ROBOTICS
AND
MECHATRONICS

DEEP LEARNING-BASED INTERPRETATION AND ANALYSIS OF ULTRASOUND RAW DATA

B. (Bangyu) Lan

MSC ASSIGNMENT

Committee:
prof. dr. ir. S. Stramigioli
dr. ir. K. Niu
prof. dr. ir. N.J.J. Verdonschot

May, 2024

012RaM2024
Robotics and Mechatronics
EEMCS
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

UNIVERSITY OF TWENTE. | TECHMED CENTRE
UNIVERSITY OF TWENTE. | DIGITAL SOCIETY INSTITUTE

DEEP LEARNING-BASED INTERPRETATION AND ANALYSIS OF ULTRASOUND RAW DATA

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Electrical Engineering
at University of Twente under the supervision of
Dr. Kenan Niu (Robotics and Mechatronics, University of Twente)

Bangyu Lan (s2933578)

May 13, 2024

Contents

Page

Abstract 4

1 Introduction 5

1.1 Research Questions 5

1.2 Thesis Outlines 6

2 Deep Learning-based Acoustic Measurement Approach 7

3 Improving Bone Tracking Precision using SIRC-UNet 15

4 Anatomical Region Perception Bone Tracking Methods 25

5 A Dual-Attention Framework to Decipher Muscle Dynamics 32

6 Discussion 43

6.1 Answers to the Questions 43

6.2 Limitation and future perspectives 44

7 Summary 46

Acknowledgements 47

List of Publications 48

Abstract

Animals excel at sensing the world around them, thereby making choices that best ensure their survival. The same is true for humans; from the moment we open our eyes each day, we are constantly receiving and processing the world's information, making decisions based on what we gather. Sensing the environment effectively is the starting of intelligence, allowing an entity or a robot to independently initiate actions. In this thesis, developing an effective algorithm to perceive and interpret the surrounding basic signals is the main topic.

The most basic signal widespread in the nature is the one-dimensional signal. From sound waves, electric and biological signals, to the earthquake signals and the temperature, one-dimensional signal represents the variations of a single parameter over time, which is a fundamental aspect of many natural and human-made processes. The typical one-dimensional signal in medical domain is the ultrasound (US), which is widely used for the non-invasive diagnosis. Among all US techniques, the A-mode type receives less attention because the one-dimensional raw signal is difficult for human to interpret. However, besides B-mode's non-invasive advantage, A-mode US is smaller, more convenient, and easier to use, showing the potential to not only be installed on the robotics system for the intelligent perception, but also having the practical values for patients' daily portable and wearable care, reflecting disease status in daily time.

In this master thesis, the ultrasound raw data will be revisited and explored using deep learning to unveil the unique features useful for medical and robotics applications. The exploration started from bone detection to muscle activity monitoring by interpreting raw US signals. The results showed high accuracy thanks to the universality, generalizability and robustness of the proposed deep-learning approaches, which is also inspiring for the intelligent robotics perception. More attempts were made to define the weakness and scope of the technical performance, providing a clearer vision for the broader application in more robotics domains of future developments.

1 Introduction

1.1 Research Questions

Nowadays, although different types of artificial intelligence (AI) theories, robotics systems and applications have been built and developed, a truly human-like robotics that can perceive the world for its own intelligence to naturally interact with the environment is still lacking. The main reason is that the whole pipeline of a life form, from the intelligence start to the end-effector movement, is still not clear and very difficult to transform to the real robotics system. If such an intelligent robotics can be built, it should first have the efficient perception capability to grasp the surrounding signals and information. The most basic signal widespread in the nature is in one-dimensional, such as sound waves, electronic signals, and bio-signals. This is because the time works as the independent variable for almost all the dynamic systems. Its pervasive nature offers a direct and efficient tool to comprehend the dynamics of physical and biological systems. Thus, this type of singular dimension forms the core of how we understand and interact with the surrounding world. Similarly, the better understanding and perception of the one-dimensional signals enable robots to grasp all sorts of information to facilitate their decisions and actions.

In the medical domain, one typical one-dimensional signal is the ultrasound (US), which was used widely as a non-invasive and safe diagnosis device. Although the B-mode US and its two-dimensional results visualization are widely used currently, the B-mode US results actually could be regarded as the combination of many A-mode ultrasound's results. The better understanding and intelligent processing of A-mode US raw data not only decrease the B-mode US cost and the device size, but also help to integrate the small size A-mode probe on the robotics arm or exoskeleton for more efficient surgery navigation or daily biometric data tracking of patients. Therefore, processing and interpreting the A-mode US signals intelligently not only have its value in the biomedical and robotics domains, but also inspire people to build a truly intelligent robotics that can process and interpret the similar types of one-dimensional signals in the surroundings.

To analyze and interpret the A-mode US raw signals from the intelligent robotics perspective, several questions need to be answered:

(1) Can this signal reflect the actual positions and track the movement of the subjects?

(2) What is the range and scope of this perception when it has a high accuracy, and what components in this method really take effects?

(3) When doing the (1) and (2), can it still classify and recognize different types of signals so that it can have a grand view of the subjects for a complete perception?

(4) Can this capability of interpreting one-dimensional signal be integrated with other forms of signals for a more complex but useful system development?

In the following, each chapter try to answer one question by solving the challenges in the medical scenarios using A-mode ultrasound. Although each chapter in the thesis used different background, the whole master thesis tries to develop the algorithm that can empower robots with the perception of the grand-view of the subjects only by one-dimensional signals.

1.2 Thesis Outlines

In chapter 2, the problem was first proposed and a preliminary deep-learning based structure was suggested. The constructed CasAtt-UNet was used to track the movement of femur & tibia in specific positions and showed sub-millimeter accuracy, surpassing previous methods in both accuracy and operation complexity, which achieved the intelligent perception without additional manual works or pre-measurement.

Based on it, the chapter 3 define the method and algorithms in a more formally and mathematically way, and explore the one-dimensional signal interpretation in a wider area instead of specific bone positions. The ablation study was done to verify the designed structures and losses in different perception bone areas.

After the problem and the approach being well-defined, the chapter 4 continuously simplify the model and enable it to recognize the signals from different channels and positions, which increase the independence of the method and make it more suitable to be integrated in a real-time robotics systems for the medical and navigation usages.

In chapter 5, the one-dimensional signal interpretation algorithm has been transformed to another type of signal (surface electromyography signals). The combination of A-mode ultrasound and the surface electromyography (sEMG) enables the robot's capability to predict mechanical movement using merely the energy information, achieving the conversion between the energy and momentum. This paves the way for the information conversion using merely the one-dimensional signal, showing the great potential perception capability for the intelligent robotics.

Through these explorations, this master thesis not only demonstrates the enhanced capabilities of A-mode ultrasound in medical imaging and robotic applications, but also paves the way for its integration into more accessible and efficient healthcare solutions. The developed deep learning algorithm give new directions on how to interpret and utilize A-mode ultrasound or other one-dimensional raw signals, expanding the applications into the fields of daily medical practice and personalized health monitoring. In addition, the generalization and adaptation of the algorithm can empower the robotics to be aware of the surroundings and make reasonable decision.

2 Deep Learning-based Acoustic Measurement Approach

In orthopedic surgery, the operation precision and convenience for the bone measurement is critical, especially in the Total Knee Replacement Arthroplasty (TKA). The traditional methods involves optical tracking systems and radioactive imaging, which have the invasive limitation and required the extended preparation times. This chapter introduces a novel deep learning model that used the A-mode US to enhance the accuracy, safety and convenience of the bone tracking. By training with the labeled data from the cadaver experiments, the method achieves sub-millimeter precision in the bone positioning, offering a safer and efficient alternative for the surgical navigation.

