Net architecture to perform segmentation. To do this I manually produced ground truth femoral head segmentation masks for the Pelvis dataset. The segmentation masks produced by the trained model were more accurate than those produced by the deterministic algorithms however it proved very difficult to produce an algorithm to compute the LCE-angle and α-angle from them. I attempted to obtain a circle around the femoral head by applying the Hough transform [10] to the predicted segmentations. However, both the segmentation and the circle had to be pixel-perfect to accurately calculate the α-angle and this was often not the case. I never managed to develop a deterministic algorithm which could calculate the LCE-angle and α-angle from segmentation masks consistently across a test set.

- **Regressing Landmarks from Images** - After the segmentation methods failed I concluded that I needed a method which predicted landmarks from images without an intermediate segmentation. I first tried to train a CNN to regress the landmark coordinates directly from images. However, it was not possible to train a model to perform this task with any accuracy. I believe that the task of predicting two numbers (x and y coordinates) from an image is too broad for a model to learn. The model has no medical knowledge and does not know that the numbers it is training to regress are coordinates on the image. It therefore cannot infer a relationship between these numbers and the pixels of the images. After this approach failed I performed a literature review and began training models on heatmaps centred at ground truth landmark positions as proposed by Tompson *et al.* [129]. This is the method I used for the experiments presented throughout my thesis.

- **Self-Supervised Learning** - During my DPhil, I experimented with methods to pre-train the weights of a CNN to improve results on landmark detection tasks. I tried self-supervised learning where the model is first pre-trained on a task which is constructed from unlabelled images (called the 'pretext' task), before being trained to perform landmark detection. In particular, I pre-trained a model to place patches of an image into the correct positions, known as the jigsaw pretext task [98]. Although pre-training the model in this way did produce better results than when weights were initialised randomly, it consistently produced less accurate landmark detection results compared to when pre-trained weights learned from ImageNet were used, as detailed in Section 5.2.4. Using pre-trained ImageNet weights is seen as the baseline method in the area and I therefore omitted the self-supervision experiment from my thesis.

References

[1] Alexander Aarvold, Daniel Perry, Jason Mavrotas, et al. “The Management of Developmental Dysplasia of the Hip in Children Aged Under Three Months: A Consensus Study from the British Society for Children’s Orthopaedic Surgery”. In: *The Bone & Joint Journal* 105.2 (2023), pp. 209–214.

[2] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, et al. “A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges”. In: *Information Fusion* 76 (2021), pp. 243–297.

[3] Rintje Agricola, Jan Waarsing, Nigel Arden, et al. “Cam Impingement of the Hip—A Risk Factor for Hip Osteoarthritis”. In: *Nature Reviews Rheumatology* 9.10 (2013), p. 630.

[4] Rintje Agricola, Jan Waarsing, Geraint Thomas, et al. “Cam Impingement: Defining the Presence of a Cam Deformity by the Alpha Angle: Data from the CHECK Cohort and Chingford Cohort”. In: *Osteoarthritis and Cartilage* 22.2 (2014), pp. 218–225.

[5] Jafar Alzubi, Balasubramaniyan Bharathikannan, Sudeep Tanwar, et al. “Boosted Neural Network Ensemble Classification for Lung Cancer Disease Diagnosis”. In: *Applied Soft Computing* 80 (2019), pp. 579–591.

[6] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, et al. “2D Human Pose Estimation: New Benchmark and State of the Art Analysis”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 3686–3693.

[7] Diego Ardila, Atilla Kiraly, Sujeeth Bharadwaj, et al. “End-to-End Lung Cancer Screening with Three-Dimensional Deep Learning on Low-Dose Chest Computed Tomography”. In: *Nature Medicine* 25.6 (2019), pp. 954–961.

[8] Sercan Ö Arik, Bulat Ibragimov, and Lei Xing. “Fully Automated Quantitative Cephalometry Using Convolutional Neural Networks”. In: *Journal of Medical Imaging* 4.1 (2017), p. 014501.

[9] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. “SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.12 (2017), pp. 2481–2495.

[10] Dana Ballard. “Generalizing the Hough Transform to Detect Arbitrary Shapes”. In: *Pattern Recognition* 13.2 (1981), pp. 111–122.

