The Australian National University
2600 ACT | Canberra | Australia

Australian National University

School of Computing

College of Engineering and Computer Science (CECS)

Automated Bone Segmentation and Knee Alignment Analysis of Long-Leg Radiographs

— 12 pt research project (S1/S2 2022)

A report submitted for the course
COMP8604, Research Project

By:
Zhe Xiong

Supervisors:
Dr. Nicolo Malagutti
Prof. Nick Barnes

November 2022

Declaration:

I declare that this work:

- upholds the principles of academic integrity, as defined in the University Academic Misconduct Rules;
- is original, except where collaboration (for example group work) has been authorised in writing by the course convener in the class summary and/or Wattle site;
- is produced for the purposes of this assessment task and has not been submitted for assessment in any other context, except where authorised in writing by the course convener;
- gives appropriate acknowledgement of the ideas, scholarship and intellectual property of others insofar as these have been used;
- in no part involves copying, cheating, collusion, fabrication, plagiarism or recycling.

November, Zhe Xiong

Acknowledgements

First, I would like to express my deepest thanks to my supervisors, Dr. Nicolo Malagutti and Prof. Nick Barnes. This project would not have been possible without their help and dedicated participation at every step of the process. I have no medical background. Through teleconferences and emails, Dr. Nicolo Malagutti has been very patient in providing me with medical engineering knowledge related to this project. When I couldn't get the data I needed for my experiments due to the network problem, he helped me download the data and walked me through the data. He also gave me valuable advice on formal academic writing. I wanted to convey my appreciation for your support and patience over the past year.

Second, I must express my heartfelt gratitude to my parents for their unfailing support and constant encouragement during my years of study at the Australian National University and throughout the project. Without them, it would have been impossible to complete this project. I am appreciative.

Abstract

In Total Knee Arthroplasty (TKA) surgery, knee alignment plays a crucial role in knee reconstruction. There is agreement that manual knee alignment is highly repetitive, time-consuming, and unreliable. In this project, a two-step method was proposed to automate the knee alignment process. First, the contours of the femur and tibia were predicted by a modified U-Net. Then the alignment was measured based on the geometry of the contours. 608 long-leg radiographs were selected from the Osteoarthritis Initiative (OAI) Project 60. Among them, 267 12-month visit radiographs were annotated and used for network training, and 341 36-month visit radiographs with readings of manual measurement were used to validate the proposed alignment measurement method. By experimenting, I found that the network performed best on the validation set when trained with the dice loss function. In the evaluation of segmentation performance, the network archived a 0.972 dice coefficient and a 0.980 mIoU on the test set. As for the evaluation of alignment measurements, the HKA angle measured by the proposed method was compared to the HKA angle measured manually in Project 60. The proposed method achieved a signed error of -0.03$\pm$0.78 degrees, and a mean absolute error of 0.58$\pm$0.52 degrees compared to the manual measurements. The Bland-Altman analysis showed a mean difference of -0.03 degrees with 95% LoA of $\pm$1.53 degrees. With the paired t-test, It could be found that there is an absence of evidence of a statistical difference between the measurements given by the proposed method and the manual measurements in Project 60.

Table of Contents

Chapter 1

Introduction

Total Knee Arthroplasty (TKA), also known as knee replacement surgery, is an effective treatment for severe knee joint disability and end-stage osteoarthritis (Carr et al. (2012)). TKA surgery involves the replacement of damaged bone with artificial implants with coronal plane alignment. Proper alignment plays a crucial role in the diagnosis stage of knee disease, artificial implant positioning, and post-surgical recovery evaluation. The mechanical Alignment (MA) (Insall et al. (1985)) method is a classical form of alignment in TKA. Due to its simplicity and consistency of operations, it rose to the status of the gold standard alignment method for TKA. There are also some skeptical voices. Bellemans et al. (2012) believed that the MA method disregards individual differences in coronal alignment. They found a natural varus pattern in 32% of men and 17% of women. In these circumstances, restoring the mechanical alignment to neutral may not be desired and cause biomechanical sequelae. However, the MA method continues to play an essential role in TKA. Hadi et al. (2015) considered the MA method to increase the longevity of implants and knee performance after TKA. MacDessi et al. (2021) used arithmetic HKA and joint line obliquity to classify the patient's knee into nine categories. They considered that patients in some types are better suited to the MA method. In addition, the MA method also guides the diagnosis of knee joint arthritic conditions. The MA draws attention to exceptional conditions, including the rapid degeneration of joints seen in arthritic joints accompanied by the obliquity of the articular surfaces (Cooke et al. (1991)).

Traditionally, knee alignment was performed manually, requiring physicians to define the related anatomical landmarks of long leg radiographs, draw the femoral and tibial axis with the anatomical landmarks, and calculate the hip-knee-ankle angle (HKA) as a general alignment measurement (Sheehy et al. (2011)). The evaluation of mechanical alignment from long-leg radiographs is highly standardized and has been very common in clinical and research practice in recent decades. However, a reliable alignment requires extensive anatomical expertise. Both Ilahi et al. (2001) and Schmidt et al. (2004)

reported significant differences in lower extremities measurements with different physicians. Additionally, evaluating a large number of radiographs requires a lot of time. Computers are well suited for this precise and highly repetitive type of work. Therefore, this project intends to leverage recent developments in computer vision and machine learning to construct a pipeline to automate the evaluation of mechanical alignment in long-leg radiographs. To complete this task, a modified U-Net (Ronneberger et al. (2015a)) is used to segment the femur and tibia on long-leg radiographs. Then four anatomical landmarks (the center of the femur, the center of the ankle, the center of the tibia, and the center of the ankle) are identified based on the geometry of the segmentation results. After that, the femur and tibia axes can be determined by the four anatomical landmarks. Lastly, the HKA is calculated by measuring the angle between the femur and tibia axes.

The summary of contributions is as follows:

1. Annotated lower extremities (femur and tibia) on 267 long-leg radiographs.
2. Trained a modified U-Net for lower extremity segmentation.
3. Identified the landmarks involved in mechanical alignment based on the geometry of the segmentation prediction.
4. Evaluated alignment measurements by comparing the HKA angle predicted by the proposed automated alignment method with the HKA angle manually measured in OAI project 60 (Sheehy et al. (2011)).

The report consists of five chapters. In the following chapter, the technical background regarding the definition of mechanical alignment is explained. Related works on automated alignment analysis and radiograph segmentation are reviewed. The Chapter 3 elaborates on the methods of segmentation and landmark location, and alignment measurement used in this project. In Chapter 4, the results of model training, segmentation, and alignment are presented and discussed. The Chapter 5 concludes the finding of this project and suggests possibilities for future works.

Chapter 2

Background

In this chapter, the background and standard of lower-extremity alignment measurement are supplemented in Section 2.1. The works of literature directly related to this project are reviewed and discussed in Section 2.2. The common segmentation methods on radiographs are reviewed and discussed in Section 2.3.

2.1 Standard of Alignment Measurements

To effectively address the problem of automated alignment analysis in lower extremity radiographs requires understanding the lower extremity's anatomical morphology. Therefore, this section aims to supplement the knowledge of common coronal plane alignment and related bone landmarks. At the end of the section, the standard of the mechanical axes used in this project is well explained.

The lower-extremity alignment in the coronal plane is described by way of two derived geometric entities, the mechanical axis of the femur (FM) and the mechanical axis of the tibia (TM). In Figure 2.1, the mechanistic axis of the femur and tibia is illustrated with a red dashed line. From the left side of the figure, the FM starts proximally at the center of the femur head, and it runs distally to the center of the knee. On the right side of the figure, the TM starts proximally in the center of the knee and runs distally to the center of the ankle.

Most of the literature favors using the center of the femoral head as the proximal point of FM and the center of the ankle as the distal point of TM. However, there are many different standards for defining the location of the center of the knee. Some literature favors using a single point to define the center of the knee. Subburaj et al. (2010) measured the mechanical alignment based on the single center point of the knee. This center point of the knee is the midpoint between the center of the femoral notch and the tibial spine notch, as illustrated in Figure 2.2. Using a single point as the center of the

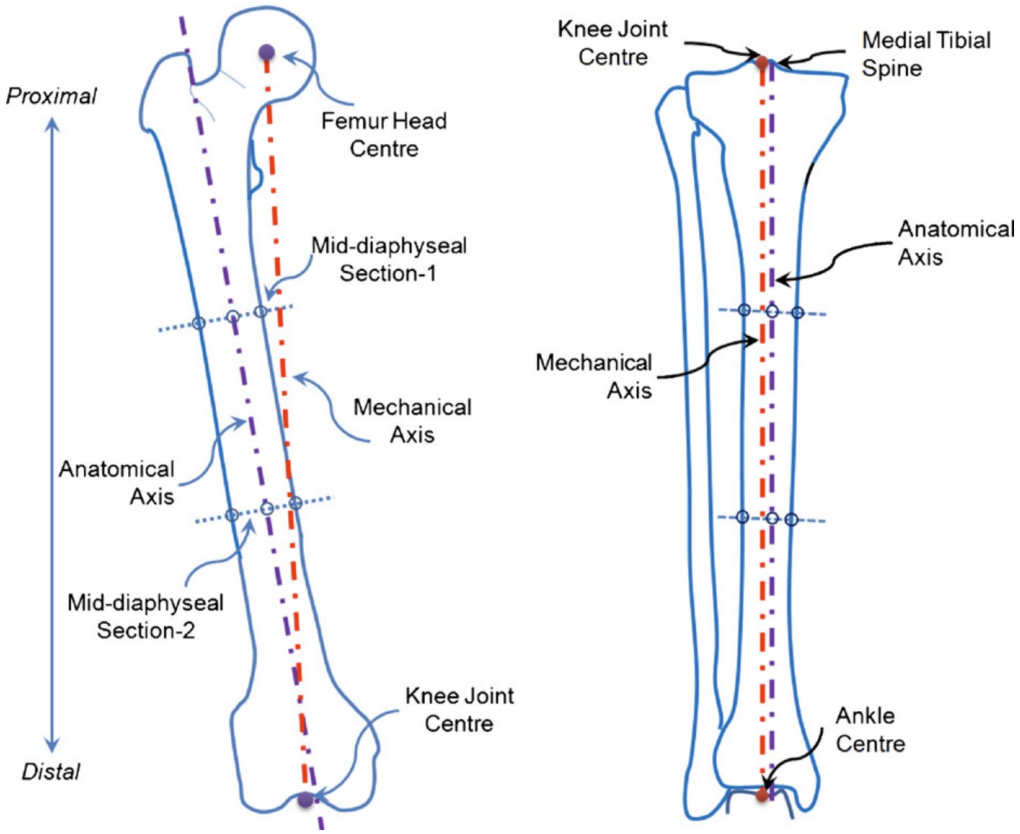


Figure 2.1: Mechanical and anatomical axis of the femur and tibia bone(Subburaj et al. (2010)).

knee for lower extremity alignment can also be found in other literatures. Kraus et al. (2005) used the midpoint of the tibial spines to construct the anatomic axis of the femur and the tibia. Brouwer et al. (2007) used the center of the tibial spines notch as one of the landmarks to measure the femorotibial angle. On the contrary, some literature supports using two separate points in the knee area for lower extremity alignments. Sled et al. (2011) suggested using the mid-condylar point of the distal femur and the center of the tibial plateau as points of reference for alignment. Their suggestion is based on previous anatomic studies in Yoshioka et al. (1987) and Yoshioka et al. (1989). Cooke et al. (2007) also agreed with this practice. They noted that using separate points for the surfaces of the femoral and tibial knee allows alignments to reveal the contributions of the femur and tibia. Their suggested method for mechanical alignment is illustrated in Figure 2.3.