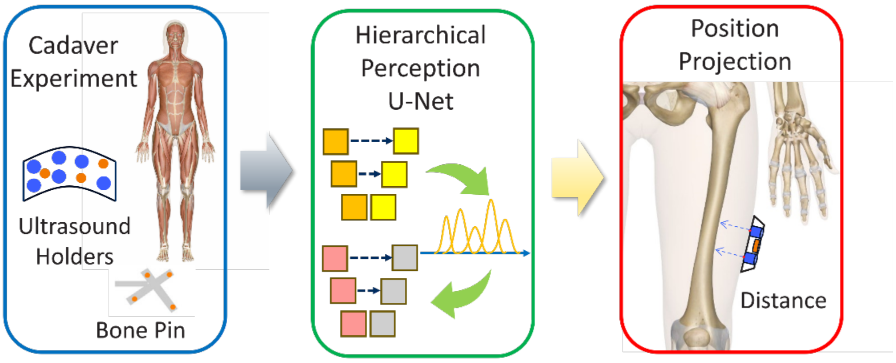


Figure 1: Graphical Abstraction of the Chapter 2

Deep Learning based acoustic measurement approach for robotic applications on orthopedics

Bangyu Lan[1], Momen Abayazid[1], Nico Verdonschot[1,2], Stefano Stramigioli[1] and Kenan Niu[1]

Abstract— **In Total Knee Replacement Arthroplasty (TKA), surgical robotics can provide image-guided navigation to fit implants with high precision. Its tracking approach highly relies on inserting bone pins into the bones tracked by the optical tracking system. This is normally done by invasive, radiative manners (implantable markers and CT scans), which introduce unnecessary trauma and prolong the preparation time for patients. To tackle this issue, ultrasound-based bone tracking could offer an alternative. In this study, we proposed a novel deep-learning structure to improve the accuracy of bone tracking by an A-mode ultrasound (US). We first obtained a set of ultrasound dataset from the cadaver experiment, where the ground truth locations of bones were calculated using bone pins. These data were used to train the proposed CasAtt-UNet to predict bone location automatically and robustly. The ground truth bone locations and those locations of US were recorded simultaneously. Therefore, we could label bone peaks in the raw US signals. As a result, our method achieved sub-millimeter precision across all eight bone areas with the only exception of one channel in the ankle. This method enables the robust measurement of lower extremity bone positions from 1D raw ultrasound signals. It shows great potential to apply A-mode ultrasound in orthopedic surgery from safe, convenient, and efficient perspectives.**

I. INTRODUCTION

Measuring bone position is important to understand the positions and kinematics of the lower extremity, for example, in robotic total knee replacement arthroplasty [1] and wearable exoskeleton [2]. However, traditional measurements, such as CT scan, skin markers, or implantable markers, unexpectedly introduce accumulated distance errors, unnecessary trauma, radiation risks, and infections.

To precisely measure distance without unnecessary trauma, previous studies used data from multiple sensors. [3] worked on measuring and evaluating fingertip distances using optical sensors and ultrasound probes. [4] regarded that merely using ultrasound was difficult to predict movements of the lower extremities; instead, they used data from both EEG and sEMG recordings. A-mode ultrasound has recently been proposed for bone model reconstruction [5], [6], registration and surgery robotics [7], [8], [9], as it is easy to deploy and feasible to collect data from different bone locations via multiple channels simultaneously. Reconstruction and registration of bony surface can be carried out using these data in different bone locations.

For example, ultrasound was studied to apply in lower extremity motion tracking and bone measurement for surgical robotics, which was a novel convenient and noninvasive method [10], [11], [12], [13]. However, in these approaches, distance measurement was mostly based on expert knowledge, by knowing the approximate range of bone peak's locations and the general shapes of the peaks, which purely based on experience and was lacked of robustness and generalization, as a little disturbance brought by processing or measurement noise could easily change profiles of the peak, making it difficult to identify. In our approach, we used the generalized and adaptability of deep learning to automatically measure bone reflection peaks in US signals without additional knowledge, making ultrasound-based measurements more applicable and automatic. In addition, it helped to ease the difficulty of deployment in total knee replacement arthroplasty and other surgical robot applications.

In previous studies, to analyze 1D medical signals using deep learning, the U-Net structure was used to recognize and identify ECG peaks to diagnose heart disease [14] by its contextual information preservation and feature localization capability. The different resolution perceptual fields can capture different sizes of peak profiles. However, it was difficult to recognize the various reflection peaks in the US, as the reflection peaks are more sparse, random, and have diverse profiles in different channels. To solve the issue, we exploited UNet's localization pattern by using a cascade U-Net structure for different perception resolutions, connected with a novel sampling-based proposal mechanism. In addition, an attention framework was introduced to filter out features that are irrelevant to the target peak range [15]. This helped to continually improve the perception of the peak profile. Through these designs, our CasAtt-UNet (**Cas**caded **Att**ention **UNet**) could easily recognize the sparse and effective peak profiles in the US signals with high accuracy. In addition, since the proposed CasAtt-UNet could infer signal peaks and calculate bone location efficiently, it could achieve real-time bone measurement in the surgical robot application, e.g. total knee replacement arthroplasty.

To the best of our knowledge, few studies focused on analyzing and interpreting the 1D ultrasound signal to detect bones using deep learning method. In this paper, we aim to introduce a deep learning-based 1D raw signal peak locating approach for accurate bone position measurement. The proposed method showed great potential for using deep learning to analyze complex and random ultrasound signals. For future development, this can be deployed on surgical robot arms to guide precise total knee replacement arthroplasty. It is worth noting that the proposed technique could also hold the potential for adaptation to other robotics surgery where bone motion tracking is essential. This is also beneficial to

[1]Robotics and Mechatronics, University of Twente, Enschede, AE, The Netherlands, [2]Orthopaedic Research Lab, Radboud University Medical Center, Nijmegen, the Netherlands

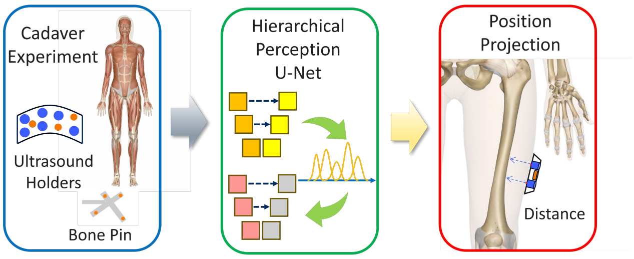


Fig. 1. Pipeline overview: Our method had three steps (from left to right): performed cadaver experiment to collect ultrasound signals **(network input)** and calculate bone positions (**network output**), Use the dataset to train CasAtt-UNet, recover the bone position and evaluate.

improve the accuracy of bone kinematics measurement.

II. MATERIAL AND METHOD

Our method contained three parts in Fig. 1: Data collection from a cadaver experiment, CasAtt-UNet training, and validation of the inference result. The US data and positions of the optical markers were collected from a full-body cadaver specimen, where bone pins were inserted into the femur and tibia for the reference locations of the bones.

Firstly, ultrasound holders were attached to the cadaver leg and tracked using the 3D optical tracking system. Each holder has multiple US transducers to collect US reflection waves in one anatomical locations on the femur or tibia. Thanks to the bone pins, the ground truth position of the two bones could be recalculated, and the reference positions of the attached US holders were also recorded concurrently. The precise bones locations were then calculated with respect to the positions of the attached US holders (i.e., that of each ultrasound probe). Subsequently, the reference distances between each ultrasound transducer and the underlying bone surface could be derived. Because bone locations and bone reflection peaks in the raw 1D ultrasound signal are correlated, the ultrasound signal with bone locations was used to train our CasAtt-UNet. In the end, the precision at different bone locations was evaluated, and the performance of our model was reported.

A. Experimental Setup and Data Acquisition

To ensure the accuracy and functioning of our method, a human cadaver specimen (male, 79kg, 179cm) was used to acquire dataset. This has been approved by Radboud University Medical Center (Radboud UMC), Nijmegen, the Netherlands. A full leg (from pelvic to foot) was CT scanned (TOSHIBA Aquilion ONE, voxel size of $0.755mm \times 0.755mm \times 0.500mm$). Subsequently, 3D geometric models of the femur and tibia were segmented using Mimics 17.0 (Materialise N.V., Leuven, Belgium).

This dataset was collected in our previous study [12]. During the experiment, each ultrasound holder contained three LED optical markers and several 7.5MHz A-mode ultrasound transducers (Imasonic SAS, Vorayl'Ognon, France), which were used to acquire ultrasound echos. There were in total 30 A-mode ultrasound transducers and 18 LED optical markers distributed on the 6 ultrasound holders (holders were designed from CAD models). Also there were 16 LED optical markers distributed on the four bone pins. The sample rate of the entire US tracking system was 20 Hz. An optical tracking system (Visualeyez VZ4000v trackers, PTI Phoenix Technologies Inc., Vancouver, Canada) was operated at 100 Hz to track the 3D locations of the US probes. The ultrasound signal was acquired and synchronized with the optical tracking system in the Diagnostic Sonar FI Toolbox (Diagnostic Sonar Ltd., Livingston, Scotland). The origin and direction of the ultrasound beam were determined from the calibration method [11].

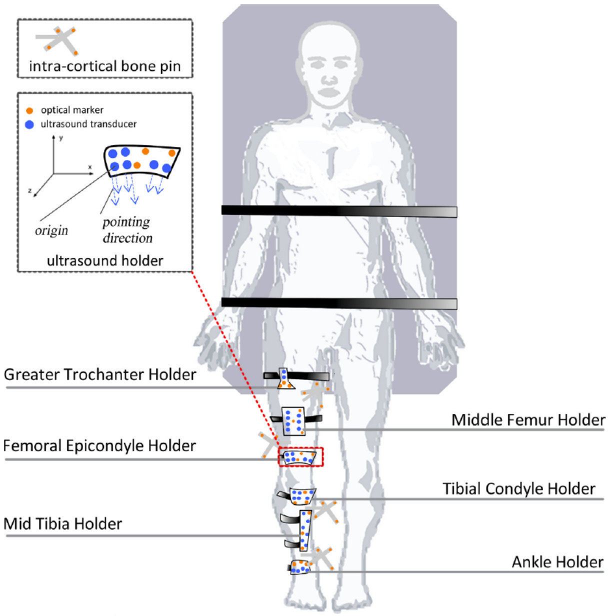


Fig. 2. Location of the US holders[12]: Our ultrasound holders were installed on the six locations of the left leg: Trochanter, Mid Tibia, Femur Epicondyle, Tibia Epicondyle, Mid Tibia and Ankle. Each holder was tied using bandages. Notice that the distribution of optical markers and transducers were different with the ones in the image.