[11] Purnajyoti Banerjee and Christopher Mclean. “Femoroacetabular Impingement: A Review of Diagnosis and Management”. In: *Current Reviews in Musculoskeletal Medicine* 4 (2011), pp. 23–32.

[12] Marcel Beetz, Jorge Acero, Abhirup Banerjee, et al. “Interpretable Cardiac Anatomy Modeling Using Variational Mesh Autoencoders”. In: *Frontiers in Cardiovascular Medicine* 9 (2022), p. 983868.

[13] Peter Belhumeur, David Jacobs, David Kriegman, et al. “Localizing Parts of Faces Using a Consensus of Exemplars”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.12 (2013), pp. 2930–2940.

[14] Bastian Bier, Mathias Unberath, Jan-Nico Zaech, et al. “X-ray-Transform Invariant Anatomical Landmark Detection for Pelvic Trauma Surgery”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 55–63.

[15] Riccardo Boscolo, Matthew Brown, and Michael McNitt-Gray. “Medical Image Segmentation with Knowledge-Guided Robust Active Contours”. In: *Radiographics* 22.2 (2002), pp. 437–448.

[16] Ian Campbell. “Chi-squared and Fisher–Irwin tests of two-by-two tables with small sample recommendations”. In: *Statistics in medicine* 26.19 (2007), pp. 3661–3675.

[17] Chongyan Chen and Guoyan Zheng. “Fully-Automatic Landmark Detection in Cephalometric X-ray Images by Data-Driven Image Displacement Estimation”. In: *Proc. ISBI International Symposium on Biomedical Imaging*. 2014.

[18] Lisha Chen, Hui Su, and Qiang Ji. “Face Alignment with Kernel Density Deep Neural Network”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 6992–7002.

[19] Runnan Chen, Yuexin Ma, Nenglun Chen, et al. “Cephalometric Landmark Detection by Attentive Feature Pyramid Fusion and Regression-Voting”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 873–881.

[20] Runnan Chen, Yuexin Ma, Lingjie Liu, et al. “Semi-Supervised Anatomical Landmark Detection via Shape-Regulated Self-Training”. In: *Neurocomputing* 471 (2022), pp. 335–345.

[21] Tao Chen, Yuxiao Zhang, Bo Wang, et al. “Development of a Fully Automated Graf Standard Plane and Angle Evaluation Method for Infant Hip Ultrasound Scans”. In: *Diagnostics* 12.6 (2022), p. 1423.

[22] Yilun Chen, Zhicheng Wang, Yuxiang Peng, et al. “Cascaded Pyramid Network for Multi-person Pose Estimation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 7103–7112.

[23] Chengwen Chu, Cheng Chen, Lutz Nolte, et al. “Fully Automatic Cephalometric X-ray Landmark Detection Using Random Forest Regression and Sparse Shape Composition”. In: *submitted to Automatic Cephalometric X-ray Landmark Detection Challenge* (2014).

[24] William G Cochran. “The $\chi 2$ test of goodness of fit”. In: *The Annals of mathematical statistics* (1952), pp. 315–345.

[25] Timothy Cootes, Andrew Hill, Christopher Taylor, et al. “Use of Active Shape Models for Locating Structures in Medical Images”. In: *Image and Vision Computing* 12.6 (1994), pp. 355–365.

[26] Timothy Cootes, Christopher Taylor, David Cooper, et al. “Active Shape Models-Their Training and Application”. In: *Computer Vision and Image Understanding* 61.1 (1995), pp. 38–59.

[27] Leda Cosmides and John Tooby. “Are Humans Good Intuitive Statisticians After All? Rethinking Some Conclusions from the Literature on Judgment Under Uncertainty”. In: *Cognition* 58.1 (1996), pp. 1–73.

[28] David Cristinacce and Tim Cootes. “Automatic Feature Localisation with Constrained Local Models”. In: *Pattern Recognition* 41.10 (2008), pp. 3054–3067.