In this project, the alignment algorithm is built based on the alignment measurement standard in Cooke et al. (2007). The ground-truth HKA values for algorithm validation are also read under such a standard. Therefore, the specifics of this standard must be clarified. The alignment standard is illustrated in Figure 2.3, FM is defined as an axis

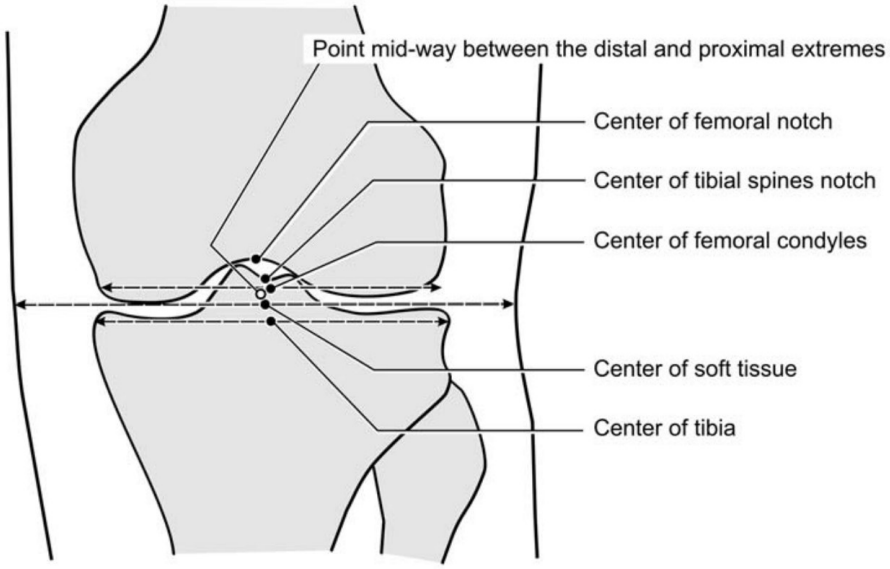


Figure 2.2: Different definitions of knee centre point(Cooke et al. (2007)).

that connects the center of the femur head and the mid-condylar point between the cruciate ligaments. TM is defined as an axis drawn proximally from the midpoint at the center of the tibial plateau distally to the center of the ankle. The hip-knee-ankle (HKA) angle is the angle between these two axes, and it is used to assess the overall alignment. The HKA in the figure is described as its angular deviation from 180 degrees. Therefore, the 0 degree HKA indicates that the alignment of the lower extremity is neutral (Figure 2.3 B). In this pattern, the FM and TM are co-linear and coincide with the load-bearing axis (LBA). In the varus pattern of the knee, FM and TM are located on the lateral side of the LBA, and the HKA value is negative (Figure 2.3 A). In the valgus pattern of the knee, FM and TM are located on the medial side of the LBA, and the HKA value is positive (Figure 2.3 C).

2.2 Works on Automated Analysis of Alignment

In recent years, automated alignment analysis on long-leg radiographs has become an active task. Accurately identifying the required anatomical landmarks plays an essential role in this task. There are two main approaches to locating landmarks. One is the learning-based approach. Another is a two-step approach that combines learning and morphology.

For the learning-based approach, a model is trained to directly predict the required landmarks on the given radiograph. Nguyen et al. (2020) proposed an automated system to measure common angles in the alignment of the lower extremities. They trained a convolutional neural network (CNN) to predict anatomic landmarks directly, then the

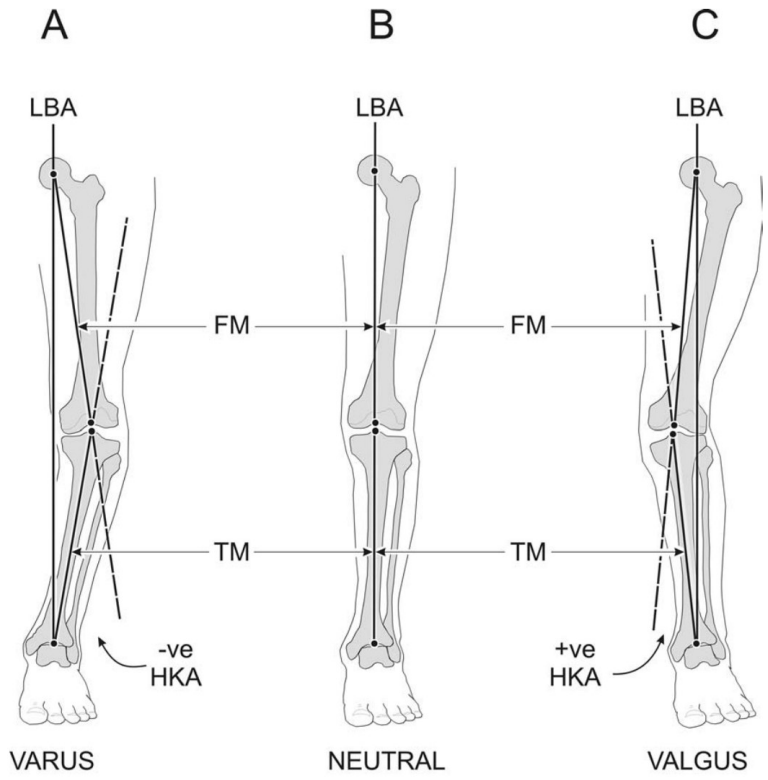


Figure 2.3: Common coronal plane alignment patterns(Cooke et al. (2007)). FM: mechanical axis of the femur. TM: mechanical axis of the tibia. LBA: load-bearing axis. HKA: hip-knee-ankle angle, which is the angle between FM and TM. The knee on the lateral side of the LBA has a negative HKA value, and the knee on the medial side of the LBA has a positive HKA value.

angles were calculated based on those landmarks. The method achieved less than 1.5 degrees deviation in 82.3% of the cases. Jo et al. (2022) trained a CNN-based model with 11,212 annotated long-leg radiographs. The annotation includes 15 anatomical landmarks, which are manually marked by orthopedic specialists. With predicted landmarks, the landmarks yielded a mean absolute error of 0.22 degrees in the HKA angle measurement.

In terms of the two-step approach, a model is trained to predict the contour or segmentation of related physiological structures, then the landmarks are located based on the geometry of the contour or segmentation. Gielis et al. (2020) proposed an automated pipeline to measure the HKA angle. Geometry-based search algorithms are used to find anatomic landmarks on long-leg radiographs. After that, the HKA angle is calculated based on the landmarks. The proposed pipeline achieved a mean absolute error of 1.8 degrees in the measurement of the HKA angle. Pei et al. (2021) trained a fine-tuned

U-Net (Ronneberger et al. (2015a)) to only segment the area of the femur head, knee, and ankle on long-leg radiographs, then three landmarks are defined as the centers of those areas. Their methods achieved a mean absolute error of 0.49 degrees.

There is a major pitfall in the learning-based approach. The model's performance in this approach highly relies on the annotation quality. Accurately marking landmarks by hand requires extensive expertise in orthopedics and anatomy. Besides, the reliability of manual landmark annotation can also be questioned. Both Ilahi et al. (2001) and Schmidt et al. (2004) documented significant differences in lower extremity alignment measurements on radiographs with inter-and intrareader. The two-step approach also requires annotating the contours of the bone structures. But, the level of refinement of the annotation is far less than that required by the learning-based approach. People who are generally familiar with the anatomic characteristics of the lower extremities are able to perform the annotation excellently. For the above reasons, the two-step approach was adopted for this project. Although there is no excessive requirement for the contour annotation of the contour, the quality of the segmentation results will directly affect the positioning of the landmarks and, thus, the accuracy of the alignment. Therefore, finding a suitable segmentation method for the automated alignment task is crucial. In the next section, the literatures on medical imaging segmentation are reviewed.

2.3 Works on Radiograph Segmentation

Image segmentation is a method used in image processing to divide an image into distinct parts based on the pixel characteristics present in the image. Segmentation can help to extract the desired object from the background. After that, further operations can be performed based on the object's morphology.

Segmentation of radiographs, such as x-ray images, is a major application in the field of image segmentation. Extracting the desired bone structure from a long-leg radiograph is challenging. First, the quality of a radiograph is poor. The image usually contains many noises, and there are low-contrast areas in the image. Then, tissues and bone structures overlapped and crowded in the hip region. Extracting the shape of the femoral head from the hip region becomes difficult. Although there are methods that effectively delineate the desired structures from radiographs, none of them are completely suitable for segmenting all types of medical images. Each method has its restrictions. The following section will introduce and analyze some common segmentation methods for radiographs.

Thresholding-Based This type of method is the most common method used in medical image segmentation. With a threshold value, the method transfers a grayscale image to a binary image. This method is simple and quick. It works well to extract the structure from the background when there is a strong contrast. However, it has a compromised performance when the edge of the structure is blurry, or the structure has overlapping areas. Bharodiya and Gonsa (2019) used an improved adaptive thresholding method

for the segmentation task on hand and arm radiographs. Kiran et al. (2019) used the Sauvola thresholding method to extract the lung field area from the chest radiographs. Maesyaroh et al. (2021) used Otsu thresholding method to extract objects from the chest radiographs.

Clustering-Based This type of method segments the image by grouping the regions with similar attributes. By the nature of clustering, this method is unsuitable for all segmentation tasks. It is more commonly used in the process of Magnetic Resonance brain images rather than radiographs. However, Ray and Sasmal (2010) experimented with combining K-means and Hierarchical clustering methods to extract bone structures from different radiographs, including hand, skull, chest, backbone, and knee radiographs. The method achieves comparable accuracy compared to existing methods. Florea et al. (2011) used Expectation The maximization (EM) algorithm determines an optimal thresholding value and then achieves a robust segmentation based on this thresholding value. Aradhya et al. (2021) used a cluster-Based method to classify chest radiographs into four conditions, including pneumonia bacterial, pneumonia virus, COVID-19, and normal.

Edge-Based This type of method detects the discontinuity in the image. It delineates the edge of the structure in the radiograph. Some notable edge detectors are Sobel, Laplacian of Gaussian, Canny, Prewitt, and Zero Cross. Goswami and Misra (2016) compared the performances of commonly used edge detectors on radiographs. The result shows that Zero Cross and Laplacian of Gaussian give the best segmentation with multiple types of radiographs. This type of method also has its constraints. They are susceptible to noise and require a well-designed threshold value to filter out redundant boundaries. Besides, the methods usually yield disjoint boundaries, thus further operations are required to connect those boundaries for closed segmentation.

Atlas-Based In this type of method, an atlas refers to a set of images with expert annotations on the desired segmentation regions. The annotations are usually obtained by manual delineation of the same anatomical structures. Compared to the segmentation methods based on deep learning, atlas-based methods require significantly fewer annotated images. Generally, the current atlas-based segmentation method uses either a probabilistic atlas or a non-probabilistic atlas (Feng (2006)). In a probabilistic atlas, the assignment of pixels to the anatomical structure is presented with probabilities, while a non-probabilistic atlas uses a hard assignment for the pixels. Because of the prior knowledge guidance, This type of method can handle complex segmentation tasks, such as segmentation of the femur from a hip radiograph (Chen et al. (2005)). In order to measure the femoral-tibial angle from the knee radiograph, Wahyuningrum et al. (2020) leveraged the active shape model to delineate the contours of the femur and tibia. To segment the tibia and femur from a knee radiograph, Seise et al. (2005) proposed a double contour active shape model. Candemir et al. (2016) used a probabilistic atlas-based method to segment rib bones from a chest radiograph. In the experiment, 25 radiographs are annotated on rib bones. With 20 annotated images, Boukala et al. (2004) extracted pelvic and femur structures from the hip radiograph.

Learning-Based In the last decade, deep learning has grown in popularity in the field of medical image processing. One of the challenges of deep learning in processing medical images is the generally small size of medical image datasets and the lack of annotation. U-Net (Ronneberger et al. (2015b)) is a pioneering work to tackle this challenge. With the U-like encoder-decoder architecture, the network improved the performance of medical image segmentation significantly. Due to the limited number of annotated training images, random elastic deformations are applied to training images. With this data augmentation method, the U-Net results average 92% IOU with 35 partially annotated training images on the "PhC-U373" dataset, and average 77% IOU with 20 partially annotated training images on "DIC-HeLa" dataset. Later, Oktay et al. (2018) improved the U-Net by integrating the attention gate into the original network. The attention gate is added to each layer of the decoder and results in improved prediction performance in various medical datasets. Other extended works based on U-Net include Recurrent Residual U-Net (Alom et al. (2019)), Inception U-Net (Punn and Agarwal (2020)), U-Net++ (Zhou et al.), and Generative Adversarial U-Net (Chen et al. (2021)). As for application on radiograph segmentation, Shu et al. (2019) used a modified U-Net, LVC-Net, to segment the pelvic and femur in the hip radiograph. Based on U-Net architectures (Ding et al. (2019)) proposed a lightweight architecture, a multi-scale convolutional neural network, to segment hand bones for hand radiographs. In order to measure the hip–knee–ankle angle, Pei et al. (2021) leveraged U-Net to segment the head of the femur, knee, and ankle structures from lower limb radiographs.