To record the location of the bones, four bone pins were inserted into different parts of the femur and tibia, and six ultrasound holders were fixed at different locations of the femur and tibia. Throughout the experiment to collect the dataset, the leg was actively maneuvered through a cyclic flexion and extension process to emulate the swing phase of the gait cycle, which simulated the condition of a person's walking. This caused changes in distance between bones and transducers, resulting in changes of peak locations in ultrasound signals.

After data collection, we gathered US signals and trajectories of the attached optical markers from the holders. These data were used to reconstruct the actual locations of US holders above bones and the directions of US waves. Consequently, the actual bone depth can be derived using the bone location, the origin of US waveform, and the US wave directions.

In total, a dataset was acquired that contained 1017 con-

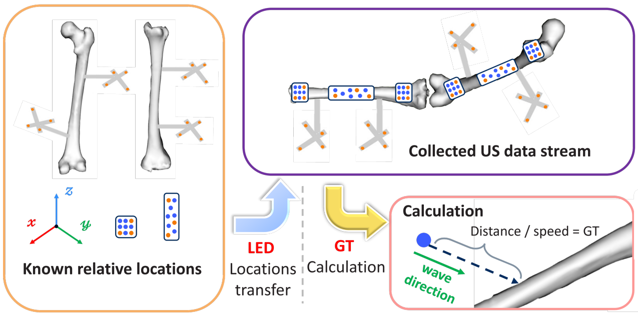


Fig. 3. Steps to determine ground truth labels: The optical markers in the predefined and the experiment case determined the transformation, which was used to transform US transducers and waves directions to the experiment coordinate frame. The label was calculated using Euclidean distance and the speed of US.

tinuous samples (moments) recorded from leg movement. In each sample, there were ultrasound echos from 30 transducer channels and 3D positions of 34 optical markers (16 bone pin markers and 18 US holder markers). We noticed that for most bone peaks in the US signals from Trochanter and Mid Tibia, they have been attenuated or disappeared. The missing number had already exceeded half of all acquired samples, which was clearly not suitable for network training. Therefore, the dataset in these two locations was directly discarded. The rest dataset was checked and screened too. Finally, the 1D ultrasound signals in four anatomical areas (Fig. 2) were collected: Femur Epicondyle, Tibia Epicondyle, Mid Tibia and Ankle. The holder at each location contained three optical markers and several transducers. Totally nine channels in the four anatomical locations with the trajectories of twelve LED optical markers were suitable for training the CasAtt-UNet.

B. Bone Location Calculation and US Signal Peak Labeling

After cadaver experiment, we labeled the ground-truth position of bone peaks in the 1D ultrasound signals. What we had were the following: 1, 3D geometric surface of femur & tibia and 3D positions of bone pins; 2, 3D positions and distributions of the US transducers and optical markers in the holders; 3, 3D positions of optical markers (bone pins and holders) in the experiment. The processing procedure was to transfer transducers, femur and tibia from the predefined coordinate frame (e.g. CT frame and CAD frames) to the experiment coordinate frame (i.e. the coordinate frame of the cadaver specimen in the experiment) in each moment. The process was shown in Fig. 3. Firstly the transformation ${}^R_H T$ that align optical markers $\{H\}$ in the predefined model with the ones in the experiment $\{R\}$ was calculated, which was $f(\{H\},\{R\})$. Then this matrix ${}^R_H T$ was used to transfer US transducers positions $\{P_H\}$ to the ones in the experiment $\{P_R\}$, which was ${}^R_H T \cdot p_H$. The calculation also kept the wave directions unchanged. The same process was done for both femur and tibia bones, where 16 optical markers on four bone pins were used to transform. The calculation was as following. This transformation process applied to all US transducers positions.

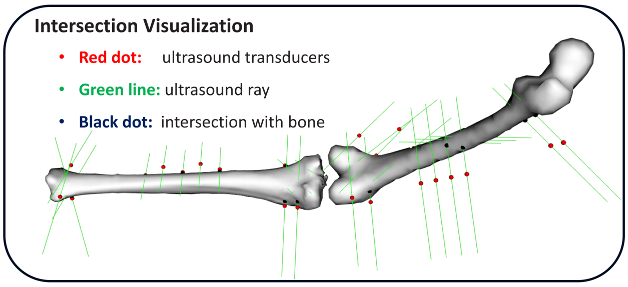

Intersection Visualization
• Red dot: ultrasound transducers
• Green line: ultrasound ray
• Black dot: intersection with bone

Fig. 4. Intersection between bones and US waves: This showed one moment in the experiment that the US waves intersected with the bony surface. The intersection positions produced ground truth locations. Red dots were transducer probes positions. Black dots were intersection positions. Green lines were waves directions. The ground truth distance (later used for labeling) was the line segments between black dots and red dots. For some waves there was no black dot as there was no intersection.

$$ {}^R_H T = f(\{H\},\{R\}) \tag{1} $$

$$ p_R = {}^R_H T \cdot p_H \tag{2} $$

After obtaining all relative positions and wave directions, we could render the US waves starting from the transducers and ending on the bony surface. The distance between intersection positions and the transducer probes were calculated and shown in Fig. 4. After knowing the 3D positions of transducer probes and the intersections, we could calculate the Euclidean distance $d(mm)$ between them (Green line segments cropped by red and black dots). To annotate the corresponding reflected peak position in 1D ultrasound signals, the bone peak location was calculated using Equation (3), represented as the index of units (idx). Here we assumed the ultrasound speed under skin was $v = 1540m/s$ from [16], the sampling rate was $f_s = 40$MS.

$$ idx = \frac{d}{(2 \times \frac{v}{f_s} \times 1000)} \tag{3} $$

, where $2 \times \frac{v}{f_s} \times 1000$ was the one unit length along the 6760 length of the 1D raw ultrasound signal. From our calculation, one unit length in the signal is equal to 0.01925 mm in the distance measurement, which is the minimum accuracy level.

After getting the depth point in the signal, the location in the 1D signal could be annotated as the training label. Using the raw US signals and labels, we could train our CasAtt-UNet.

D. Overview of Cascade Attention U-Net

The proposed CasAtt-Unet was shown in Fig. 5. It composed of coarse attention U-Net, sampling-based proposal, and refined attention U-Net. As the raw US signal occupied 130mm while the original peak annotation was only one index interval (1 index = 0.01925mm), a hierarchical structure was required for narrowing down the range. To do this, a coarse attention U-Net was first used to determine the existence of bone peak and capture the approximate range

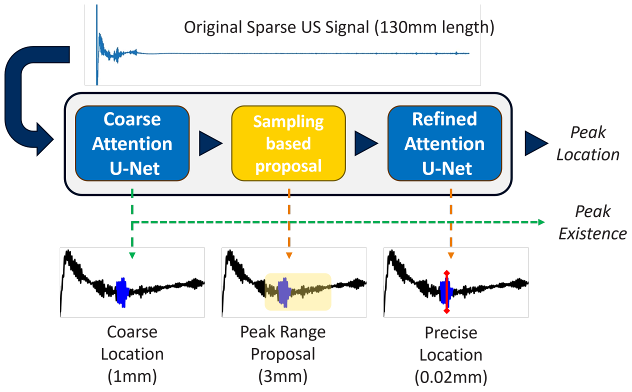


Fig. 5. Pipeline of Algorithm: Given a 130mm US raw signal, the first coarse U-Net captured the approximate range (1mm) of bone peak. The output was used to segment a continuous length (3mm) of signal region, which is the input of the refined U-Net to determine the exact position of the bone peak.

(1mm). From this region, a novel sampling-based mechanism proposed the most likely region of the bone peak. Based on the proposal, the refined U-Net predicted the exact peak location.

E. Structure of Attention U-Net

The proposed coarse and refined attention U-Net was shown in Fig. 6, which was inspired from [14]. Compared with the normal U-Net in image domain, the convolution kernels had been replaced from two dimension to one dimension. In addition, an attention block similar to [17] was inserted in every skip connection between two sides of U-Net. The output from the left side worked as the input feature of the attention block, filtered by the attention signal from the deeper layers, as the features from deeper layers contained abstract and general space knowledge that can filter out large irrelevant areas in the signal. In this way, the CasAtt-Unet performed more robustly when dealing with more complex and random signals. During training, because the peak summit was only one unit length (0.01925 *mm*) along the whole range, to facilitate network training, the range of peak was increased to 5 units (about 0.1 *mm*) for the Refined location. For the coarse UNet, the coarse range has been increased to 50 units (approx. 1 *mm*). The network output had the whole signal length consisting of bone peak probability in each unit. This would be threshold by 0.5 probability to the bone peak segment. To recover the exact peak position (i.e., bone depth), we used the middle point of this segment to calculate the depth distance by multiplying the middle point unit index number with the unit interval length.