[29] Xiubin Dai, Hao Zhao, Tianliang Liu, et al. “Locating Anatomical Landmarks on 2D Lateral Cephalograms Through Adversarial Encoder-Decoder Networks”. In: *IEEE Access* 7 (2019), pp. 132738–132747.

[30] Navneet Dalal and Bill Triggs. “Histograms of Oriented Gradients for Human Detection”. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 1. IEEE. 2005, pp. 886–893.

[31] Jia Deng, Wei Dong, Richard Socher, et al. “ImageNet: A Large-Scale Hierarchical Image Database”. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Ieee. 2009, pp. 248–255.

[32] Carol Dezateux and Karen Rosendahl. “Developmental Dysplasia of the Hip”. In: *The Lancet* 369.9572 (2007), pp. 1541–1552.

[33] Kunio Doi. “Computer-Aided Diagnosis in Medical Imaging: Historical Review, Current Status and Future Potential”. In: *Computerized Medical Imaging and Graphics* 31.4-5 (2007), pp. 198–211.

[34] Xuanyi Dong and Yi Yang. “Teacher Supervises Students How to Learn from Partially Labeled Images for Facial Landmark Detection”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 783–792.

[35] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *arXiv preprint arXiv:2010.11929* (2020).

[36] Thomas Ebner, Darko Stern, Rene Donner, et al. “Towards Automatic Bone Age Estimation from MRI: Localization of 3D Anatomical Landmarks”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2014, pp. 421–428.

[37] David Felson. “Epidemiology of Hip and Knee Osteoarthritis”. In: *Epidemiologic Reviews* 10 (1988), pp. 1–28.

[38] Ronald Fisher. “Statistical Methods for Research Workers”. In: *Breakthroughs in Statistics*. Springer, 1992, pp. 66–70.

[39] Christian Fraitzl, Thomas Kappe, Friederike Pennekamp, et al. “Femoral Head-Neck Offset Measurements in 339 Subjects”. In: *Knee Surgery, Sports Trauma, Arthroscopy* 21.5 (2013), pp. 1212–1217.

[40] Yarin Gal and Zoubin Ghahramani. “Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning”. In: *International Conference on Machine Learning*. PMLR. 2016, pp. 1050–1059.

[41] Reinhold Ganz, Javad Parvizi, Martin Beck, et al. “Femoroacetabular Impingement: A Cause for Osteoarthritis of the Hip”. In: *Clinical Orthopaedics and Related Research* 417 (2003), pp. 112–120.

[42] Christian Garbin, Xingquan Zhu, and Oge Marques. “Dropout vs. Batch Normalization: An Empirical Study of Their Impact on Deep Learning”. In: *Multimedia Tools and Applications* 79.19 (2020), pp. 12777–12815.

[43] Arkadiusz Gertych, Aifeng Zhang, James Sayre, et al. “Bone Age Assessment of Children Using a Digital Hand Atlas”. In: *Computerized Medical Imaging and Graphics* 31.4-5 (2007), pp. 322–331.

[44] Ross Girshick. “Fast R-CNN”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 1440–1448.

[45] Ross Girshick, Jeff Donahue, Trevor Darrell, et al. “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 580–587.

[46] David Golan, Yoni Donner, Chris Mansi, et al. “Fully Automating Graf’s Method for DDH Diagnosis Using Deep Convolutional Neural Networks”. In: *Deep Learning and Data Labeling for Medical Applications*. Springer, 2016, pp. 130–141.

[47] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, et al. “Generative Adversarial Networks”. In: *Communications of the ACM* 63.11 (2020), pp. 139–144.

[48] Robert Grupp, Mathias Unberath, Cong Gao, et al. “Automatic Annotation of Hip Anatomy in Fluoroscopy for Robust and Efficient 2D/3D Registration”. In: *International Journal of Computer Assisted Radiology and Surgery* (2020), pp. 1–11.

[49] Nitesh Gundavarapu, Divyansh Srivastava, Rahul Mitra, et al. “Structured Aleatoric Uncertainty in Human Pose Estimation”. In: *CVPR Workshops*. Vol. 2. 2019, p. 2.

[50] Chuan Guo, Geoff Pleiss, Yu Sun, et al. “On Calibration of Modern Neural Networks”. In: *International Conference on Machine Learning*. PMLR. 2017, pp. 1321–1330.