According to the reviews on existing radiographs segmentation methods, the learning-based segmentation method is more aligned with the goal of this project. The unsupervised methods, such as thresholding-based, clustering-based, and edge-based methods, are applicable to segment well-defined, non-overlapping physiological structures, such as the hand, arm, and chest. Guided by prior knowledge, the atlas-based methods can handle the more delicate segmentation tasks. However, the method performs compromised when segmenting the bone structures from an overlapping area, such femoroacetabular joint. Besler et al. (2017) found that the atlas-based method causes femur segmentation leaks into the pelvis in their experiments. In this project, accurate extraction of the femoral head contour from the femoroacetabular joint is important for the subsequent localization of the center of the femoral head. Therefore, the atlas-based methods are not adopted. In recent works (Ding et al. (2019), Shu et al. (2019), and Pei et al. (2021)), the U-Net-based architecture was used to segment the femur and tibia on long-leg radiographs. The segmentation results are superior to previous work. Therefore, this project adopts a modified U-Net architecture for the lower extremity segmentation.

Chapter 3

Proposed Methods

This chapter elaborates on the methods used in the project. The source and characteristics of the data are presented in Section 3.1. The criteria of data inclusion and the pipeline of data preprocessing are introduced in Section 3.2. The architecture of the network for segmentation is illustrated and explained in Section 3.3. The training details including loss functions, training scheme, and hyperparameters are presented in Section 3.4. The algorithms for landmark locating are explained in Section 3.5.

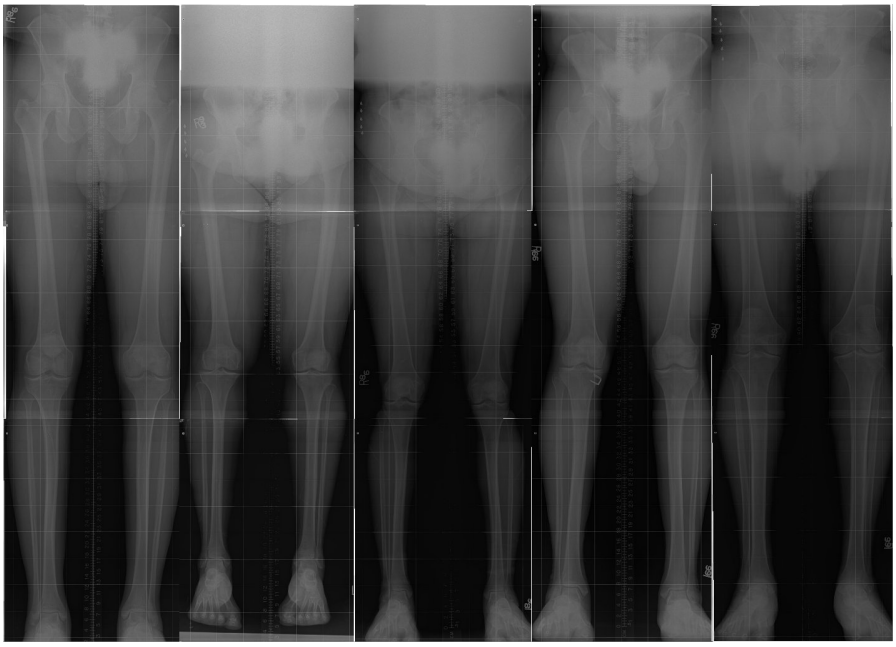
3.1 Dataset

The Osteoarthritis Initiative (OAI) study[1] is a multi-center, longitudinal, prospective observational study of knee osteoarthritis (OA). The dataset used in this project is acquired from the OAI Project 60 (Sled et al. (2011)), which is an OAI-funded and OAISYS-executed alignment assessment reading study. The dataset includes the central assessments of coronal plane lower extremity mechanical alignment (HKA angle) from OAI long-leg radiographs. The dataset in OAI Project 60 contains repeated participant monitoring visits at varied time intervals, including 12-month visit, 24-month visit, 36-month visit, and 48-month visit. The varying time interval between visits enables the data used to compare the values of a variable at different periods. In the experiments, the 12-month visit data are used for training the network to do bone segmentation, and the 36-month visit data are used to validate the alignment measurements. This dataset was chosen for the experiments because it is a publicly available dataset for the study of knee osteoarthritis, thus it is easily accessible. It contains many long-leg radiographs and the corresponding alignment measurement data, which is required to validate the proposed method.

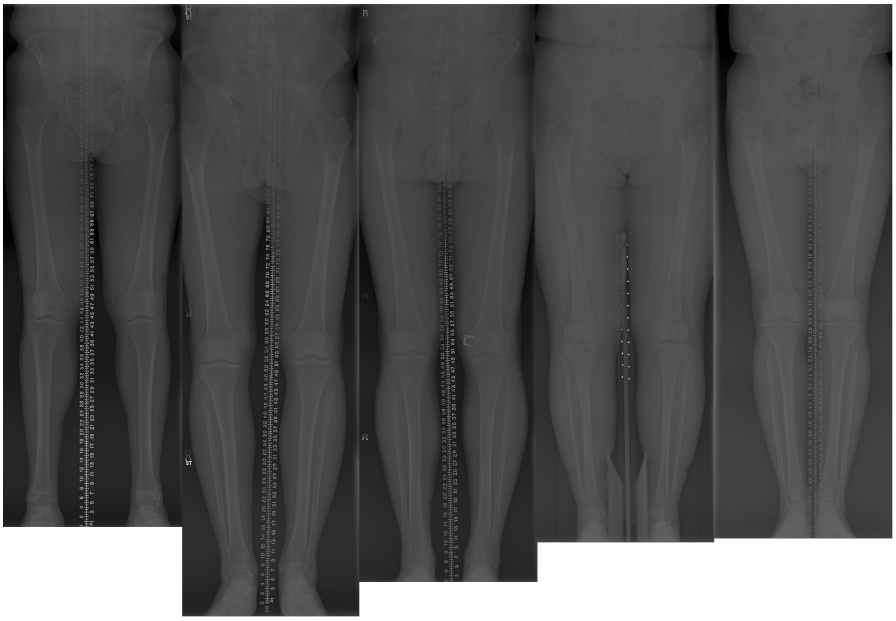
There are two types of radiographs in the dataset. The scanned radiographs are shown

[1] https://nda.nih.gov/oai/

in Figure 3.1a, and the digital radiographs are shown in Figure 3.1b. In terms of image size, the digital radiographs are larger in width. The digital radiographs have an image pixel width of 501, and scanned radiographs have an image pixel width of 256. Besides, both types of radiographs have varied image heights, but scanned radiographs have a smaller difference in height than digital radiographs. In terms of image quality, digital radiographs show higher resolution and more details. The brightness and contrast in scanned radiographs are less uniform than in digital radiographs. Due to the differences, it is necessary to standardize the size and quality of the radiographs in the pre-processing.



(a) Scanned radiographs



(b) Digital radiographs

Figure 3.1: Two types of radiographs in the dataset. Overall, the quality of scanned radiographs is poorer than digital radiographs.

3.2 Preprocessing

Data Exclusion After data collection, the initial step of data preprocessing is to exclude low-quality radiographs. Low-quality radiographs can be categorized into seven types, as shown in Figure 3.2. The types of overexposure, underexposure, and soft tissue are exclusively seen in scanned radiographs (Figure 3.1a). The remaining types existed in both scanned and digital radiographs. This step ensures the annotation's quality and the segmentation's accuracy. The reasons for excluding low-quality radiographs are the following. First, the low-quality radiographs contain bone with ambiguous or discontinued contours, which are difficult to annotate objectively. Second, artificial implants can hide the natural bone contours and show a different brightness in radiographs. These differences interfere with the prediction of the network on bone segmentation.

The radiographs inclusion criteria are as follows:

1. The radiograph clearly shows the contour of the femur and tibia, especially in the area of the femur head, knee, and ankle.
2. The bone structure is complete on at least one side.
3. No artificial implants in the bone structure.

Figure 3.2: Seven types of low-quality radiographs to exclude.

Data Preprocessing After excluding the low-quality radiographs, there were 267 radiographs left in the 12-month visit dataset, and 341 radiographs left in the 36-month visit dataset. Then the raw 12-month visit dataset was preprocessed for training the network to perform the segmentation task. The process of preprocessing is illustrated in Figure 3.3.

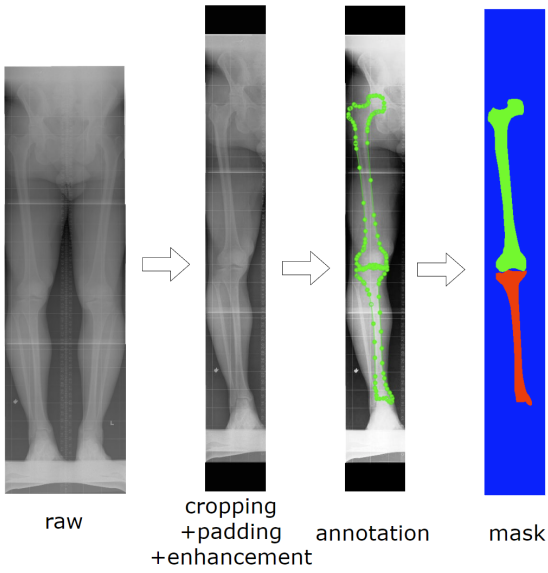


Figure 3.3: Process of radiograph preprocessing and target mask generation.

Because of the inconsistency of raw radiograph size, the first step was to resize all the radiographs to size 256 x 1024 with black pads on the boards if necessary. Then the radiographs were copped to unilateral ones. If the left side of the lower limb was kept, the radiograph was horizontally flipped. After the operations, the size of the radiograph was 128 x 1024. Next, the operation of histogram equalization was used to reduce unevenness in brightness and contrast in raw radiographs. After that, the contours of the femur and tibia were annotated manually with Labelme (Wada (2018)). Last, the 3-channel masks were generated from the points of the contours. In the mask, the green channel was used to label the femur, the red channel was used to label the tibia, and the blue channel was used to label the background. The preprocessed radiographs were used as the input to train the network, and the masks were used as labels to guide the training.

Data Augmentation The data augmentation methods based on spatial transformations, such as rotation, shifting, and zooming, are common practices in data augmentation. Teschock2020automated the authors augmented the input radiographs with a random rotation ranging from -10° to 10°, and scaled randomly by a coefficient ranging from 0.8 to 1.2. A similar practice can also be found in Shen et al. (2021). The authors randomly rotated the input radiographs from -45° to 45°, followed by random shifting and scaling operations. Inspired by previous work, rotation, shifting, and zooming operations have also been experimentally applied to input radiographs in the project. First, the radiograph is expanded to 1024x1024 with black pads surrounding the radiograph. The purpose of the expansion is to avoid losing pixels during the spatial transformation. Next, each radiograph and the corresponding mask were rotated at an angle randomly

selected from the uniform distribution between -30° and 30°, then randomly scaled by a factor between 0.9 and 1.1. The spatial transforms were performed 20 times on radiographs in the OAI 12-month visit dataset. The data augmentation operations expanded the data size from 267 to 5,340. Then, 4,000 augmented radiographs were sampled to train the network and the rest were 1,340 radiographs for testing. The data augmentation process is illustrated in Figure 3.4. In addition, data augmentation methods based on non-spatial transformations are also applied in the experiment. The contrast and brightness of the input radiographs were adjusted by a factor randomly selected from the uniform distribution between 0.7 and 1.3. This augmentation method ensures that the model performs robustly for different contrast and brightness radiographs.

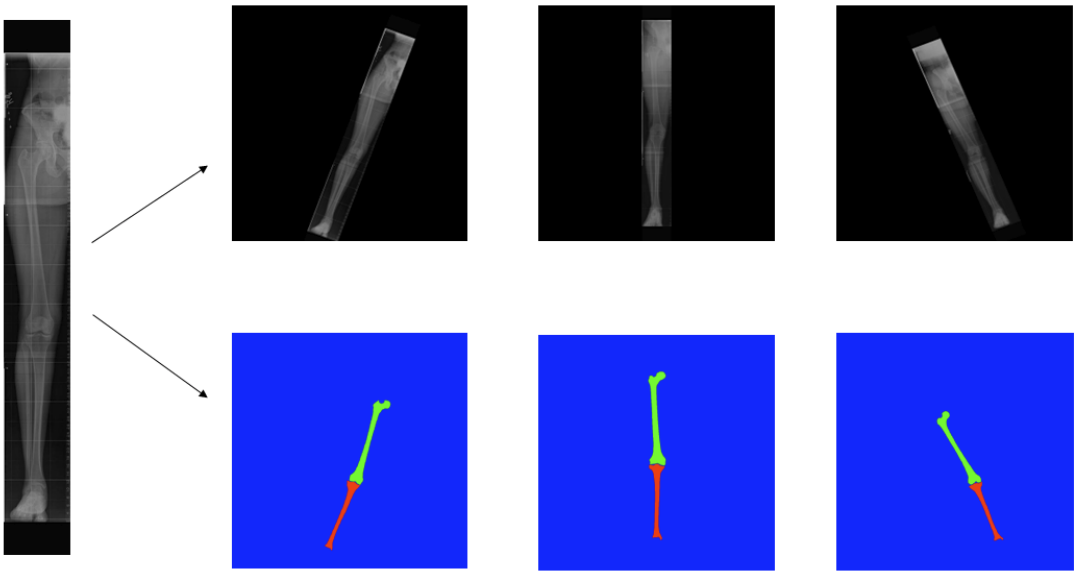


Figure 3.4: Illustration on the process of data augmentation. Each preprocessed radiograph was expanded to 1024x1024 with black pads, then randomly rotated, shifted, and zoomed.