F. Mechanism of Sampling-based Proposal

To have a candidate region from the coarse U-Net output, a sampling-based proposal method was required and shown in Fig. 7. A standard Gaussian distribution (mean=1.0, std=1.0) curve, whose amplitude was determined by the predicted probability, was built on each unit of the whole signal. They were combined and summed to have a mixture possibility density curve. Based on this PDF, a continuous fixed-length region was probabilistic sampled. The start and end location cropped the original US signal to be the input of refined attention U-Net, which continuously identified the bone peak in this local region. The benefit of probabilistic approach instead of the deterministic one was that: Each time the refined U-Net can be confronted with a random region even if the coarse U-Net output was the same, the refined U-Net was forced to learn the meaningful peak profile instead of memorizing peak position. This enabled CasAtt-UNet's robust detection when only trained using a limited dataset. During inference, instead of sampling from the PDF, the peak position was directly decided by the largest probability position in this region.

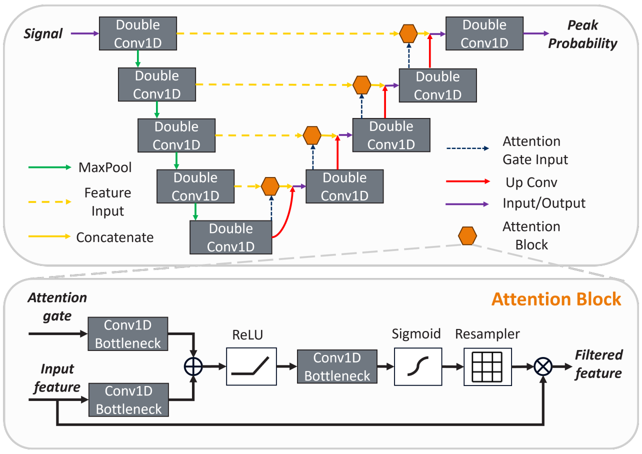


Fig. 6. Attention U-Net Structure: We replaced the 2D convolution in normal U-Net to 1D convolution, and added additional attention block to increase the perception ability. In each orange hexagon connection (attention block), the attention signal was from the deeper layer, while the input feature was from the left side of U-Net. The output was concatenated with the up convolution result of the deeper layer and input to the double conv1D block.

G. Training and Evaluation

To train a dataset with highly unbalanced foreground and background labels, we introduced dice loss and cross-entropy loss simultaneously. The dice loss was written in Equation (4). It is defined as one minus dice coefficient, which is widely used in medical image segmentation: The numerator is defined as twice of intersection between ground truth and prediction, while the denominator is defined as the total probability of prediction and ground truth. The ε can stabilize the training and prevent zero division. For cross-entropy loss, we constructed one-hot vectors for each unit prediction, then did a binary class softmax operation on the U-Net output. The loss was calculated between onehot labels and softmax results. These two losses were used for both the coarse and refined U-Nets' training.

$$DiceLoss = 1 - \frac{2\sum_{i=0}^{n}(p_i^{pred} * p_i^{true}) + \varepsilon}{\sum_{i=0}^{n} p_i^{pred} + \sum_{i=0}^{n} p_i^{true} + \varepsilon} \qquad (4)$$

We split the whole dataset into 8:2 for training and testing.

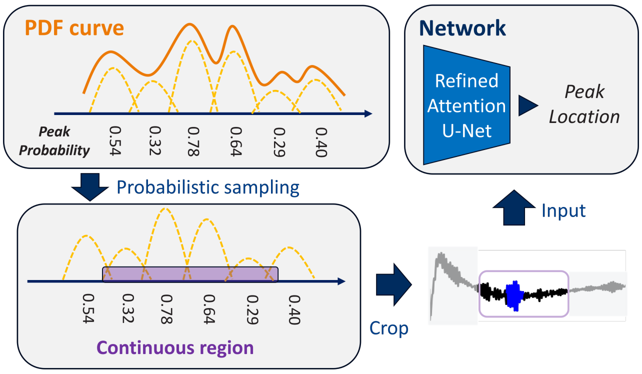


Fig. 7. Sampling-based Proposal: This structure connected the coarse and refined attention U-Net. The Gaussian distribution was built on each unit position of the prediction. Then a mixture PDF curve was built, on which a continuous region was sampled and cropped the original raw signal. The result was the input of the refined U-Net.

For network training, we augmented for 10 times by shifting the signal x-axis up to 1000 units. To train the model, the first coarse U-Net was trained for first 30 epochs so that a relatively accurate coarse region could be generated, then the second refined U-Net was trained for another 20 epochs using the output of Sampling-based Proposal to get the exact positions of bone peaks.

For evaluation, the inference result from the refined UNet could be represented as a sequence of bone peak probability: $S_{prob} = \{p_1, p_2, ..., p_m\}$. After filtering the outputs using the softmax and 0.5 threshold probability, we had a feasible segment of indices in the signal to represent peak positions: $S_{index}\{idx_1, idx_2, ..., idx_n\}$. The exact peak prediction was defined as the middle of this segment, which was expressed as $(max(S_{index}) + min(S_{index}))/2$. With the unit index length, the bone peak depth prediction could be easily calculated by multiplying the position unit index with the interval length from Equation (3): $2 \times \frac{v}{f_s} \times 1000$. The whole step was in Equation (5).

$$Position = Interval \times (max(S_{index}) + min(S_{index}))/2 \quad (5)$$

With the prediction and the ground truth bone depth, our method performance could be evaluated by counting the percentage when the bias of ground truth and prediction was lower than 0.5mm (ACCURACY). We also calculated the average bias (BIAS) and the detection rate whether the bone peak existed or not (FIND PEAK%). We reported the results in the following sections.

III. EXPERIMENTAL RESULTS

Here the method's performance was compared with the traditional method. For each body area (the US holder location), we selected 2 to 3 typical channels, in which the peaks were clear to see and the number of peaks was over half of total 1017 samples. This allowed for the training and facilitated the learning process.

TABLE I
WINDOW WIDTH AND POSITION DETERMINED BY THE EXPERT

LOCATION	CHANNEL	WINDOW POSITION	WINDOW WIDTH
Femur Epicondyle	11	12.427mm	8mm
	12	15.879mm	9mm
Tibia Epicondyle	16	9.205mm	7mm
	17	7.939mm	5mm
	19	12.287mm	7mm
Mid Tibia	24	3.011mm	5mm
	26	6.559mm	5mm
Ankle	28	7.642mm	6mm
	29	5.211mm	5mm

A. Traditional Method Result

For the traditional method to detect bone peaks, the highest peak from a certain window in the signal is annotated and determined as the bone peak using past experience. Here we used the expert knowledge from a previous study [12]. The window position p_{win} and window width w_{win} was shown in TABLE I. As the interval l_{int} was already known in Equation (3), the start and end position of the window (represented as the unit index) was determined using the following Equations.

$$idx_{start} = (p_{win} - w_{win}/2)/l_{int} \quad (6)$$
$$idx_{end} = (p_{win} + w_{win}/2)/l_{int} \quad (7)$$

After cropping the meaningful region, a simple peak detection algorithm from the python library ($scipy.signal.find\_peaks$) was used. It detected the peak that has a certain height range, prominence, and threshold. In our experiment, the height range was determined using the same training dataset, and tested using the same testing dataset of neural network. The result was shown in the left gray background of TABLE II. Noticed that the bias between the prediction and the ground truth was between 1 to 3 mm, showing that the error was large and the accuracy (below 0.48mm) was quite low.

B. Deep Learning Method Result

The result of CasAtt-UNet was shown in the bright right background of TABLE II. Noticed that without any expert knowledge, our model could automatically locate bone peak locations in different areas within sub-millimeter accuracy (except for Channel 28). The average accuracy (below 0.48mm) in all the channels could perform 71.19% on average, which demonstrated the advantage of our method. It was evident that the traditional method could almost achieve 100% peak recognition rate. This is because the traditional method could always find a peak in the window region, as long as all samples in the test dataset existed peaks. Thus, it did not provide useful indication of the performance. However for the CasAtt-UNet, it indeed showed a sensitivity issue, as different static probability threshold could produce various length of possible bone peak segments, directly influencing final peak prediction. A better method would be

TABLE II
RESULTS COMPARISON WITH TRADITIONAL METHOD
(LEFT GRAY PART WAS FROM TRADITIONAL METHOD, RIGHT BRIGHT PART WAS FROM OUR CASATT-UNET.)