[51] Ali Harakeh, Michael Smart, and Steven Waslander. “BayesOD: A Bayesian Approach for Uncertainty Estimation in Deep Object Detectors”. In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2020, pp. 87–93.

[52] Abhilash Hareendranathan, Dornoosh Zonoobi, Myles Mabee, et al. “Toward Automatic Diagnosis of Hip Dysplasia from 2D Ultrasound”. In: *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. IEEE. 2017, pp. 982–985.

[53] Houssam El-Hariri, Kishore Mulpuri, Antony Hodgson, et al. “Comparative Evaluation of Hand-Engineered and Deep-Learned Features for Neonatal Hip Bone Segmentation in Ultrasound”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 12–20.

[54] Kaiming He, Georgia Gkioxari, Piotr Dollár, et al. “Mask R-CNN”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 2961–2969.

[55] Kaiming He, Xiangyu Zhang, Shaoqing Ren, et al. “Deep Residual Learning for Image Recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[56] Yihui He, Chenchen Zhu, Jianren Wang, et al. “Bounding Box Regression with Uncertainty for Accurate Object Detection”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 2888–2897.

[57] Xindi Hu, Limin Wang, Xin Yang, et al. “Joint Landmark and Structure Learning for Automatic Evaluation of Developmental Dysplasia of the Hip”. In: *IEEE Journal of Biomedical and Health Informatics* 26.1 (2021), pp. 345–358.

[58] Bulat Ibragimov, Boštjan Likar, Franjo Pernus, et al. “Automatic Cephalometric X-ray Landmark Detection by Applying Game Theory and Random Forests”. In: *Proc. ISBI Int. Symp. on Biomedical Imaging*. 2014.

[59] Fabian Isensee, Jens Petersen, Andre Klein, et al. “nnU-Net: Self-Adapting Framework for U-Net-Based Medical Image Segmentation”. In: *arXiv preprint arXiv:1809.10486* (2018).

[60] Ranjeet Jha, Gaurav Jaswal, Divij Gupta, et al. “PixISegNet: Pixel-level Iris Segmentation Network Using Convolutional Encoder–Decoder with Stacked Hourglass Bottleneck”. In: *IET Biometrics* 9.1 (2020), pp. 11–24.

[61] Heinrich Jiang, Been Kim, Melody Guan, et al. “To Trust or Not to Trust a Classifier”. In: *Advances in neural information processing systems* 31 (2018).

[62] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. “Snakes: Active Contour Models”. In: *International Journal of Computer Vision* 1.4 (1988), pp. 321–331.

[63] Robert Keys. “Cubic convolution interpolation for digital image processing”. In: *IEEE transactions on acoustics, speech, and signal processing* 29.6 (1981), pp. 1153–1160.

[64] Kostiantyn Khabarlak and Larysa Koriashkina. “Fast Facial Landmark Detection and Applications: A Survey”. In: *arXiv preprint arXiv:2101.10808* (2021).

[65] David Killock. “AI Outperforms Radiologists in Mammographic Screening”. In: *Nature Reviews Clinical Oncology* 17.3 (2020), pp. 134–134.

[66] Joost Kok, Egbert Boers, Walter Kosters, et al. “Artificial Intelligence: Definition, Trends, Techniques, and Cases”. In: *Artificial Intelligence* 1 (2009), pp. 270–299.

[67] Andrey Kolmogorov. “Sulla Determinazione Empirica di una Legge di Distribuzione”. In: *Giorn Dell’ist Ital Degli Att* 4 (1933), pp. 89–91.

[68] Florian Kordon, Peter Fischer, Maxim Privalov, et al. “Multi-Task Localization and Segmentation for X-ray Guided Planning in Knee Surgery”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 622–630.

[69] Florian Kordon, Ruxandra Lasowski, Benedict Swartman, et al. “Improved X-ray Bone Segmentation by Normalization and Augmentation Strategies”. In: *Bildverarbeitung für die Medizin 2019*. Springer, 2019, pp. 104–109.

[70] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Communications of the ACM* 60.6 (2017), pp. 84–90.