3.3 Architecture

A modified U-Net (Ronneberger et al. (2015a)) architecture is leveraged for automated bone segmentation in the project. The architecture of the network is illustrated in Figure 3.5. It is a mirrored encoder-decoder architecture. The encoder retrieves features of the pixels from the input radiograph, and then the decoder reconstructs a predicted mask from the retrieved features. The The encoder consists of 4 DoubleConv blocks. Each block comprises two successive 3x3 convolution layers with stride of 1 and padding of 1, followed by a BatchNorm layer and a ReLU activation layer. At the end of each DoubleConv block, a 2x2 max pooling layer with stride of 2 downsamples the output features

from the DoubleConv block. The depth of the features doubled with the operation of downsampling . The depth started at 32 at the first DoubleConv block and increased to 512 at bottleneck. Similar to the structure of the encoder, the decoder consists of 4 DoubleConv blocks. Each block is identical to the block in the encoder. But at the end of each block in the decoder, there is a 2x2 transposed convolution with stride of 2 and no padding for upsampling. Each upsampling operation halves the depth of the features. The features output from the DoubleConv block in the encoder are concatenated onto the corresponding features after the operation of upsampling via skip connection. After decoding the features, a 1x1 convolution layer converts the depth of the features from 32 to 3, followed by a softmax layer to output a raw mask with the size of 3x1024x128. Last, the RGB mask can be reconstructed based on the raw mask.

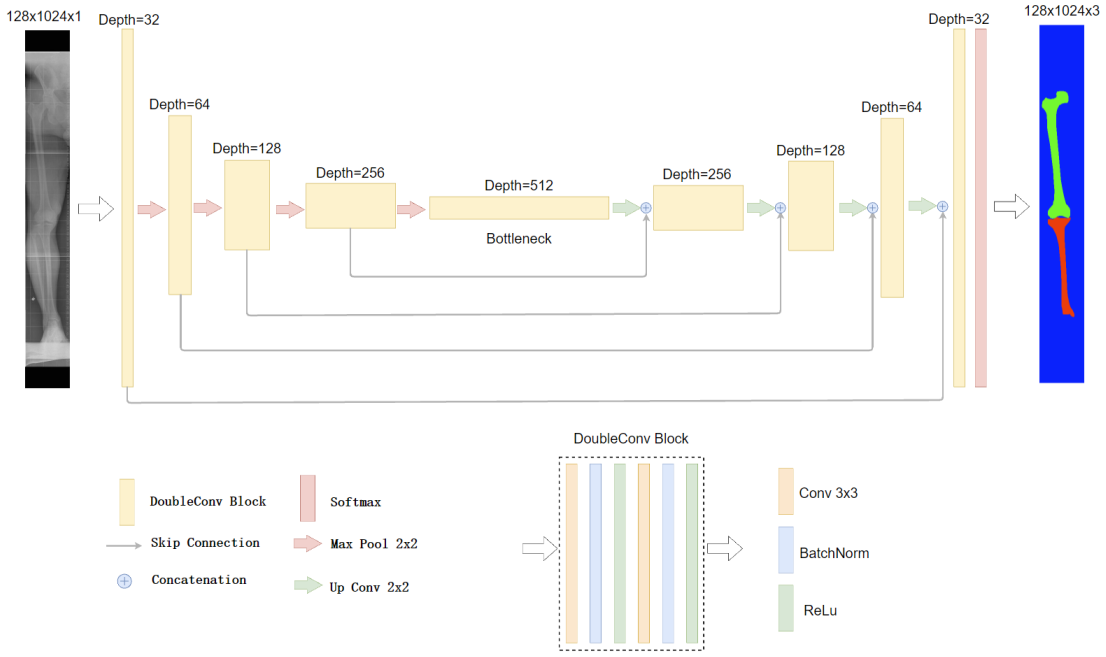


Figure 3.5: U-Net architecture used for bone segmentation. The blocks and layers are illustrated as rectangles of different colors, and the operations are illustrated as arrows. The numbers above the block are the output depth numbers of the blocks. The network inputs grayscale radiographs and outputs a 3-channel RGB mask.

Compared to the original implementation in Ronneberger et al. (2015a), The architecture of the network in this project has some notable modifications. First, there is no cropping operation before skip connections. In the original work of U-Net, the task of the network was to segment the cells in microscopical images. The authors proposed an overlap-tile strategy to improve segmentation accuracy on the image's border. In this strategy, the input image is extrapolated by reflective padding. Therefore the cropping operation

before the skip connection allows the feature map to concatenate to the upsampling outputs in the decoder. In addition, the BatchNorm layer is added after each convolution for more stable training. Last, the base depth (the depth of the feature after the first DoubleConv block) is 32 rather than 64 in the original implementation. The reduction in the depth reduces the total number of parameters in the network from 31.03M to 7.76M. The purpose of this modification is to avoid overfitting.

3.4 Training Details

Loss Function The loss function is used as a training guide for networks. The function maps the differences between the network's prediction and the target to a real number, which is usually called loss (or cost). The smaller the value of the loss, the closer the network's prediction is to the target. In the segmentation task, smaller loss value represents a more accurate segmentation prediction. There are three loss functions tested in the experiments. They are cross entropy loss, focal loss, and dice loss. The original U-Net (Ronneberger et al. (2015a)) used cross entropy loss for training. Shen et al. (2021) used dice loss to train a lightweight U-Net for bone segmentation. With the concern of foreground-background imbalance in radiographs, Schock et al. (2020) used the focal loss to train a modified U-Net for bone segmentation. The cross entropy loss is a distribution-based Loss, which measures the dissimilarity between two distributions (Ma (2020)). The cross entropy is given by

$$L_{\text{CE}} = -\sum_{c=1}^{C} y_c \log\left(p_c\right) \tag{3.1}$$

where C is the number of the classes, and y_c is the target of the class c. The p_c is the softmax probability of class c. In this project, the C is equal to 3, since the segmentation task involved 3 classes (femur, tibia, and background).

Based on the cross entropy loss, the focal loss (Lin et al. (2017)) is proposed to handle the problem of foreground-background class imbalance. The focal loss is given by

$$L_{focal} = -\sum_{c=1}^{C} y_c \left(1 - p_c\right)^{\gamma} \log\left(p_c\right) \tag{3.2}$$

where C is the number of the classes, y_c is the target of the class c, the p_c is the softmax probability of class c, and the γ is the modulator, which gives more weights to hard-classified examples, and fewer weights to easy-classified examples. For instance, when p_c is close to 1, the $(1 - p_c)$ is close to 0, thus the weight of the easy-classified example is small. When γ is equal to zero, the function of focal loss (Equation 3.2) is identical to that of entropy loss (Equation 3.1).

The dice loss (Milletari et al. (2016)) respects the nature of the segmentation. It directly optimizes the dice coefficient, which is one of the common metrics in the segmentation task. The dice loss is given by

$$L_{dice} = 1 - \sum_{c=1}^{C} \frac{|P_c \bigcap T_c| \times 2 + \text{s}}{|P_c| + |T_c| + \text{s}} \tag{3.3}$$

where C is the number of the classes, P_c is the predicted pixels of class c, T_c is the predicted pixels of class c, and s is the smooth coefficient, and it set to $1e^{-8}$ in the implementation. The smooth coefficient is intended to prevent the denominator from being 0 when both P_c and T_c are equal to 0 (the object of class c is absent from the radiographs).

To determine which loss function to train the network with will give the best segmentation performance, an experiment was conducted in Section 4.1.

Training Setup The OAI 12-month visit dataset was used to train the network. There were 267 long-leg radiographs in the dataset after data exclusion. The radiographs were preprocessed and annotated with the operations illustrated in Section 3.2. According to the different sets of augmentation methods, the size of the dataset is different. With the set of brightness change and contrast change, the augmented methods were randomly applied to the input radiographs as explained in 3.2, and the size of the dataset remained the same. With a total of 267 examples, the dataset consisted of 194 training examples, 48 validation examples, and 25 test examples. With the set of rotation, shifting, zooming, brightness change, and contrast change, the original 267 examples were augmented and expanded to 5,340 examples as explained in Section 3.2. Then 3,200 examples were used for training, 800 examples were used for validation, and the rest 1,340 examples were used for testing. The cross-validation method was not adopted in training, thus the examples in training set, validation set, and test set were fixed during the training process. The network was implemented with PyTorch, and trained on a rented server[2] with a single NVIDIA GeForce RTX 2080 Ti 11GB GPU, 6× Xeon E5-2678 v3 CPU, and 62GB RAM.

Training Scheme The total number of training epochs was 100, the batch size was 8, and the learning rate was 0.003. The optimizer was the Adam gradient descent algorithm with $\beta_1 = 0.9$, and $\beta_2 = 0.999$. The optimizer was used to minimize the dice loss (Equation 3.3). During the validation and testing stage, the network was evaluated by dice coefficient and mIoU. The dice coefficient is equal to $1 - L_{dice}$. The mIoU is defined as

$$mIoU = \sum_{c=1}^{C} \frac{P_c \bigcap T_c}{P_c \bigcup T_c} \tag{3.4}$$

[2] https://matpool.com/

where C is the number of the classes, P_c is the predicted pixels of class c, and the T_c is the target pixels of class c.

3.5 Identification of Anatomical Landmark Locations

In Chapter 2, the landmarks in this project are defined. There are 4 landmarks involved in the HKA computation. They are the center of the femur head, the center of the knee, the center of the tibia, and the center of the ankle. Among them, the line connecting the first two landmarks is defined as the mechanical axis of the femur (FM), and the line connecting the last two landmarks is defined as the mechanical axis of the tibia (TM). The hip-knee-ankle (HKA) angle between the FM and TM can be calculated to assess the overall alignment.

The identification of anatomical landmark locations was performed on the predicted mask. Generally, landmark identification requires four steps. First, the binary mask of the femur or tibia is extracted from the predicted RGB mask. Second, the binary mask is denoised by a morphological opening with a 10x10 kernel. Third, the regional mask is located based on the spatial location of the bone. Last, the landmark is identified based on the regional mask's geometry. The landmark identification process is illustrated in Figure 3.6.

To identify the center of the femur head, the largest circle fits the regional mask of the femur head, as shown in Figure 3.6 A, and the center of the circle is the center of the femur head. Specifically, the region mask was created by cropping the upper right of the femoral region (the first 10% of the pixels in the vertical direction and the last 65% of the pixels in the horizontal direction). Then the operator of distance transformation (Borgefors (1986)) was used to generate a distance map with the original mask. The operator calculates the euclidean distance from each pixel in the regional mask to the nearest pixel with a value of zero. The coordinate of the largest value in the distance map is the center of the circle (or the approximate center of the femur head), and the largest value is the circle's radius.

To identify the center of the knee, a line fits the contour of the middle femoral shaft as illustrated in 3.6 B. The region mask was created by cropping the middle of the femoral region (cutting the upper and lower 10% pixels in the vertical direction). Then the contour was obtained by subtracting the mask with morphological erosion from the original mask. The least squares method fitted the line to the contour points by minimizing the sum of the horizontal distances from all contour points to the line. The intersection between the line and the distal-most pixel of the femur is the approximate center of the knee (the red point as shown in Figure 3.7 A).