LOCATION	CHANNEL	ACCURACY	BIAS	FIND PEAK%	ACCURACY	BIAS	FIND PEAK%
Femur Epicondyle	11	38.30%	1.455mm	92.20%	81.18%	0.348mm	91.18%
	12	27.84%	2.334mm	86.27%	90.07%	0.337mm	92.16%
Tibia Epicondyle	16	14.21%	2.778mm	100.00%	80.22%	0.437mm	92.62%
	17	4.41%	2.808mm	100.00%	63.59%	0.561mm	95.59%
	19	19.80%	2.488mm	96.57%	81.35%	0.275mm	94.53%
Mid Tibia	24	32.35%	1.302mm	100.00%	66.51%	0.554mm	94.77%
	26	26.96%	1.380mm	100.00%	63.52%	0.648mm	89.53%
Ankle	28	28.43%	1.294mm	100.00%	55.25%	1.616mm	85.72%
	29	20.59%	2.211mm	100.00%	59.06%	0.623mm	89.77%

studied later to automatically adjust the probability threshold to achieve the optimal peak recognition performance.

IV. DISCUSSION

This work constructed a deep-learning-based method for bone peak detection in 1D US signals, to measure the bone positions with A-mode ultrasound transducers when applied in the robotic orthopedic application and exoskeletons. The contextual information preservation and feature localization capability of U-Net were greatly improved by cascading different perception resolution U-Nets together, connected by a novel sampling-based proposal mechanism. The introduced attention blocks enabled the learnt patterns more robust and adapt to more channels situations.

one limitation of the work is that we only used one cadaver specimen to collect the dataset, the result may be lack of the variability in patient's anatomy, and had not considered other impacts of the surgical environment factors that can directly impact model's performance. Another limitation is that, compared with other bone registration and reconstruction study, our work only provided and evaluated the distance between skin and bones without completing the whole registration process. However, it is worth noting that the advantage of our method lies in the high precision and automatic bone position measurement under skin, without additional trauma or expert knowledge. Even in the registration process, the previous study [11] reported the registration error as 2.81 mm, which was much worse than our measurement of bone positions.

Besides, this technique could further provide real-time bone locations when installed on the surgical robot arms, as the normal time to process 2D ultrasound images has been removed. Its sub-millimeter accuracy can guide the robot to do high-precision alignment of total knee replacement, also in other similar surgery that had strict requirement. In addition, since the neural network can achieve high precision in identifying the special and sparse bone peaks, this work convinced that deep learning technique was capable to identify and find profiles for very special and irregular signal peaks in the 1D raw signal, which could also inspire other works that require identification of the non-evident and sparse peaks in the 1D raw signal.

V. CONCLUSIONS

In this study, we proposed a novel deep-learning structure to detect highly random and sparse bone peaks in the 1D raw signals with high precision. The measured distance can be used for later bone registration and bone position recovery in real-time, which could be installed on the robotics arms and guide the surgery with sub-millimeter accuracy. The experiment results demonstrated the optimal precision we can achieve, which shows the promising prospective of our system to apply in not only the total knee replacement arthroplasty but also other similar robotics surgeries.

REFERENCES

[1] C. Li, Z. Zhang, G. Wang, C. Rong, W. Zhu, X. Lu, Y. Liu, and H. Zhang, "Accuracies of bone resection, implant position, and limb alignment in robotic-arm-assisted total knee arthroplasty: a prospective single-centre study," *Journal of Orthopaedic Surgery and Research*, vol. 17, no. 1, p. 61, 2022.

[2] T. B. Meier, N. A. Goldfarb, C. J. Nycz, and G. S. Fischer, "Evaluating knee exoskeleton design based on movement with respect to underlying bone structure using mri," *IEEE Transactions on Medical Robotics and Bionics*, 2023.

[3] C. Fang, D. Wang, D. Song, and J. Zou, "The second generation (g2) fingertip sensor for near-distance ranging and material sensing in robotic grasping," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 1506–1512.

[4] K. Shi, R. Huang, F. Mu, Z. Peng, K. Huang, Y. Qin, X. Yang, and H. Cheng, "A novel multimodal human-exoskeleton interface based on eeg and semg activity for rehabilitation training," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 8076–8082.

[5] X. Chen, "Reconstruction individual three-dimensional model of fractured long bone based on feature points," *Computational and Applied Mathematics*, vol. 39, no. 2, p. 131, 2020.

[6] C. Gebhardt, L. Göttling, L. Buchberger, C. Ziegler, F. Endres, Q. Wuermeling, B. M. Holzapfel, W. Wein, F. Wagner, and O. Zettinig, "Femur reconstruction in 3d ultrasound for orthopedic surgery planning," *International Journal of Computer Assisted Radiology and Surgery*, pp. 1–8, 2023.

[7] O. Guinebretiere and J. Giles, "Feasability of a-mode ultrasound based registration to track scapula motion: A simulation study," *EPiC Series in Health Sciences*, vol. 4, pp. 97–102, 2020.

[8] C. Zhang, Y. Liu, Y. Zhang, and H. Li, "A hybrid feature-based patient-to-image registration method for robot-assisted long bone osteotomy," *International Journal of Computer Assisted Radiology and Surgery*, vol. 16, no. 9, pp. 1507–1516, 2021.

[9] C. Liu, Y. Song, X. Ma, and T. Sun, "Accurate and robust registration method for computer-assisted high tibial osteotomy surgery," *International Journal of Computer Assisted Radiology and Surgery*, vol. 18, no. 2, pp. 329–337, 2023.

[10] K. Niu, V. Sluiter, J. Homminga, A. Sprengers, and N. Verdonschot, "A novel ultrasound-based lower extremity motion tracking system," *Intelligent Orthopaedics: Artificial Intelligence and Smart Image-guided Technology for Orthopaedics*, pp. 131–142, 2018.

[11] K. Niu, J. Homminga, V. I. Sluiter, A. Sprengers, and N. Verdonschot, “Feasibility of a-mode ultrasound based intraoperative registration in computer-aided orthopedic surgery: A simulation and experimental study,” *Plos One*, vol. 13, no. 6, p. e0199136, 2018.

[12] K. Niu, T. Anijs, V. Sluiter, J. Homminga, A. Sprengers, M. A. Marra, and N. Verdonschot, “In situ comparison of a-mode ultrasound tracking system and skin-mounted markers for measuring kinematics of the lower extremity,” *Journal of biomechanics*, vol. 72, pp. 134–143, 2018.

[13] K. Niu, J. Homminga, V. Sluiter, A. Sprengers, and N. Verdonschot, “Measuring relative positions and orientations of the tibia with respect to the femur using one-channel 3d-tracked a-mode ultrasound tracking system: A cadaveric study,” *Medical engineering & physics*, vol. 57, pp. 61–68, 2018.

[14] V. Moskalenko, N. Zolotykh, and G. Osipov, “Deep learning for ecg segmentation,” in *Advances in Neural Computation, Machine Learning, and Cognitive Research III: Selected Papers from the XXI International Conference on Neuroinformatics, October 7-11, 2019, Dolgoprudny, Moscow Region, Russia*. Springer, 2020, pp. 246–254.

[15] P.-H. Chen, C.-H. Huang, W.-T. Chiu, C.-M. Liao, Y.-R. Lin, S.-K. Hung, L.-C. Chen, H.-L. Hsieh, W.-Y. Chiou, M.-S. Lee, *et al.*, “A multiple organ segmentation system for ct image series using attention-lstm fused u-net,” *Multimedia Tools and Applications*, vol. 81, no. 9, pp. 11 881–11 895, 2022.

[16] H. Azhari, “Appendix a: Typical acoustic properties of tissues,” 2010.

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

3 Improving Bone Tracking Precision using SIRC-UNet

Compared with the previous chapter, this chapter targets at a similar challenge but with a different solution and evaluation. The introduced SIRC-UNet can recognize the bone peaks in local bone areas instead of the specific positions, and more mathematical equations and algorithm were stated to provide a formal and complete view of the technique. In the analysis, the bias of each area between the prediction and the ground truth has been demonstrated, compared with the traditional method. In addition, the ablation study was done to verify the effectiveness of the proposed loss and the Sampling-based Proposal mechanism.