[71] Abhinav Kumar, Tim Marks, Wenxuan Mou, et al. "LUVLi Face Alignment: Estimating Landmarks' Location, Uncertainty, and Visibility Likelihood". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 8236–8246.

[72] Satish Kutty, Prism Schneider, Peter Faris, et al. "Reliability and Predictability of the Centre-Edge Angle". In: *International Orthopaedics* 36.3 (2012), pp. 505–510.

[73] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. "Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles". In: *Advances in Neural Information Processing Systems* 30 (2017).

[74] Michael Le, Frederik Diehl, Thomas Brunner, et al. "Uncertainty Estimation for Deep Neural Object Detectors in Safety-Critical Applications". In: *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE. 2018, pp. 3873–3878.

[75] Hansang Lee, Minseok Park, and Junmo Kim. "Cephalometric Landmark Detection in Dental X-ray Images Using Convolutional Neural Networks". In: *Medical Imaging 2017: Computer-Aided Diagnosis*. Vol. 10134. SPIE. 2017, pp. 494–499.

[76] Jeong-Hoon Lee, Hee-Jin Yu, Min-ji Kim, et al. "Automated Cephalometric Landmark Detection with Confidence Regions Using Bayesian Convolutional Neural Networks". In: *BMC Oral Health* 20.1 (2020), pp. 1–10.

[77] Si-Wook Lee, Hee-Uk Ye, Kyung-Jae Lee, et al. "Accuracy of New Deep Learning Model-Based Segmentation and Key-Point Multi-Detection Method for Ultrasonographic Developmental Dysplasia of the Hip (DDH) Screening". In: *Diagnostics* 11.7 (2021), p. 1174.

[78] Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, et al. "Leveraging Uncertainty Information from Deep Neural Networks for Disease Detection". In: *Scientific Reports* 7.1 (2017), pp. 1–14.

[79] Haofu Liao, Addisu Mesfin, and Jiebo Luo. "Joint Vertebrae Identification and Localization in Spinal CT Images by Combining Short- and Long-Range Contextual Information". In: *IEEE Transactions on Medical Imaging* 37.5 (2018), pp. 1266–1275.

[80] Tsung-Yi Lin, Michael Maire, Serge Belongie, et al. "Microsoft COCO: Common Objects in Context". In: *CoRR* abs/1405.0312 (2014). URL: `http://arxiv.org/abs/1405.0312`.

[81] Claudia Lindner, Paul Bromiley, Mircea Ionita, et al. "Robust and Accurate Shape Model Matching Using Random Forest Regression-Voting". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.9 (2014), pp. 1862–1874.

[82] Claudia Lindner, Ching-Wei Wang, Cheng-Ta Huang, et al. "Fully Automatic System for Accurate Localisation and Analysis of Cephalometric Landmarks in Lateral Cephalograms". In: *Scientific Reports* 6.1 (2016), pp. 1–10.

[83] Chuanbin Liu, Hongtao Xie, Sicheng Zhang, et al. "Misshapen Pelvis Landmark Detection by Spatial Local Correlation Mining for Diagnosing Developmental Dysplasia of the Hip". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 441–449.

[84] Ze Liu, Yutong Lin, Yue Cao, et al. "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 10012–10022.

[85] Gwilym Lodwick, Cosmo Haun, Walton Smith, et al. "Computer Diagnosis of Primary Bone Tumors: A Preliminary Report". In: *Radiology* 80.2 (1963), pp. 273–275.

[86] Jonathan Long, Evan Shelhamer, and Trevor Darrell. "Fully Convolutional Networks for Semantic Segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.

[87] David Lowe. "Object Recognition from Local Scale-Invariant Features". In: *Proceedings of the Seventh IEEE International Conference on Computer Vision*. Vol. 2. IEEE. 1999, pp. 1150–1157.

[88] James McCouat and Irina Voiculescu. "Contour-Hugging Heatmaps for Landmark Detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 20597–20605.

[89] James McCouat, Irina Voiculescu, and Siôn Glyn-Jones. "Automatically Diagnosing HIP Conditions from X-Rays Using Landmark Detection". In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2021, pp. 179–182.