To identify the center of the tibia and the center of the ankle, a line fits the contour of the middle tibial shaft as illustrated in 3.6 C. The region mask was created by cropping the middle of the tibial region (cutting the upper and lower 10% pixels in the vertical direction). Then, the least squares method fitted a line to the contour points. The

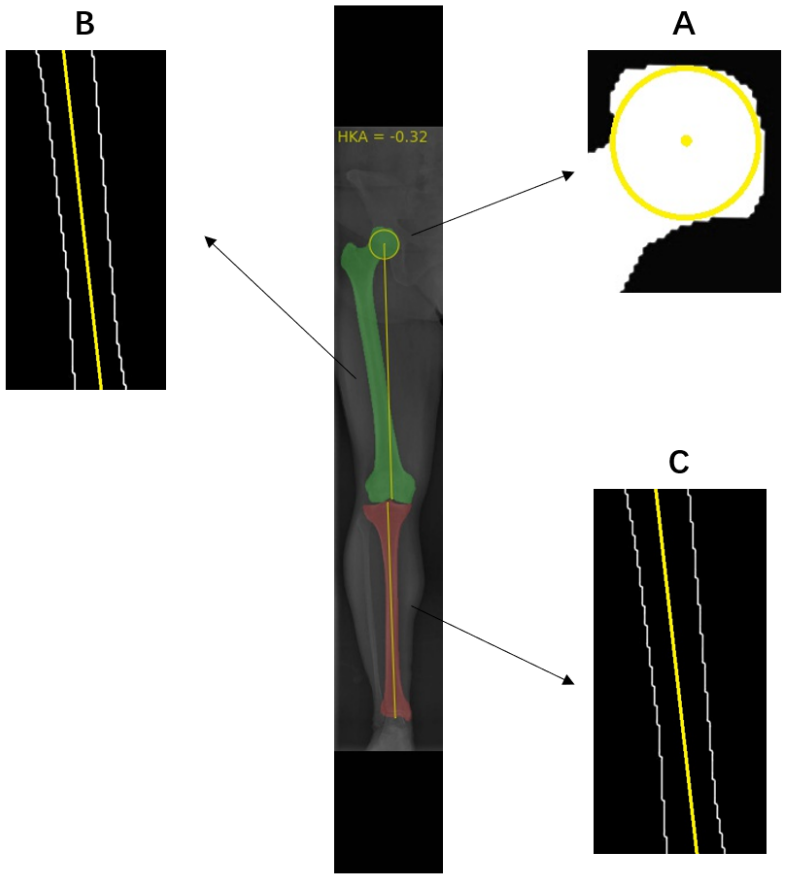


Figure 3.6: Process of landmarks identification. A: fitting a circle to the contour of the femur head, B: fitting a line to the contour of the middle femoral shaft, C: fitting a line to the contour of the middle tibial shaft.

intersection between the line and the proximal end pixel of the tibia is the approximate center of the tibia (the red point as shown in Figure 3.7 B), and the intersection between the line and the distal-most pixel of the tibia is the approximate center of the ankle (the red point as shown in Figure 3.7 C).

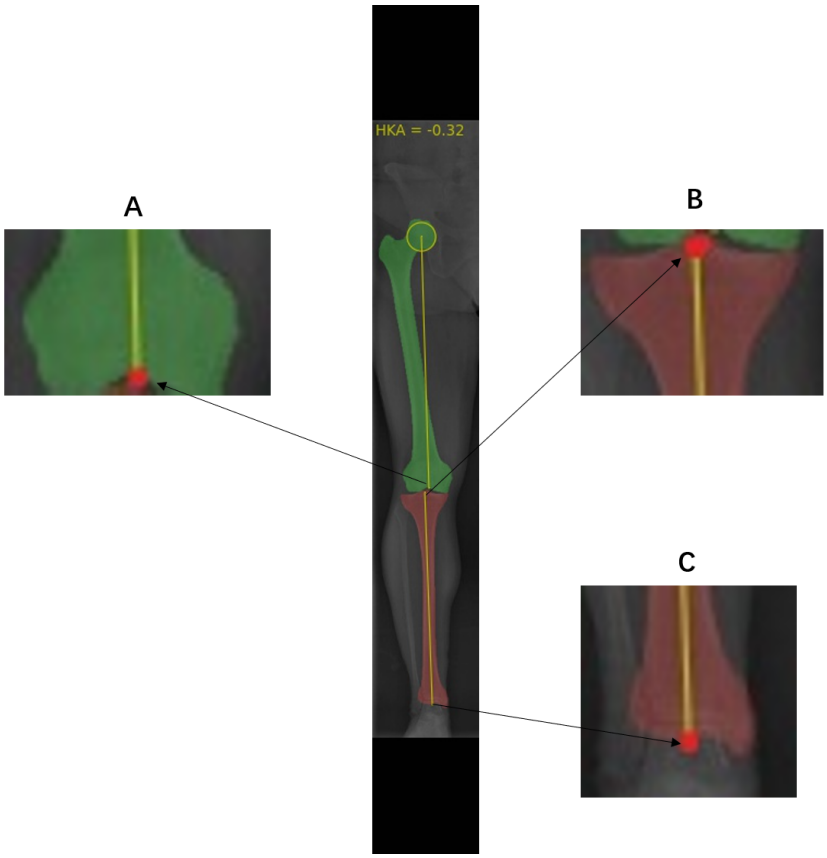


Figure 3.7: Illustration of the landmarks identified by the intersection between line and pixel. A: center of the knee, B: center of the tibia, C: center of the ankle.

Chapter 4

Results and Discussion

In this chapter, the results of the project are presented and discussed. In Section 4.1, the experiments on data augmentation methods, batch size, and loss functions are presented, followed by an analysis of the training results. The quantitive and qualitative results of automated alignment measurements are presented and discussed in Section 4.2.

4.1 Results of Training

Experiment on Data Augmentation Methods The data augmentations used in experiments include rotation, shifting, zooming, brightness change, and contrast change. The details of the data augmentation process are explained in Section 3.2. The training scheme was identical to the scheme presented in Section 3.4, and the dice loss was used to train the network. From the performance of the network on the validation set, the methods of brightness change together with contrast change yielded the best dice coefficient of 0.978. When spatial transformations, such as rotation, shifting, and zooming, are involved in data augmentation, it was observed that the network failed to converge and produced unstable dice coefficient and mIoU scores during the training. The augmentation methods of rotation, shifting, and zooming yielded the best dice coefficient of 0.933. In common, spatial transformations are often used in data augmentation in segmentation tasks to improve the network's performance. However, these methods deteriorated the performance of the network in the experiment. The unexpected training process may arise from the specificity of the segmentation task. In this task, the spatial location of the segmentation targets (femur and tibia) is relatively fixed. The extensive spatial transformations on the input radiographs can interfere with the training by presenting the segmentation targets in an unrealistic position. Therefore, only non-spatially transformed data augmentation methods are applied to the input radiographs in this project.

Experiment on Batch Size Compared to the previous works, this project uses a different batch size in network training. In previous works (Schock et al. (2020), Pei et al. (2021), and Shen et al. (2021)) on radiograph segmentation, the authors advocate the batch size of 1 in network training, but none of them gave a reason for choosing this batch size. The popularity of this batch size should have started with the original U-Net (Ronneberger et al. (2015a)). In the original work, the authors traded the size of the input image for the batch size. The authors believe that a high-resolution input will result in better segmentation. However, the size of the input radiograph is 128x1024 in this project. Therefore, the larger batch size can be chosen if graphics memory allows it. Due to the limited graphics memory, the batch sizes that can be chosen in experiments are 1, 4, and 8. Through experiments, it was found that batch size of 8 and 4 yielded significantly better performance on the validation set than the batch size of 1. The performance differences produced by the batch size of 8 and 4 were very close. But training was much faster with a batch size of 8. Therefore, the batch size of 8 was used for the training in the project.

Experiment on Loss Functions Cross entropy loss, dice loss, and focal loss are three common loss functions used in segmentation tasks. In the experiment, the network was trained with each of these three loss functions first. Then the network was trained with any combination of these three loss functions. The training scheme was identical to the scheme presented in Section 3.4. The best metrics achieved in the validation set were recorded for loss function rating. The best performance of the network with different loss function(s) is listed in Table 4.1.

Table 4.1: Network performance with different function(s). The network achieved the best performance on the validation set with the dice loss.

Loss Function	Dice Coefficient	mIoU	Epoch
Cross Entropy Loss	0.969	0.976	67
Dice Loss	**0.978**	**0.982**	52
Focal Loss	0.936	0.952	97
Cross Entropy Loss+Dice Loss	0.969	0.976	67
Cross Entropy Loss+Focal Loss	0.962	0.973	85
Dice Loss+Focal Loss	0.964	0.973	47
Cross Entropy Loss+Dice Loss+Focal Loss	0.968	0.975	47

What stands out in the table is dice loss. With dice loss, the network achieved the highest dice coefficient of 0.978 and mIoU of 0.982 at epoch 52. With cross entropy loss, the network performed second best and achieved a dice coefficient of 0.969 and mIoU of 0.976. With the focal loss, the network performed worst among loss functions, only achieving a dice coefficient of 0.936 and mIoU of 0.952. What is interesting about the data in this table is that the network with cross entropy loss only and with the combination of cross entropy loss and dice loss achieved identical metrics at identical

epoch. The reason for this can be that the scale of the cross entropy loss is larger than the scale of the dice loss when the network is trained with the combination of those two losses. The network is biased to decrease the cross entropy loss since it contributes more to the total loss. Therefore, The design of this experiment is deficient. Normalizing the losses to the same scale before combining them is necessary.

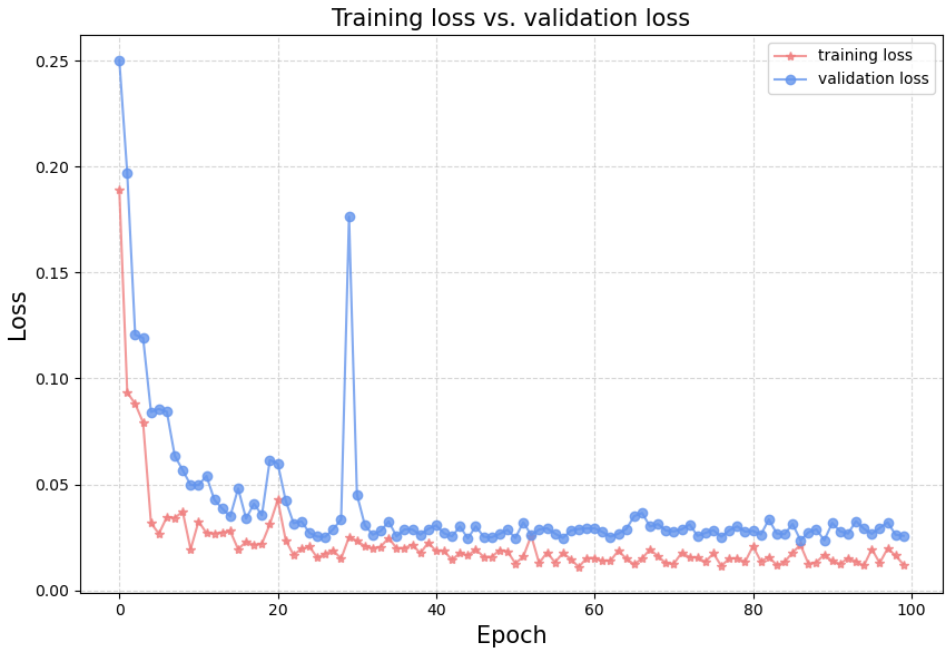
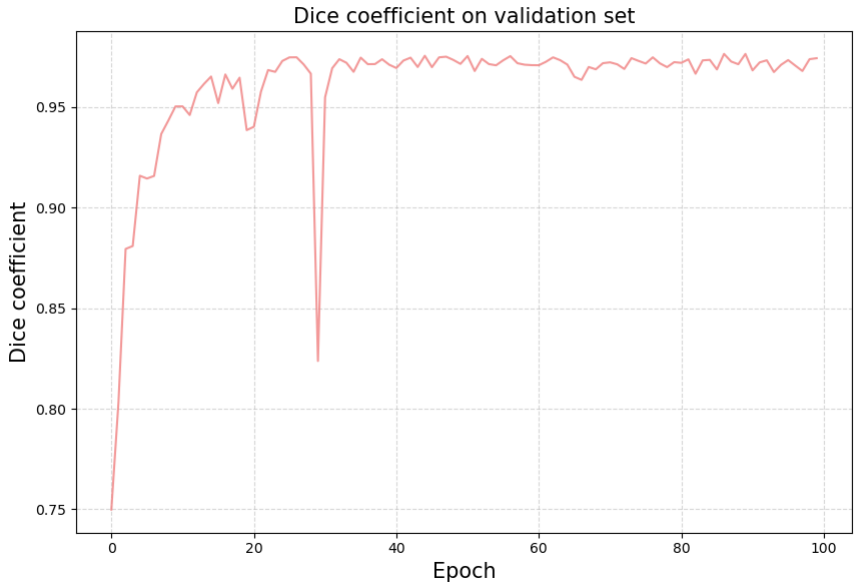


Figure 4.1: Training loss and validation loss. The line with star markers shows the training loss, and the line with dot markers shows the validation loss.