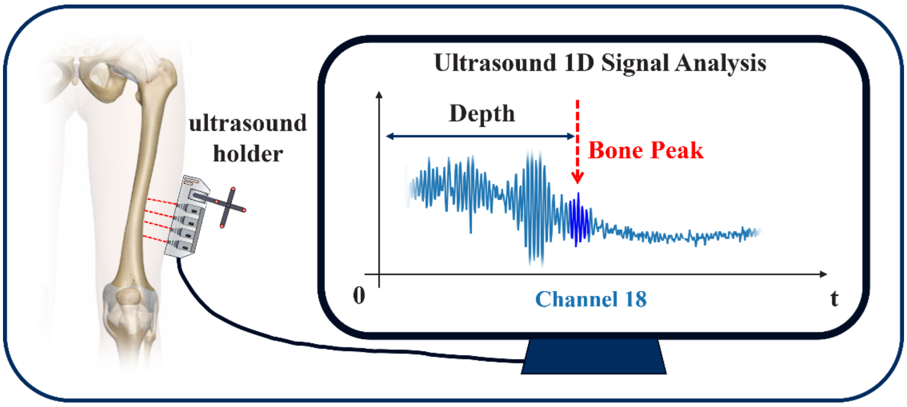
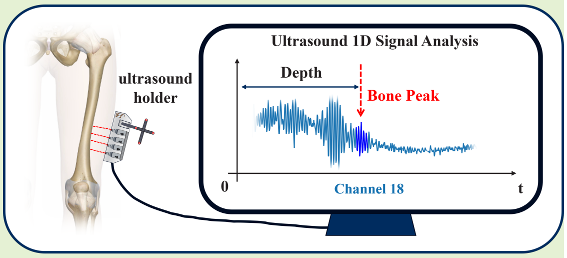


Figure 2: Graphical Abstraction of the Chapter 3

SIRC-UNet: Improving Bone Tracking Precision of A-mode Ultrasound Signals by Decoding Hierarchical Resolution Features

Bangyu Lan, *Student Member, IEEE*, Momen Abayazid *Member, IEEE*, Nico Verdonschot, Stefano Stramigioli *fellow, IEEE*, and Kenan Niu *Member, IEEE*

***Abstract*— A-mode ultrasound (US) has not been widely used in medical applications compared to B-mode ultrasound. The primary reason is that the data representation, being 1-dimensional (1D), is less intuitive to users and harder to interpret. However, A-mode ultrasound transducers have several advantageous features, such as faster data acquisition to allow real-time sensing, direct distance measurement from raw RF data, and a smaller size. A-mode ultrasound has been used to measure biometric distances. However, current distance measurement algorithms are crude, relying on conventional signal processing for peak detection. Especially when the tracking task is under dynamic conditions, it becomes challenging to maintain high accuracy and robustness. In this study, we introduce a novel method to enhance the tracking reliability of the A-mode US under dynamic conditions. This approach aims to improve the accuracy of A-mode US for bone tracking application. SIRC-UNet is designed to enhance the perceptual resolution of the received A-mode ultrasound RF data. It allows for the accurate analysis of the significant signal region, leading to more precise peak detection. The method performance is evaluated by analyzing the bias between the predicted and the ground truth peak locations, and the capability to distinguish bone peaks from irrelevant peaks. The results demonstrate that our method can perform real-time high-precision (sub-millimeter accuracy) bone measurements on the cadaver experiment. It showcases the potential to provide accurate dynamic bone tracking and bone position detection, with possibilities to extend applications to surgical robots and rehabilitation exoskeletons, where real-time bone tracking is crucial.**



***Index Terms*— A-mode ultrasound, dynamic bone tracking, deep learning, peak detection, SIRC-UNet**

I. Introduction

MEDICAL ultrasound (US) is commonly used in diagnostic examinations. It is a safe, non-invasive, and inexpensive diagnostic method. In medical imaging, different types of ultrasound are used, such as A-mode (“A” for amplitude), B-mode (“B” for brightness), M-mode (“M” for motion) ultrasound and Doppler sonography. Among these, the A-mode US is not as popular as others [1]. This is due to the fact that the received signal of the A-mode is visualized as one-dimensional echos plotted as the function of depth, which is less intuitive for the user to interpret compared to 2D images [2]. The A-mode peak profiles are diverse and difficult to analyze [3]. However, the A-mode has its advantages. Compared to the B-mode, which is integrated with multiple A-mode transducers [4], A-mode is more convenient and portable [2], [3], [5]. When measuring, the A-mode US allows real-time sensing and data acquisition, which enables the direct measurement of the distance from the RF data.

In orthopedics, one of the clinical applications for ultrasound diagnosis is bone tracking [6], where traditional methods can introduce unnecessary trauma or measurement deviation. The use of skin-mounted markers is currently widely used to measure the kinematics of the lower extremity [7], [8], but soft tissue artifacts (STA) introduce large errors for measurement [9], [10]. Fluoroscopic systems use radiographic images and model-based methods [11]–[14] to achieve the measurement accuracy of 1mm translation and a 2 degree rotation [15]–[17], but irradiation and high cost hinder the application in reality. Although the B-mode ultrasound-based system can estimate knee joint kinematics without invasion and radiation risk [18], it is more expensive and larger in size. However, to realize reliable joint kinematics tracking, integrating multiple B-mode transducers to cover various anatomical locations is inevitable. Moreover, the A-mode US is more

Bangyu Lan is with the Robotics and Mechatronics group in the Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS), University of Twente, Drienerlolaan 5, 7522 NB Enschede, Netherlands. (email: b.lan@student.utwente.nl)

Momen Abayazid, Stefano Stramigioli, and Kenan Niu are with the Robotics and Mechatronics group, the Faculty of EEMCS in University of Twente, P.O. Box 217, 7500 AE Enschede, Netherlands. (email: m.abayazid@utwente.nl, S.Stramigioli@ieee.org, and k.niu@utwente.nl)

Nico Verdonschot is with the Orthopaedic Research Lab, Radboud University Medical Center, 263 Centraal magazijn, Geert Grooteplein Zuid 30, Nijmegen, Netherlands (email:n.verdonschot@utwente.nl)

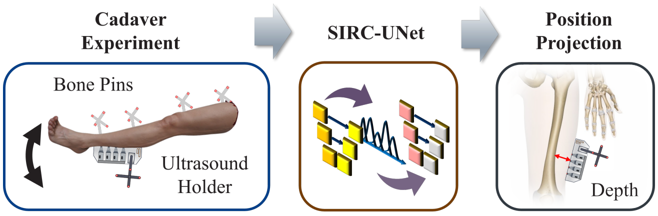


Fig. 1. The overview of the workflow. It includes three parts (from left to right): cadaver experiment for getting a dataset (used as the ground truth position), develop SIRC-UNet for peak recognition and detection, transform peak distance to bone position for validation.

accurate for biometric depth measurement [19], [20]. It is also faster, cheaper, and smaller in size. With these features, the A-mode US is capable of tracking the bone in real time and more precisely.

In bone measurement, the exact three-dimensional (3D) coordinate of bone position can be derived from two sources of information [21]: the distance (or depth) between US transducers and the bony surface, as well as the 3D position of beam origins and directions emitted from the A-mode ultrasound transducer. To perform a distance measurement using A-mode US, the common approach is to detect peak locations empirically [22]. However, the problem that the peak has an unclear profile or a rapid position change cannot be avoided, which is due to the ambiguous interface between the tendon and the bony surface, or the relative shift of the skin and soft tissues [23]. Furthermore, it requires the necessary knowledge of the general positions of bone peaks. A new method is required to better resolve these problems and automatically capture bone peak patterns.

Therefore, the challenge of A-mode US for bone measurement lies in the automatic and precise detection of bone peak in 1D signal. To develop an ideal system, we increase the current precision as described in [24] by using a Deep Learning framework. This framework is inspired by the study of 1D ECG signals in [25], where peak profiles are automatically recognized by UNet to identify specific heart diseases. UNet is capable of preserving contextual information and locating hierarchical characteristics of effective segmentation, which can be used to capture the regular peak shapes in the ECG signal that represent symptoms. However, for the A-mode US signal, there is no such regular profile. The various shapes of the US peaks are the results of stretching and contraction of the tissues [3].

To develop an effective method, different resolution features of the A-mode US signal are exploited for better feature decoding. For the proposed method, **SIRC-UNet** represents **S**ampling-based **I**ncreased **R**esolution **C**ascaded **UNet**, as the resolution of US signal is different for the two UNets. The perceptual region for the second UNet is specified by a novel sampling-based method, which connects two UNets. These designs attempt to make the proposed method more powerful in recognizing the random and dynamic peak profiles in the A-mode US signal.

The cascaded UNets design not only enables dynamic control of the perceptual region, but also is beneficial for peak recognition and detection tasks. As the larger perceptual region is better for peak recognition and the smaller region is better for peak detection, an algorithm is introduced to improve the recognition and detection performance following two UNets' characteristics. Combining peak recognition and detection, our method is expected to have the correct bone peak recognition and the better accuracy for bone tracking.

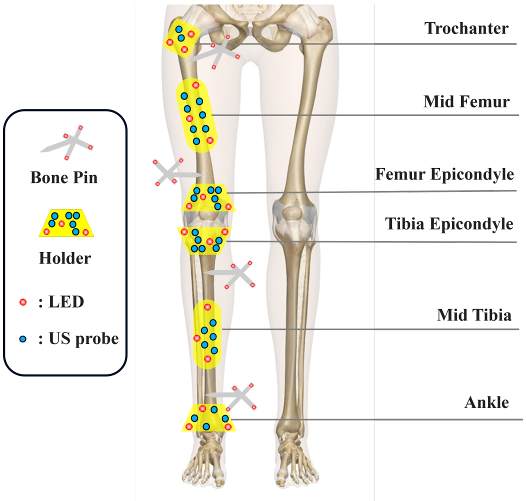


Fig. 2. Schematic overview of the Experimental Setup: The cadaver experiment was performed on a specimen leg. The four bone pins were installed to record the movement of ground truth position, while six US transducer holders were attached to record the ultrasound signal.

In summary, the proposed SIRC-UNet can simultaneously recognize and detect the location of bone peaks in US signals for high-precision bone tracking. In addition, its universal network structure allows other applications to accurately track the motions of various parts of the body, for example, upper limbs, thorax, and neck.