[90] James McCouat, Irina Voiculescu, and Sion Glyn-Jones. "Fast and accurate automatic measurements of the hip from X-rays using AI". In: (2022).

[91] Bjoern Menze, Andras Jakab, Stefan Bauer, et al. "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)". In: *IEEE Transactions on Medical Imaging* 34.10 (2014), pp. 1993–2024.

[92] Phillip Meyers, Charles Nice Jr, Hal Becker, et al. "Automated Computer Analysis of Radiographic Images". In: *Radiology* 83.6 (1964), pp. 1029–1034.

[93] Hengameh Mirzaalian and Ghassan Hamarneh. "Automatic Globally-Optimal Pictorial Structures with Random Decision Forest Based Likelihoods for Cephalometric X-ray Landmark Detection". In: *Proc. ISBI International (a) Landmark* 4 (2014).

[94] Jishnu Mukhoti and Yarin Gal. "Evaluating Bayesian Deep Learning Methods for Semantic Segmentation". In: *arXiv preprint arXiv:1811.12709* (2018).

[95] Angela Murphy, Beate Ehrhardt, Celia Gregson, et al. "Machine Learning Outperforms Clinical Experts in Classification of Hip Fractures". In: *Scientific Reports* 12.1 (2022), p. 2058.

[96] Alejandro Newell, Kaiyu Yang, and Jia Deng. "Stacked Hourglass Networks for Human Pose Estimation". In: *European conference on computer vision*. Springer. 2016, pp. 483–499.

[97] Alexandru Niculescu-Mizil and Rich Caruana. "Predicting Good Probabilities with Supervised Learning". In: *Proceedings of the 22nd International Conference on Machine Learning*. 2005, pp. 625–632.

[98] Mehdi Noroozi and Paolo Favaro. "Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles". In: *European Conference on Computer Vision*. Springer. 2016, pp. 69–84.

[99] Kanghan Oh, Il-Seok Oh, Dae-Woo Lee, et al. "Deep Anatomical Context Feature Learning for Cephalometric Landmark Detection". In: *IEEE Journal of Biomedical and Health Informatics* 25.3 (2020), pp. 806–817.

[100] Hakan Omeroğlu, Ali Biçimoğlu, Suha Koparal, et al. "Assessment of Variations in the Measurement of Hip Ultrasonography by the Graf Method in Developmental Dysplasia of the Hip." In: *Journal of Pediatric Orthopedics. Part B* 10.2 (2001), pp. 89–95.

[101] David Opitz and Richard Maclin. "Popular Ensemble Methods: An Empirical Study". In: *Journal of Artificial Intelligence Research* 11 (1999), pp. 169–198.

[102] Antony Palmer, Vandana Ayyar-Gupta, Susan Dutton, et al. "Protocol for the Femoroacetabular Impingement Trial (FAIT) a Multi-Centre Randomised Controlled Trial Comparing Surgical and Non-Surgical Management of Femoroacetabular Impingement". In: *Bone & Joint Research* 3.11 (2014), pp. 321–327.

[103] Christian Payer, Darko Štern, Horst Bischof, et al. "Regressing Heatmaps for Multiple Landmark Localization Using CNNs". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2016, pp. 230–238.

[104] Christian Payer, Darko Štern, Horst Bischof, et al. "Integrating Spatial Configuration into Heatmap Regression-Based CNNs for Landmark Localization". In: *Medical Image Analysis* 54 (2019), pp. 207–219.

[105] Christian Payer, Martin Urschler, Horst Bischof, et al. "Uncertainty Estimation in Landmark Localization Based on Gaussian Heatmaps". In: *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis*. Springer, 2020, pp. 42–51.

[106] Jiahong Qian, Ming Cheng, Yubo Tao, et al. "CephaNet: An Improved Faster R-CNN for Cephalometric Landmark Detection". In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE. 2019, pp. 868–871.

[107] Quan Quan, Qingsong Yao, Jun Li, et al. "Which Images to Label for Few-shot Medical Landmark Detection?" In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 20606–20616.

[108] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, et al. "Zero-shot Text-to-Image Generation". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 8821–8831.