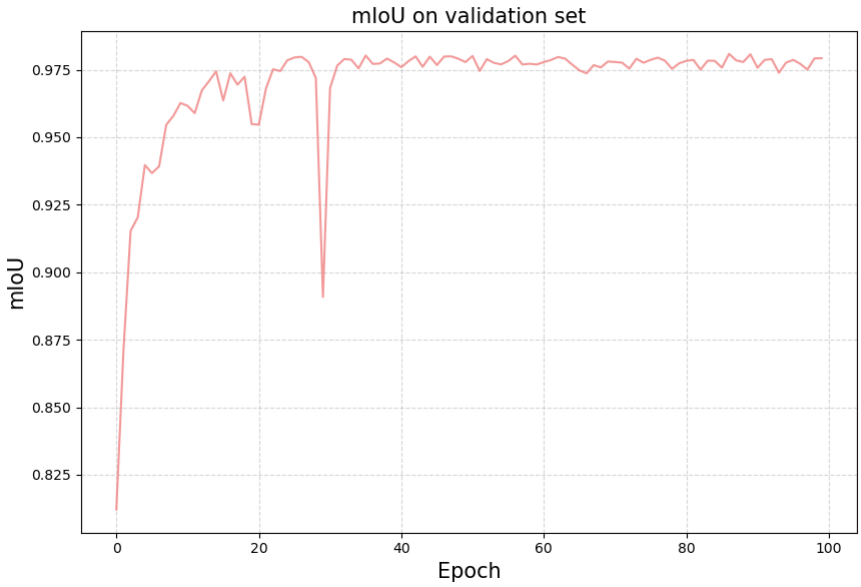
Training Results It has been shown that the best segmentation performance was achieved when the network is trained with dice loss. Therefore, this project used the dice loss function as the loss function in network training. The values of training loss and validation loss are presented in Figure 4.1. The figure compares the values of training loss and validation loss at different epoch. The line with star markers shows the changes of loss on the training set, and the line with dot markers shows the changes of loss on the validation set. Overall, training loss and validation loss rapidly decline in the first 10 epochs from 0.18 and 0.25 to 0.03 and 0.05 respectively, followed by a steady decrease to 0.02 and 0.04 respectively until epoch 22. Then the training loss fluctuates around 0.02 till the end of the training. On the contrary, the validation loss quickly rebounds to 0.17 at epoch 28 and drops back below 0.05 at epoch 30, followed by a fluctuation around 0.03 till the end of the training. Regarding the changes in metrics on the validation set, the Figure 4.2a shows the changes in the dice coefficient on the validation set during the training. The value of the dice coefficient starts at 0.75. Then it rapidly increased to

0.95 in the first 10 epochs, followed by a steady increment to 0.97 at epoch 25. Over the next 5 epochs, the value drops sharply to 0.83 and rapidly rebounds to 0.97. After that, the value fluctuates around 0.97 and reaches a peak of 0.978 at epoch 52. The changes of mIoU on the validation set share a similar pattern to the changes of dice coefficient as shown in Figure 4.2b. The value of mIoU starts at 0.81. Then it rapidly increases to 0.96 in the first 10 epochs, followed by a gradual increment to 0.98 at epoch 25. Over the next 5 epochs, the value steeply drops to 0.89 and rapidly rebounds to about 0.98. After that, the value fluctuates around 0.98 and reaches the highest value of 0.982 at epoch 52.

regarding the performance on the test set, the network achieved a dice coefficient of 0.972, and a mIoU of 0.980, which were slightly lower than the best values from the validation set. As for quality results of training. The first 10 data were sampled from the test set to show the quality of the segmentation. In Figure 4.3, the first row is the input radiographs, the second row is the ground truth, and the third row is the prediction. Overall, the network performs outstandingly on the segmentation tasks. A closer observation of the prediction results shows that there are small pits and leaks at the edge of the bones. But, these minor defects do not significantly affect the identification of the landmarks. The examples of predictions with more serious defects are shown in Figure 4.4. The incomplete femur head in Example B may impact the identification of the center of the femur head. The pixel leak in Example D may interfere with determining the femur position. The other examples also show pixel leaks or incomplete bone structures, but these defective areas will not be used as a reference for the identification of landmarks. Therefore, these defects are less likely to affect the identification of landmarks.



(a) Dice coefficient



(b) mIoU

Figure 4.2: Changes in the dice coefficient and mIoU on the validation set during the training. Both metrics converge around epoch 20. The value of the dice coefficient reaches a peak of 0.978 at epoch 52, and the value of mIoU reaches a peak of 0.982 at epoch 52.

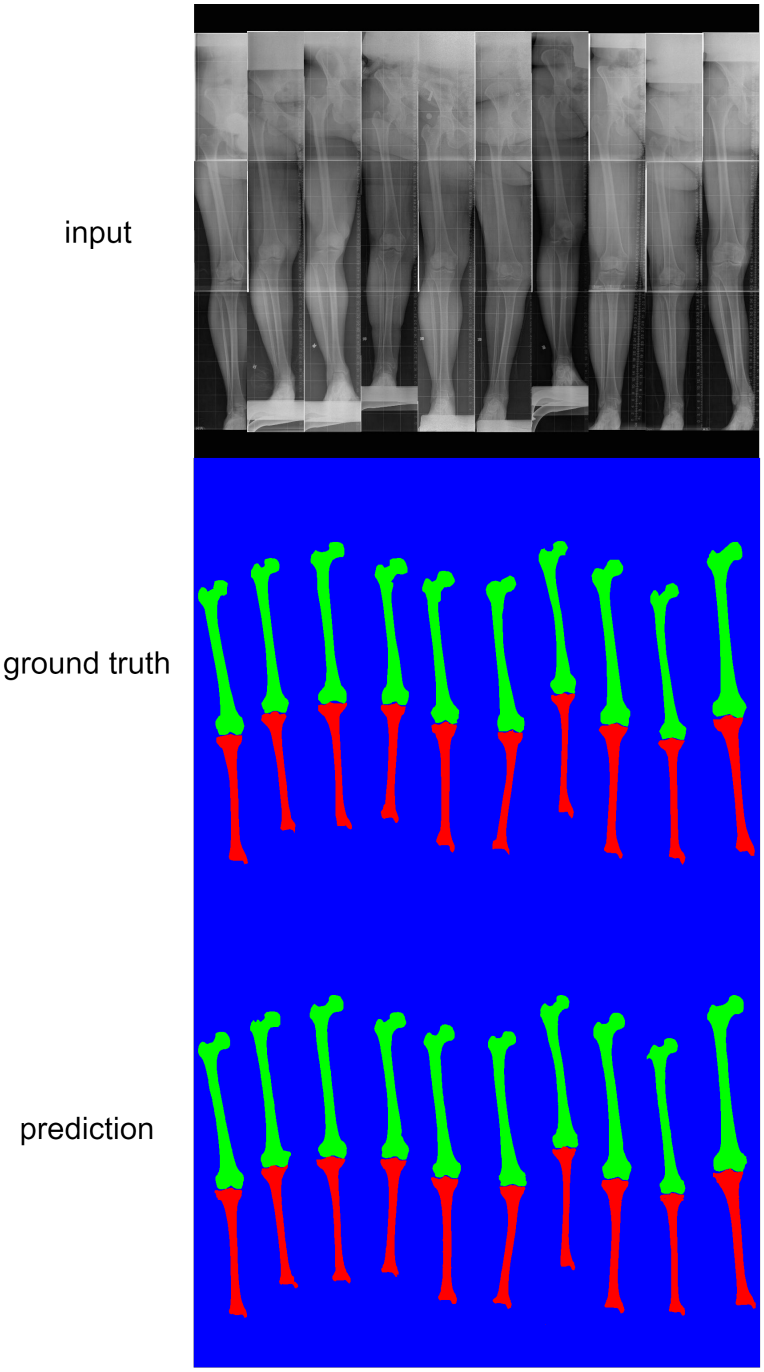


Figure 4.3: Comparative quality results from training. Ten radiographs and corresponding ground truth segmentations were sampled from the test set to present the quality results. The first row shows the input radiographs, the second row shows the ground truths of segmentation, and the third row shows the segmentation predictions.

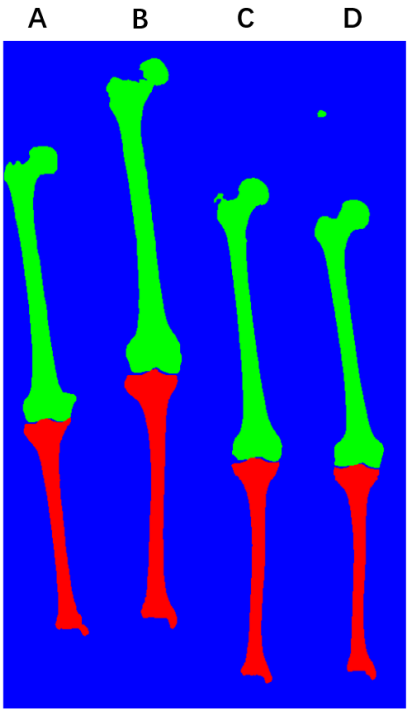


Figure 4.4: Examples of the inaccurate test set predictions. A: incomplete medial malleolus, B: incomplete femur head, C: pixel leak at the greater trochanter, D: pixel leak above the femur.

4.2 Results of Alignment Measurements

The HKA angle is used to assess the overall alignment. Therefore, To validate the alignment measurements by the proposed method, the predicted HKA angle is compared to the HKA angle manually measured in OAI project 60 (Sled et al. (2011)). The definition of the HKA angle is aligned with the definition used in OAI project 60. The HKA angle is defined as its angular deviation from 180°. As illustrated in Figure 2.3, when the knee alignment is neutral, the HKA angle is equal to 0 degree. The varus pattern of the knee alignment gives a negative HKA angle, and the valgus pattern of the knee alignment gives a positive HKA angle.

The OAI 36-month visit data were used to validate the alignment measurements. There were 341 long-leg radiographs in the dataset after data exclusion. The radiographs were preprocessed with the operations illustrated in Section 3.2. Each radiograph had a corresponding manually measured HKA angle as ground truth. In Figure 4.5, the distributions of the prediction and the ground truth are presented. The red triangles are the values of the HKA angle predicted by the proposed method, and the blue dots are the values of the HKA angle given by manual measurements. In general, there is not

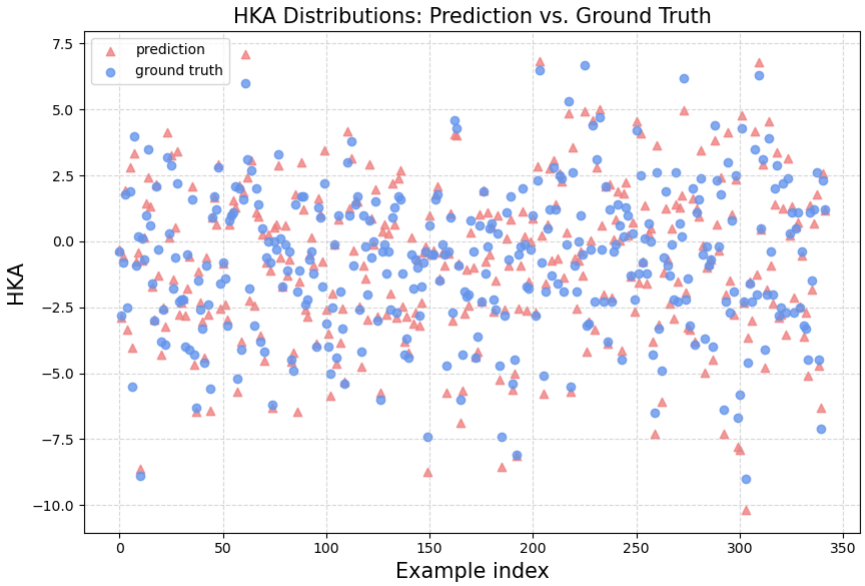


Figure 4.5: Distributions of predictions and ground truths. The red triangles show the HKA angle measured by the proposed method, and the blue dots show the HKA angle manually measured in Project 60.

a significant difference between the two distributions. To more accurately evaluate the alignment measurements, further quantitative and qualitative evaluations are required.

To quantitatively evaluate the results of alignment measurements, the mean signed error, the mean absolute error, the Bland-Altman analysis, and the paired t-test were used. A similar evaluation practice can also be found in Gielis et al. (2020) and Nguyen et al. (2020).