II. Material and Method

This work consists of three stages in Fig. 1: a cadaver experiment to collect the training dataset, the SIRC-UNet architecture for peak recognition and detection, and validation experiments to evaluate the precision of the measurement. The cadaver experiment was carried out on a leg of cadaver samples, where four bone pins were installed to record bone movement, while six US transducer holders were attached to the skin to detect the US signal reflected by the bony surfaces. 3D positions of the bone pins and holders were recorded using an additional optical tracking device. After the experiment, the exact locations of the bones could be retrospectively calculated by calculating the relative positions between the bones and the transducers. These locations correspond to the received US signal. In the second step, they were paired and processed to remove the unqualified US channels, where bone peaks were attenuated or disappeared. The remaining channels formed the dataset for peak detection. The proposed network was trained and validated in the recognition and detection of the bone peaks. In the last step, since the peak location corresponds to the bone position, the inference results could be used to establish the existence of bone below specific skin

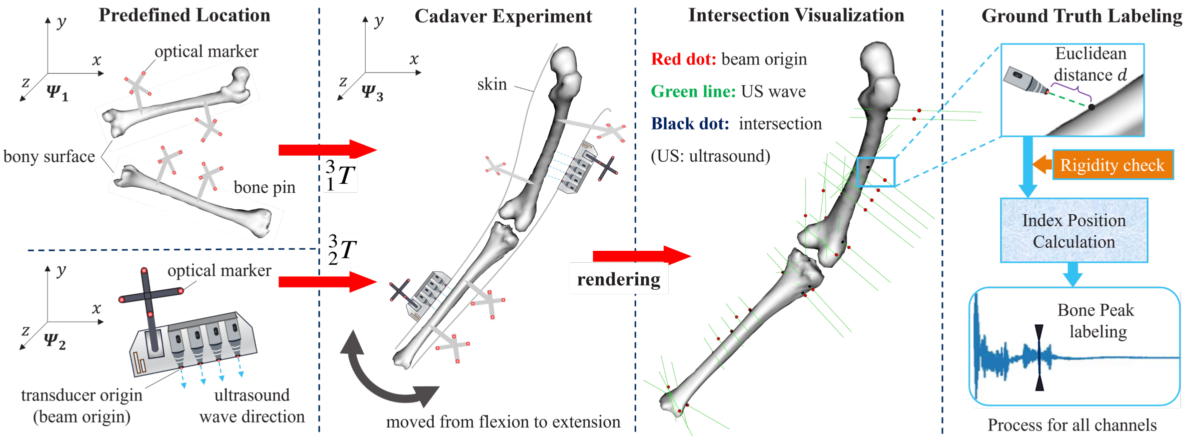


Fig. 3. Ground truth calculation and labeling process: We first define the 3D positions of the bony surface and the holders in Ψ_1 and Ψ_2. With transformation ^3_1T and ^3_1T, the 3D positions of the bones and holders in the experiment Ψ_3 were recovered. The origins of the ultrasound beam and wave directions can be rendered, so the intersection between the US waves and the bony surface can be found. The Euclidean distance representing the bone position is calculated and the corresponding peak position is labeled in the signal.

positions and the bone depth from these locations. The results were compared to the traditional method to demonstrate the advantages of our method.

A. Experimental Setup and Data Acquisition

The cadaver experiment was from a previous research project [26], where a human cadaver specimen (male, 79kg, 179cm) was used to acquire all essential data. This experiment was approved by Radboud University Medical Center (UMCRPS02PRD_PRTC0119). A full leg (from the pelvic to foot) was CT scanned using a TOSHIBA Aquilion ONE scanner, with a voxel size of 0.755mm×0.755mm×0.500mm. The geometrical models of the femur and tibia were then segmented using Mimics 17.0 (Materialise N.V., Leuven, Belgium).

To collect the ultrasound dataset, a total of 30 A-mode US transducers (7.5MHz, Imasonic SAS, Vorayl'Ognon, France) and 18 LED optical markers were installed in the six ultrasound holders. The US holders were strategically placed in key positions on the specimen leg according to a previous study [21]. This was to ensure the optimal reception of the US signals from transducers and accurate positioning of optical markers. Meanwhile, 16 LED optical markers were placed on the four bone pins. The bone pins were installed on the femur and tibia to record the bone movement. Additionally, the sampling rate for the entire US tracking system was 20 Hz. The optical tracking system (Visualeyez VZ4000v trackers, PTI Phoenix Technologies Inc., Vancouver, Canada) that tracked 3D location LED markers was operated at 100 Hz. The US signal was acquired and synchronized in the Diagnostic Sonar FI Toolbox (Diagnostic Sonar Ltd., Livingston, Scotland).

During data collection, the leg was manually moved cyclically from flexion to extension to simulate the swing phase in the gait cycle. On the skin of the leg, A-mode ultrasound transducers were attached. This movement changed the distance between bony surface and US transducers, which introduced the change of bone position that SIRC-UNet predicts. During the movement, the A-mode signal and the trajectories of the attached optical markers were collected. They were used to reconstruct the actual locations of the US transducers and the directions of US waves, as well as the movements of the femur and tibia. Finally, the actual position of the bone was derived from the Euclidean distance between the bony surface and the transducers in the US beam directions. The calculation in details is Equation (5).

In total, a dataset containing continuous leg movement was collected. The movement sequence was divided into 1017 samples (moments). Each sample has 30 transducer channels including the received US signals and 34 optical marker positions for bones and transducer movement recording. Note that the Trochanter and Mid Tibia datasets were discarded due to the attenuation and disappearance of signal peaks, making them unsuitable for peak detection. The rest of channels were screened in four typical locations (the number of bone peaks should be greater than 70% of the total samples for better training). In the end, the four anatomical areas (Femur Epicondyle, Tibia Epicondyle, Mid Tibia, and Ankle) provided 12 available US channels and 18 LED optical marker trajectories. They are used for peak detection of the A-mode ultrasound.

B. Ground Truth Calculation and Labeling

The prediction target of SIRC-UNet is the distance between the transducer origin and the intersection between the US wave and the bony surface. To derive the distance, the required data from the cavader experiment includes three parts: ($\{*\}$ means a series of points)

- In the bone coordinate frame Ψ_1, bony surface $\{P^{1,B}\}$ segmented from CT scan and the bone pins installed with the optical markers $\{P^{1,M}\}$;
- In the holder coordinate frame Ψ_2, six ultrasound holders including optical markers $\{P^{2,M}\}$, origins of the transducers $\{P^{2,O}\}$ and the directions of the US waves;
- In the experiment coordinate frame Ψ_3, the optical markers $\{P^{3,M}\}$ of both bone pins and US holders in the cadaver experiment.

The whole process is illustrated in Fig. 3.

The idea is to use optical markers of bone pins and holders as intermediaries to transfer the bony surface of the frame Ψ_1 and the US transducers of the frame Ψ_2 to the experimental coordinate frame Ψ_3. In this way, the positions of the transducers, the directions of the US waves, and the surface of the femur and tibia are expressed in the same coordinate frame Ψ_3.

The derivation algorithm is as follows. The positions of the optical markers in Ψ_2 are $\{P^{2,M}\}$, which correspond to $\{P^{3,M}\}$ in the experiment. The transform matrix ^3_2T can be derived using $f(\{P^{2,M}\},\{P^{3,M}\})$. Then the origins of US transducers can be transformed from $\{P^{2,O}\}$ to $\{P^{3,O}\}$ using the same transformation ^3_2T. Similarly, the bone positions $\{P^{3,B}\}$ in Ψ_3 can be derived from $\{P^{1,B}\}$ using ^3_1T, which is from $f(\{P^{1,M}\},\{P^{3,M}\})$. To align the two coordinates correctly, VTK Landmark Transformation Library [27] is used to perform the alignment $f(P^{source},P^{target})$.

$$^3_1T = f(\{P^{1,M}\},\{P^{3,M}\}) \tag{1}$$

$$\{P^{3,B}\} =^3_1 T \cdot \{P^{1,B}\} \tag{2}$$

$$^3_2T = f(\{P^{2,M}\},\{P^{3,M}\}) \tag{3}$$

$$\{P^{3,O}\} =^3_2 T \cdot \{P^{2,O}\} \tag{4}$$

After knowing the bony surface $\{P^{3,B}\}$ and the origins of the US transducer $\{P^{3,O}\}$, the intersection between the US waves and the bony surface can be derived. This is calculated for every moment in the experiment. The Euclidean distance ($d(mm)$) between the transducer origin and the intersection with the US wave can infer the peak location in the 1D signal, as the US waveform hitting and reflection process will create the signal peak in the echo at the corresponding location. However, as sometimes the optical tracking device cannot detect the optical marker positions due to the occlusion, a rigidity check is performed to ensure that the change of bone peak location is aligned with the actual bone movement: The positions of optical markers $\{P^{3,M}\}$ are examined to ensure a rigid transformation during the experiment, so that the calculated transformations ^3_1T and ^3_2T are correct. The peak index (idx) in the signal is determined using knowledge from the ultrasound device. The equation is as follows:

$$idx = \frac{d}{d_{unit}} = \frac{d}{(2 \times \frac{v}{f_s} \times 1000)} \tag{5}$$

where $v = 1540m/s$ refers to the speed of the US in soft tissues (mainly muscle), and $f_s = 40 \times 10^6 Hz$ refers to the sampling rate of the US. The formula $d_{unit} = 2 \times \frac{v}{f_s} \times 1000$ is the length of one unit over the 6760 units length of the signal. The length of one unit is calculated as $0.01925mm$, which is the minimum precision of the measurement.