[109] Joseph Redmon, Santosh Divvala, Ross Girshick, et al. "You Only Look Once: Unified, Real-time Object Detection". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 779–788.

[110] Shaoqing Ren, Kaiming He, Ross Girshick, et al. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". In: *Advances in Neural Information Processing Systems*. 2015, pp. 91–99.

[111] 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*. Springer. 2015, pp. 234–241.

[112] Daniel Rueckert, Luke Sonoda, Carmel Hayes, et al. "Nonrigid Registration Using Free-Form Deformations: Application to Breast MR Images". In: *IEEE Transactions on Medical Imaging* 18.8 (1999), pp. 712–721.

[113] Fernando Ruiz Santiago, Alicia Santiago Chinchilla, Afshin Ansari, et al. "Imaging of Hip Pain: From Radiography to Cross-Sectional Imaging Techniques". In: *Radiology Research and Practice* 2016 (2016).

[114] Lawrence Schobs, Shuo Zhou, Marcella Cogliano, et al. "Confidence-Quantifying Landmark Localisation For Cardiac MRI". In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2021, pp. 985–988.

[115] Lawrence Schöbs, Andrew J Swift, and Haiping Lu. "Uncertainty Estimation for Heatmap-based Landmark Localization". In: *IEEE Transactions on Medical Imaging* (2022).

[116] Ramprasaath Selvaraju, Michael Cogswell, Abhishek Das, et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 618–626.

[117] Fahad Shamshad, Salman Khan, Syed Waqas Zamir, et al. "Transformers in Medical Imaging: A Survey". In: *arXiv preprint arXiv:2201.09873* (2022).

[118] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps". In: *arXiv preprint arXiv:1312.6034* (2013).

[119] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-scale Image Recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[120] Abhinav Singh, James McCouat, Daniel Perry, et al. "Automatic Identification of Clinical Landmarks and Calculation of Graf Angles in Newborn Hip Screening: A Pilot Study". In: (2022).

[121] Nikolai Smirnov. "Estimate of Deviation Between Empirical Distribution Functions in Two Independent Samples". In: *Bulletin Moscow University* 2.2 (1939), pp. 3–16.

[122] Thorvald Sorensen. "A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons". In: *Biologiske Skrifter* 5 (1948), pp. 1–34.

[123] Darko Štern, Thomas Ebner, and Martin Urschler. "From Local to Global Random Regression Forests: Exploring Anatomical Landmark Localization". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2016, pp. 221–229.

[124] Ke Sun, Bin Xiao, Dong Liu, et al. “Deep High-Resolution Representation Learning for Human Pose Estimation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 5693–5703.

[125] Christian Szegedy, Wei Liu, Yangqing Jia, et al. “Going Deeper with Convolutions”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 1–9.

[126] Xiaoli Tang. “The Role of Artificial Intelligence in Medical Imaging Research”. In: *BJR| Open* 2.1 (2019), p. 20190031.

[127] Terje Terjesen. “Residual Hip Dysplasia as a Risk Factor for Osteoarthritis in 45 Years Follow-up of Late-Detected Hip Dislocation”. In: *Journal of Children’s Orthopaedics* 5.6 (2011), pp. 425–431.

[128] Franz Thaler, Christian Payer, Martin Urschler, et al. “Modeling Annotation Uncertainty with Gaussian Heatmaps in Landmark Localization”. In: *Machine Learning for Biomedical Imaging* 1.UNSURE2020 special issue (2021), pp. 1–10.

[129] Jonathan Tompson, Arjun Jain, Yann LeCun, et al. “Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation”. In: *Advances in Neural Information Processing Systems* 27 (2014), pp. 1799–1807.

[130] Alexander Toshev and Christian Szegedy. “DeepPose: Human Pose Estimation via Deep Neural Networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 1653–1660.

[131] Martin Urschler, Thomas Ebner, and Darko Štern. “Integrating Geometric Configuration and Appearance Information into a Unified Framework for Anatomical Landmark Localization”. In: *Medical Image Analysis* 43 (2018), pp. 23–36.

[132] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. “Attention is All You Need”. In: *Advances in Neural Information Processing Systems* 30 (2017).