The quantitive evaluation of alignment measurements with the proposed method on the dataset revealed a high agreement with the manual measurements in Project 60. The Figure 4.6 shows the analysis of the signed error. The mean signed error is -0.03±0.78 degrees. In the figure, the distribution of the signed error is shown as blue dots. The black dashed line represents the mean level, and the two red dashed lines define the range of standard deviations. As shown in Figure 4.7, the mean absolute error is 0.58±0.52 degrees. In the figure, the distribution of the absolute error is displayed with blue dots. The black dashed line represents the mean level, and the two red dashed lines define the range of standard deviations. Except for a few outliers, most error values are below 1 degree. The Bland-Altman analysis as shown in Figure 4.8 presents agreement between the measurements by the proposed method and the measurements in Project 60. The black dashed line shows the mean difference of -0.03 degrees between the measurements given by the proposed method and the measurements in Project 60, with 95% limits of

Figure 4.6: Analysis of the signed error between the HKA angle measured by the proposed method and measured manually in Project 60. The mean signed error is -0.03 degrees, as shown with the black dashed line, and the standard deviation is 0.78 degrees. The red dashed lines show the levels of mean±standard deviation.

agreement (LoA) of ±1.53 degrees. The definition of the limits of agreement (LoA) is defined as the mean difference ±1.96 standard deviations. Through the paired t-test, the statistic is equal to -0.69, and the p-value is equal to 0.49. Therefore, there is an absence of evidence of a statistical difference between the measurements given by the proposed method (prediction) and the manual measurements in Project 60 (ground truth).

According to the absolute error of the measurements, the best and worst 5 measurements are presented as the qualitative results of the alignment measurements as shown in Figure 4.9. The left side of the figure presents the best 5 measurements in evaluation, and the right side presents the worst 5 measurements. From the sample results, it can be concluded that the segmentation quality directly affects the accuracy of the alignment measurements. In the set of worst measurements, the first example shows that the femur segmentation leaks in the pelvic area cause the algorithm to incorrectly locate the leaked pixels as the femur head. In the second and fifth examples, the leaks of femur segmentation in the lower tibia area cause the algorithm to incorrectly locate the center of the knee. The unsatisfied measurements in the third and fourth examples are caused by the missing pixels on the lower half tibia.

Through the analysis of the examples of worst measurements, it can be concluded that

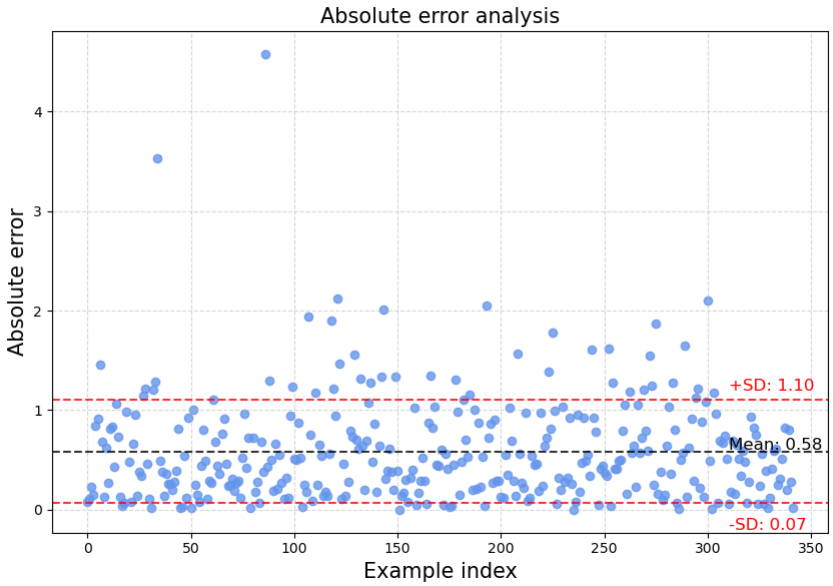


Figure 4.7: Analysis of the absolute error between the HKA angle measured by the proposed method and measured manually in Project 60. The mean absolute error is 0.58 degrees, as shown with the black dashed line, and the standard deviation is 0.52 degrees. The red dashed lines show the levels of mean±standard deviation.

the measurements with large errors mainly result in segmentation discontinuity. Further observation of the measurement results reveals that this discontinuity is more common in digital radiographs (Figure 3.1b). Radiographs in the dataset did not have a differentiating label between digital and scanned radiographs, and this did not allow a systematic sub-cohort analysis of the two groups. Therefore, the worst 5 and worst 10 measurements were sampled to support the observation. In the set of 5 worst measurements in Figure 4.9, there are 4 measurements on digital radiographs. In an extensive experiment with 10 worst measurements, 7 of the 10 were from measurements on digital radiographs. One of the reasons for this could be the domain gap between the dataset for training and the dataset for alignment measurement validation. In the dataset for training the network, there is only a very small proportion of digital radiographs. But, about half of the radiographs in the dataset of alignment measurements validation are digital. Therefore, compared to the scanned radiographs (Figure 3.1a), the network performs less well for digital ones. To experimentally address this problem, the morphological closing and opening operations are used to post-process the predicted segmentation masks. Post-processing shows a minor improvement in mean absolute error from 0.61 to 0.58 in mean absolute error. But the improvement is hardly observed in qualitative results.

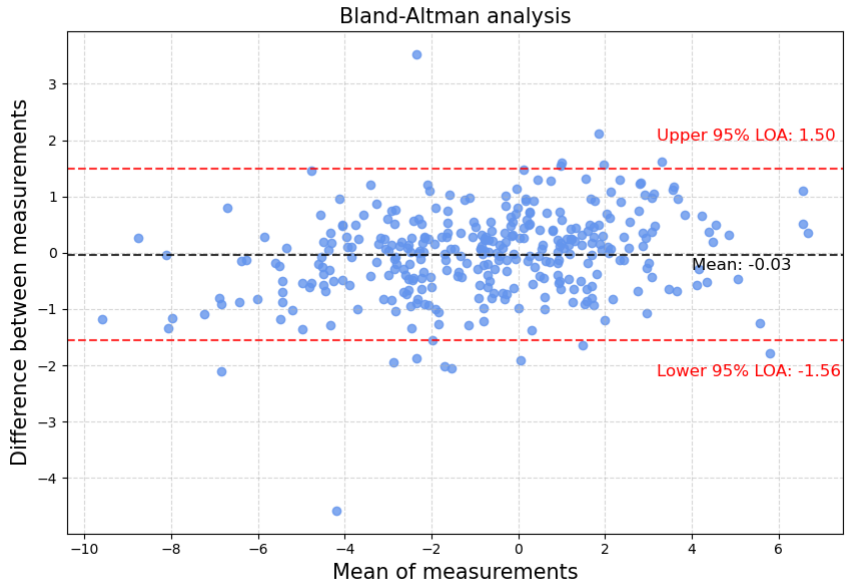


Figure 4.8: Bland-Altman analysis presents agreement between the HKA angle measured by the proposed method and measured manually in Project 60. The mean difference of -0.03 degrees is shown with the black dashed line. The red dashed lines show the upper and lower 95% LOA.

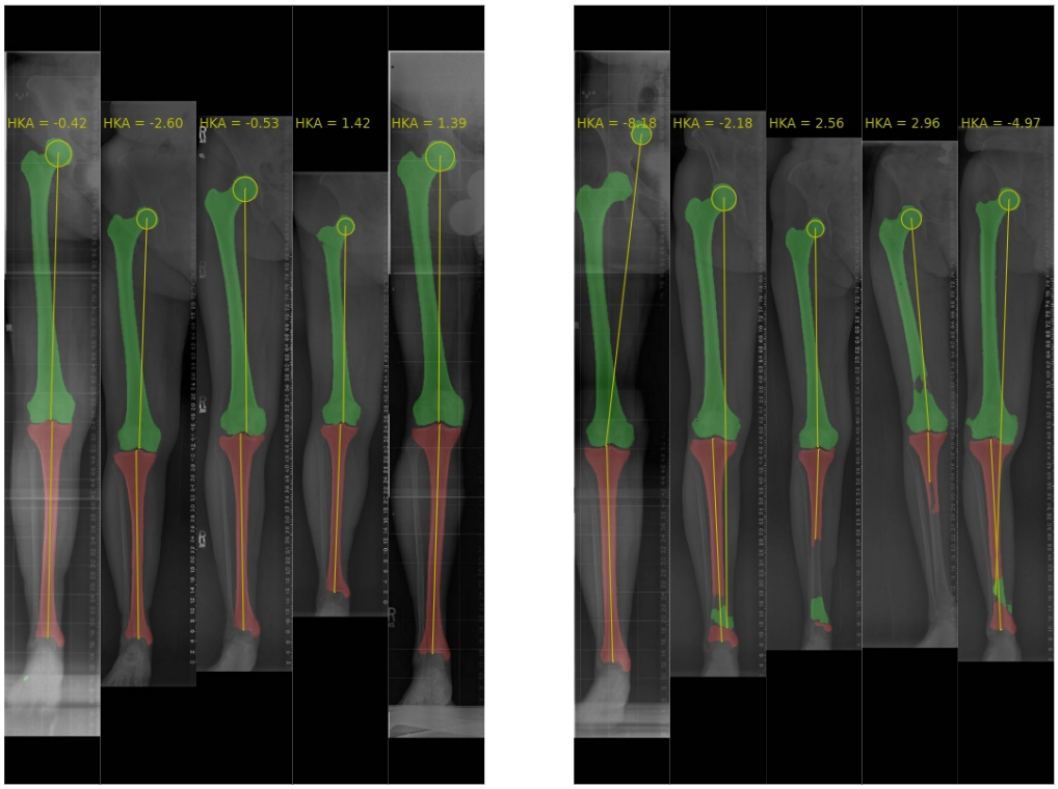


Figure 4.9: Best and worst measurements. The inaccurate measurements mainly result in segmentation discontinuity.

Chapter 5

Future Work and Conclusions

5.1 Future Work

The work in this project provides the foundation for automated bone segmentation and alignment measurements on long-leg radiographs. There are several possible extensions to this work in the future. First, the experiment on the loss function needs to be re-conducted. When experimenting with the combination of the losses, normalizing the losses to the same scale is necessary before training with them. Second, the reliable segmentation prediction on lower extremities in this project provides a platform that can support fast morphological analysis. Some other morphological characteristics, such as the joint line angle used in CPAK classification (MacDessi et al. (2021)), can be identified based on the segmentation results in this project. Last, there is still a requirement for future work to reduce the errors of alignment measurements even further. Since the accuracy of landmark identification and segmentation prediction are highly correlated, future work can improve the accuracy of landmark identification by strengthening the network's performance on the segmentation task. In terms of architecture, some modules, such as residual and attention modules, can be introduced into the network, thus enhancing the network's generalization ability. To improve the network's performance on digital radiographs, the training set needs to include more digital radiographs, so that the number of digital and scanned radiographs in the training set is balanced. A post-processing method could be applied to identify, exclude, or mitigate the results of grossly incorrect segmentation, such as segmentation discontinuity in the qualitative alignment measurement results. A set of constraints can be used to keep the landmark identification algorithm from being interfered with by outliers.

5.2 Conclusions

In this project, a two-step method was proposed to complete the task of automated knee alignment analysis on long-leg radiographs. First, the modified U-Net accurately segmented the femur and tibia from the plain radiograph. Then, the proposed landmarks identification method automatedly computed the HKA angle for overall alignment evaluation. As for the evaluation of network performance, the network achieved a 0.972 dice coefficient and 0.980 mIoU on the test set. As for the evaluation of alignment measurements, the HKA angle measured by the proposed method was compared to the HKA angle measured manually in Project 60. The proposed method achieved a signed error of -0.03±0.78 degrees, and a mean absolute error of 0.58±0.52 degrees compared to the manual measurements. The Bland-Altman analysis showed a mean difference of -0.03 degrees with 95% LoA of ±1.53 degrees. With the paired t-test, It could be found that there is an absence of evidence of a statistical difference between the measurements given by the proposed method and the manual measurements in Project 60.

Bibliography

Alom, M. Z.; Yakopcic, C.; Hasan, M.; Taha, T. M.; and Asari, V. K., 2019. Recurrent residual u-net for medical image segmentation. *Journal of Medical Imaging*, 6, 1 (2019), 014006. [Cited on page 9.]

Aradhya, V.; Mahmud, M.; Guru, D.; Agarwal, B.; and Kaiser, M. S., 2021. One-shot cluster-based approach for the detection of covid–19 from chest x–ray images. *Cognitive Computation*, 13, 4 (2021), 873–881. [Cited on page 8.]

Bellemans, J.; Colyn, W.; Vandenneucker, H.; and Victor, J., 2012. The chitranjan ranawat award: is neutral mechanical alignment normal for all patients?: the concept of constitutional varus. *Clinical Orthopaedics and Related Research®*, 470, 1 (2012), 45–53. [Cited on page 1.]

Besler, B. A.; Michalski, A. S.; Forkert, N. D.; and Boyd, S. K., 2017. Automatic full femur segmentation from computed tomography datasets using an atlas-based approach. In *International Workshop on Computational Methods and Clinical Applications in Musculoskeletal Imaging*, 120–132. Springer. [Cited on page 9.]

Bharodiya, A. K. and Gonsa, A. M., 2019. An improved segmentation algorithm for x-ray images based on adaptive thresholding classification. *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH*, 8, 09 (2019). [Cited on page 7.]

Borgefors, G., 1986. Distance transformations in digital images. *Computer vision, graphics, and image processing*, 34, 3 (1986), 344–371. [Cited on page 20.]

Boukala, N.; Favier, E.; Laget, B.; and Radeva, P., 2004. Active shape model based segmentation of bone structures in hip radiographs. In *2004 IEEE International Conference on Industrial Technology, 2004. IEEE ICIT'04.*, vol. 3, 1682–1687. IEEE. [Cited on page 8.]

Brouwer, G.; Tol, A. V.; Bergink, A.; Belo, J.; Bernsen, R.; Reijman, M.; Pols, H.; and Bierma-Zeinstra, S., 2007. Association between valgus and varus alignment and the development and progression of radiographic osteoarthritis of the knee. *Arthritis & rheumatism*, 56, 4 (2007), 1204–1211. [Cited on page 4.]

Candemir, S.; Jaeger, S.; Antani, S.; Bagci, U.; Folio, L. R.; Xu, Z.; and Thoma, G., 2016. Atlas-based rib-bone detection in chest x-rays. *Computerized Medical Imaging and Graphics*, 51 (2016), 32–39. [Cited on page 8.]

Carr, A. J.; Robertsson, O.; Graves, S.; Price, A. J.; Arden, N. K.; Judge, A.; and Beard, D. J., 2012. Knee replacement. *The Lancet*, 379, 9823 (2012), 1331–1340. [Cited on page 1.]

Chen, X.; Li, Y.; Yao, L.; Adeli, E.; and Zhang, Y., 2021. Generative adversarial u-net for domain-free medical image augmentation. *arXiv preprint arXiv:2101.04793*, (2021). [Cited on page 9.]

Chen, Y.; Ee, X.; Leow, W. K.; and Howe, T. S., 2005. Automatic extraction of femur contours from hip x-ray images. In *International Workshop on Computer Vision for Biomedical Image Applications*, 200–209. Springer. [Cited on page 8.]

Cooke, T.; Scudamore, R.; Bryant, J.; Sorbie, C.; Siu, D.; and Fisher, B., 1991. A quantitative approach to radiography of the lower limb. principles and applications. *The Journal of Bone and Joint Surgery. British volume*, 73, 5 (1991), 715–720. [Cited on page 1.]

Cooke, T. D. V.; Sled, E. A.; and Scudamore, R. A., 2007. Frontal plane knee alignment: a call for standardized measurement. *Journal of Rheumatology*, 34, 9 (2007), 1796. [Cited on pages 4, 5, and 6.]

Ding, L.; Zhao, K.; Zhang, X.; Wang, X.; and Zhang, J., 2019. A lightweight u-net architecture multi-scale convolutional network for pediatric hand bone segmentation in x-ray image. *IEEE Access*, 7 (2019), 68436–68445. [Cited on page 9.]

Feng, D., 2006. Segmentation of bone structures in x-ray images. *School of Computing National University of Singapore*, (2006). [Cited on page 8.]

Florea, L.; Florea, C.; Vertan, C.; and Sultana, A., 2011. Automatic tools for diagnosis support of total hip replacement follow-up. *Advances in Electrical and Computer Engineering*, 11, 4 (2011), 55–62. [Cited on page 8.]

Gielis, W. P.; Rayegan, H.; Arbabi, V.; Ahmadi Brooghani, S. Y.; Lindner, C.; Cootes, T. F.; de Jong, P. A.; Weinans, H.; and Custers, R. J., 2020. Predicting the mechanical hip–knee–ankle angle accurately from standard knee radiographs: a cross-validation experiment in 100 patients. *Acta orthopaedica*, 91, 6 (2020), 732–737. [Cited on pages 6 and 30.]

Goswami, B. and Misra, S. K., 2016. Analysis of various edge detection methods for x-ray images. In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, 2694–2699. IEEE. [Cited on page 8.]

Hadi, M.; Barlow, T.; Ahmed, I.; Dunbar, M.; McCulloch, P.; and Griffin, D., 2015. Does malalignment affect revision rate in total knee replacements: a systematic review of the literature. *Springerplus*, 4 (Dec. 2015), 835. doi: 10.1186/s40064-015-1604-4. [Cited on page 1.]

Ilahi, O. A.; Kadakia, N. R.; and Huo, M. H., 2001. Inter-and intraobserver variability of radiographic measurements of knee alignment. *The American journal of knee surgery*, 14, 4 (2001), 238–242. [Cited on pages 1 and 7.]

Insall, J. N.; Binazzi, R.; Soudry, M.; and Mestriner, L. A., 1985. Total knee arthroplasty. *Clinical orthopaedics and related research*, , 192 (1985), 13–22. [Cited on page 1.]

Jo, C.; Hwang, D.; Ko, S.; Yang, M. H.; Lee, M. C.; Han, H.-S.; and Ro, D. H., 2022. Deep learning-based landmark recognition and angle measurement of full-leg plain radiographs can be adopted to assess lower extremity alignment. *Knee Surgery, Sports Traumatology, Arthroscopy*, (2022), 1–10. [Cited on page 6.]

Kiran, M.; Ahmed, I.; Khan, N.; and Reddy, A. G., 2019. Chest x-ray segmentation using sauvola thresholding and gaussian derivatives responses. *Journal of Ambient Intelligence and Humanized Computing*, 10, 10 (2019), 4179–4195. [Cited on page 8.]

Kraus, V. B.; Vail, T. P.; Worrell, T.; and McDaniel, G., 2005. A comparative assessment of alignment angle of the knee by radiographic and physical examination methods. *Arthritis & Rheumatism*, 52, 6 (2005), 1730–1735. [Cited on page 4.]

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P., 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988. [Cited on page 18.]

Ma, J., 2020. Segmentation loss odyssey. *arXiv preprint arXiv:2005.13449*, (2020). [Cited on page 18.]

MacDessi, S. J.; Griffiths-Jones, W.; Harris, I. A.; Bellemans, J.; and Chen, D. B., 2021. Coronal plane alignment of the knee (cpak) classification: a new system for describing knee phenotypes. *The Bone & Joint Journal*, 103, 2 (2021), 329–337. [Cited on pages 1 and 35.]

Maesyaroh, U.; Munawaroh, L.; Sumarti, H.; and Adrial, R., 2021. Analysis of chest x-ray (cxr) images in covid-19 patients based on age using the otsu thresholding segmentation method. *Journal of Natural Sciences and Mathematics Research*, 7, 2 (2021), 59–65. [Cited on page 8.]

Milletari, F.; Navab, N.; and Ahmadi, S.-A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, 565–571. IEEE. [Cited on page 19.]

Nguyen, T. P.; Chae, D.-S.; Park, S.-J.; Kang, K.-Y.; Lee, W.-S.; and Yoon, J., 2020. Intelligent analysis of coronal alignment in lower limbs based on radiographic image with convolutional neural network. *Computers in biology and medicine*, 120 (2020), 103732. [Cited on pages 5 and 30.]

Oktay, O.; Schlemper, J.; Folgoc, L. L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N. Y.; Kainz, B.; et al., 2018. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, (2018). [Cited on page 9.]

Pei, Y.; Yang, W.; Wei, S.; Cai, R.; Li, J.; Guo, S.; Li, Q.; Wang, J.; and Li, X., 2021. Automated measurement of hip–knee–ankle angle on the unilateral lower limb x-rays using deep learning. *Physical and Engineering Sciences in Medicine*, 44, 1 (2021), 53–62. [Cited on pages 6, 9, and 24.]

Punn, N. S. and Agarwal, S., 2020. Inception u-net architecture for semantic segmentation to identify nuclei in microscopy cell images. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16, 1 (2020), 1–15. [Cited on page 9.]

Ray, C. and Sasmal, K., 2010. A new approach for clustering of x-ray images. *International Journal of Computer Science Issues (IJCSI)*, 7, 4 (2010), 22. [Cited on page 8.]

Ronneberger, O.; Fischer, P.; and Brox, T., 2015a. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv*, (May 2015). doi:10.48550/arXiv.1505.04597. [Cited on pages 2, 7, 16, 17, 18, and 24.]

Ronneberger, O.; Fischer, P.; and Brox, T., 2015b. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer. [Cited on page 9.]

Schmidt, G. L.; Altman, G. T.; Dougherty, J. T.; and DeMeo, P. J., 2004. Reproducibility and reliability of the anatomic axis of the lower extremity. *Journal of Knee Surgery*, 17, 3 (Jul. 2004), 141–143. https://pubmed.ncbi.nlm.nih.gov/15366268. [Cited on pages 1 and 7.]

Schock, J.; Truhn, D.; Abrar, D. B.; Merhof, D.; Conrad, S.; Post, M.; Mittelstrass, F.; Kuhl, C.; and Nebelung, S., 2020. Automated analysis of alignment in long-leg radiographs by using a fully automated support system based on artificial intelligence. *Radiology: Artificial Intelligence*, 3, 2 (2020), e200198. [Cited on pages 18 and 24.]

Seise, M.; McKenna, S. J.; Ricketts, I. W.; and Wigderowitz, C. A., 2005. Segmenting tibia and femur from knee x-ray images. In *Proc. of Medical Image Understanding and Analysis*, 103–106. [Cited on page 8.]

Sheehy, L.; Felson, D.; Zhang, Y.; Niu, J.; Lam, Y.-M.; Segal, N.; Lynch, J.; and Cooke, T. D. V., 2011. Does measurement of the anatomic axis consistently predict hip-knee-ankle angle (HKA) for knee alignment studies in osteoarthritis? Analysis of long limb radiographs from the multicenter osteoarthritis (MOST) study. *Osteoarthritis Cartilage*, 19, 1 (Jan. 2011), 58–64. doi:10.1016/j.joca.2010.09.011. [Cited on pages 1 and 2.]

Shen, W.; Xu, W.; Zhang, H.; Sun, Z.; Ma, J.; Ma, X.; Zhou, S.; Guo, S.; and Wang, Y., 2021. Automatic segmentation of the femur and tibia bones from x-ray images based on pure dilated residual u-net. *Inverse Problems & Imaging*, 15, 6 (2021), 1333. [Cited on pages 15, 18, and 24.]

Shu, Y.; Wu, X.; and Li, W., 2019. Lvc-net: Medical image segmentation with noisy label based on local visual cues. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 558–566. Springer. [Cited on page 9.]

Sled, E. A.; Sheehy, L. M.; Felson, D. T.; Costigan, P. A.; Lam, M.; and Cooke, T. D. V., 2011. Reliability of lower limb alignment measures using an established landmark-based method with a customized computer software program. *Rheumatology international*, 31, 1 (2011), 71–77. [Cited on pages 4, 11, and 29.]

Subburaj, K.; Ravi, B.; and Agarwal, M., 2010. Computer-aided methods for assessing lower limb deformities in orthopaedic surgery planning. *Computerized Medical Imaging and Graphics*, 34, 4 (2010), 277–288. [Cited on pages 3 and 4.]

Wada, K., 2018. labelme: Image polygonal annotation with python. https://github.com/wkentaro/labelme. [Cited on page 15.]

Wahyuningrum, R. T.; Purnama, I. K. E.; Verkerke, G. J.; van Ooijen, P. M.; and Purnomo, M. H., 2020. A novel method for determining the femoral-tibial angle of knee osteoarthritis on x-ray radiographs: data from the osteoarthritis initiative. *Heliyon*, 6, 8 (2020), e04433. [Cited on page 8.]

Yoshioka, Y.; Siu, D.; Cooke, T.; et al., 1987. The anatomy and functional axes of the femur. *J Bone Joint Surg Am*, 69, 6 (1987), 873–880. [Cited on page 4.]

Yoshioka, Y.; Siu, D. W.; Scudamore, R. A.; and Cooke, T. D. V., 1989. Tibial anatomy and functional axes. *Journal of orthopaedic research*, 7, 1 (1989), 132–137. [Cited on page 4.]

Zhou, Z.; Siddiquee, M.; Tajbakhsh, N.; and Liang, J. U. A nested u-net architecture for medical image segmentation. arxiv 2018. *arXiv preprint arXiv:1807.10165*. [Cited on page 9.]