[133] Ching-Wei Wang, Cheng-Ta Huang, Meng-Che Hsieh, et al. “Evaluation and Comparison of Anatomical Landmark Detection Methods for Cephalometric X-ray Images: A Grand Challenge”. In: *IEEE Transactions on Medical Imaging* 34.9 (2015), pp. 1890–1900.

[134] Ching-Wei Wang, Cheng-Ta Huang, Jia-Hong Lee, et al. “A Benchmark for Comparison of Dental Radiography Analysis Algorithms”. In: *Medical Image Analysis* 31 (2016), pp. 63–76.

[135] Guotai Wang, Wenqi Li, Michael Aertsen, et al. “Aleatoric Uncertainty Estimation with Test-Time Augmentation for Medical Image Segmentation with Convolutional Neural Networks”. In: *Neurocomputing* 338 (2019), pp. 34–45.

[136] Jingdong Wang, Ke Sun, Tianheng Cheng, et al. “Deep High-Resolution Representation Learning for Visual Recognition”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.10 (2020), pp. 3349–3364.

[137] Sanghyun Woo, Jongchan Park, Joon-Young Lee, et al. “Cbam: Convolutional block attention module”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 3–19.

[138] Hai Wu, Hongtao Xie, Chuanbin Liu, et al. “CircleNet for Hip Landmark Detection”. In: *AAAI*. 2020, pp. 12370–12377.

[139] Steven Wyatt, Rowena Bailey, Patrick Moore, et al. "Equity of Access to NHS-Funded Hip Replacements in England and Wales: Trends From 2006 to 2016". In: *The Lancet Regional Health–Europe* 21 (2022).

[140] Dong Yang, Shaoting Zhang, Zhennan Yan, et al. "Automated Anatomical Landmark Detection on Distal Femur Surface Using Convolutional Neural Network". In: *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2015, pp. 17–21.

[141] Qingsong Yao, Zecheng He, Hu Han, et al. *Miss the Point: Targeted Adversarial Attack on Multiple Landmark Detection*. 2020. arXiv: `2007.05225 [cs.CV]`.

[142] Qingsong Yao, Quan Quan, Li Xiao, et al. "One-shot Medical Landmark Detection". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2021, pp. 177–188.

[143] Varduhi Yeghiazaryan and Irina Voiculescu. "Family of Boundary Overlap Metrics for the Evaluation of Medical Image Segmentation". In: *Journal of Medical Imaging* 5.1 (2018), p. 015006.

[144] Fisher Yu and Vladlen Koltun. "Multi-Scale Context Aggregation by Dilated Convolutions". In: *arXiv preprint arXiv:1511.07122* (2015).

[145] Minmin Zeng, Zhenlei Yan, Shuai Liu, et al. "Cascaded Convolutional Networks for Automatic Cephalometric Landmark Detection". In: *Medical Image Analysis* 68 (2021), p. 101904.

[146] Baihua Zhang, Shouliang Qi, Patrice Monkam, et al. "Ensemble Learners of Multiple Deep CNNs for Pulmonary Nodules Classification Using CT Images". In: *IEEE Access* 7 (2019), pp. 110358–110371.

[147] Guoyan Zheng. "Fully Automatic Segmentation and Landmarking of Hip CT Images". In: *EPiC Series in Health Sciences* 4 (2020), pp. 312–315.

[148] Zhusi Zhong, Jie Li, Zhenxi Zhang, et al. "An Attention-Guided Deep Regression Model for Landmark Detection in Cephalograms". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 540–548.

[149] Zongwei Zhou, Mahfuzur Siddiquee, Nima Tajbakhsh, et al. "UNet++: A Nested U-Net Architecture for Medical Image Segmentation". In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain* 11045 (2018), pp. 3–11.

[150] Heqin Zhu, Qingsong Yao, Li Xiao, et al. "You Only Learn Once: Universal Anatomical Landmark Detection". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2021, pp. 85–95.

[151] Xu Zou, Sheng Zhong, Luxin Yan, et al. "Learning Robust Facial Landmark Detection via Hierarchical Structured Ensemble". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 141–150.