When conducting a visualization analysis to validate the calculations, we observed a clear relationship between the calculated index idx and the location of the peak in the signal. This shows that the signal peak can represent the distance of bone depth under the skin. These calculated indexes are the ground truth labels of the US signal, and used as the target of the peak detection.

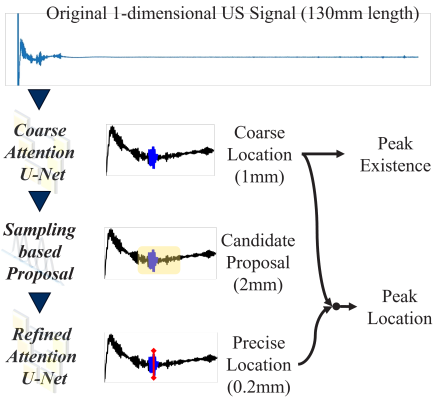


Fig. 4. Cascaded Attention design in SIRC-UNet. It includes two attention UNets that focus on different regions of signal. The coarse UNet recognizes the peak and give a coarse detection. The intermediate mechanism determines the input of the refined UNet from coarse UNet. The refined UNet predict more precise peak location.

C. Overview of the SIRC-UNet architecture

From the cadaver experiment and the ground truth labeling process, ultrasound signals and labels of 12 transducer channels were collected. In addition to detect the peak position, SIRC-UNet is also required to identify the peak existence, as sometimes US waves do not intersect with any bony surfaces. At these moments, the segmentation of SIRC-UNet should be only the background. Since the bone peak label (index) is only one unit, which is difficult for network segmentation, we increase the length of label to a sufficient width (coarse label: 50 units, refined label: 10 units) for peak recognition and introduce a mechanism to increase the input resolution. In post-processing, we recover the exact one-unit bone peak location from the detection result.

The proposed network has two attention UNets, and a sampling-based strategy to select the region of signal for the second UNet. The first UNet identifies the existence of bone peaks and performs coarse position detection. The sampling-based method increases the signal resolution by selecting the most likely region from the coarse probability. This region of the signal is used as the input of the second UNet, which predicts the precise peak position. The final result is determined by both the coarse and refined predictions. The overall architecture is shown in Fig. 4.

D. Cascaded Attention UNet

The UNet for peak recognition and detection is inspired by the study of ECG signal [25]. An attention block similar to [28] is incorporated to highlight the prominent features of the skip connections. In the A-mode signal, there are other tissue peaks and random noise that confuse the bone peak recognition. The coarser scale of features in the attention block can filter out irrelevant and noisy features in the skip connections. In this way, only the relevant features of the bone peaks are merged into the concatenation operations.

To be concrete, the attention UNet has the structure in Fig. 5, where the attention block uses coarser scales of contextual features as the gate signal to filter irrelevant features in the

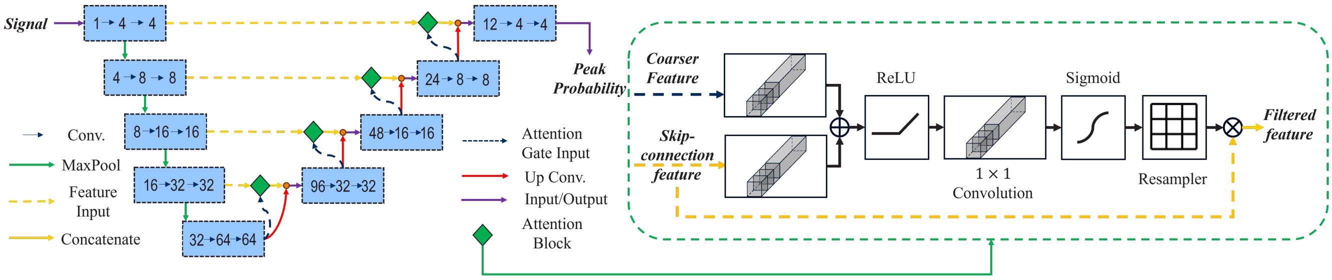


Fig. 5. Attention UNet structure. The UNet with an attention block is shown here. The light blue box includes two 1-dimensional convolution, represented as the blue solid arrow, connecting the number of convolution kernels before and after. The detail in the 1×1 convolution in the attention block is from [28].

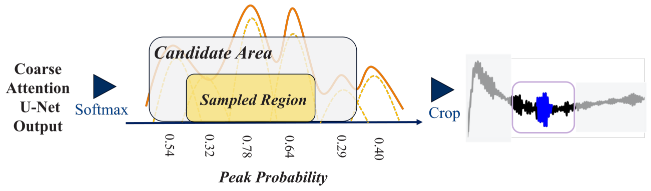


Fig. 6. Sampling-based Proposal. The softmax is applied to get the coarse peak probability. Then a candidate area is proposed and a region is sampled on this area. This region determines the input of the refined UNet.

skip connections [28]. The output is concatenated with the up convolution of the coarser features and goes into the left feature decoder. This helps suppress irrelevant signal areas and highlight salient peak features.

Although the two attention UNets share the same structure, they have different priorities for recognition and detection. The coarse UNet identifies the bone peak existence in full resolution and gives a coarse peak probability. On the basis of that, a sampling-based mechanism proposes a candidate region by probability-based sampling. The proposed region is the input of the second UNet for the precise detection of the bone peak. The cascaded UNet pipeline increases signal resolution to improve peak detection accuracy.

As the signal resolution is different for two UNets, different labels should be applied accordingly. The peak label for the coarse UNet is expanded to 50 units of length (approximately 1 mm), while the peak label for the refined UNet is only expanded to 10 units of length (approximately 0.2 mm). This complements the information of bone peaks when inference, as it is difficult to quantify the actual width of the bone peak. The peak profile is influenced not only by the medium layers between bone and soft tissues but also by the relative movement of multiple soft tissue layers when muscles contract or stretch. The method to determine the exact peak location and the existence of the peak from two UNets prediction is illustrated in the II-F Peak Location Determination.

E. Sampling-based Proposal

To define the input signal of the refined UNet for precise detection, a direct strategy is to use the segment of signal where the peak is most likely to appear. This segment can be determined directly by the most probable peak region of the coarse UNet, but it has two problems: First, not every time the coarse UNet can detect exactly the peak position, sometimes the peak is far away. In this situation, the proposed region does not contain any peak. The second problem occurs when the coarse UNet is well trained, and each time the peak is exactly in the middle of the output (the expansion of the label is symmetrical around the peak location). Then it is difficult for the refined UNet to learn effectively, as it can directly select the middle position of the proposed region as the location of bone peak.

An improved strategy is to change the proposed region dynamically: sample a segment based on the probability of the coarse UNet output. As the proposed region keeps changing, the learning of coarse UNet is disentangled with the refined UNet. The refined UNet is required to learn without any reference from the coarse UNet. In this way, training becomes stable and test performance is improved. Inspired by the region proposal in [29], we develop a proposal method based sampling for the selection of the effective region.

The detail is illustrated in Fig. 6. During training, the Softmax operation is applied to the coarse UNet output to get the peak probability. Based on it, the sampled region is determined as a typical area for the efficient computation instead of the entire signal range (approximately 130 mm): we initially identify the peak with the highest probability. Following this, the sampled region is created around this identified peak to ten times its width. In this area, a region is sampled to decide the start and end of the index. These indexes of proposal crop the original signal as the input of the refined UNet. In the testing (validation), since there is no requirement for sampling, the proposed region is the segment that has the highest probability peak from the coarse UNet output.

F. Peak Location Determination

As both UNets predict bone peaks as their detection results, a strategy is required to determine the existence of the peak and pinpoint the most precise location. To make the final decision, an algorithm is designed in **Algorithm 1**. The idea is that the coarse UNet recognizes the existence of the peak, while the refined UNet detects the location of the peak. There are two abnormal situations: (1) If the coarse UNet does not find the peak while the refined UNet does, the bone peak is treated as non-existent. (2) If the coarse UNet find the peak,

Algorithm 1 Peak Location Determination

```
Input: coarse segmentation, refined segmentation
if coarse segmentation $\neq 0$ then
    peak exist
    if refined segmentation $\neq 0$ then
        location = middle of refined segment
    else
        location = middle of coarse segment
    end if
else
    peak does not exist
end if
Output: peak existance, peak location
```

while the refined UNet does not, the peak is treated as existing and the location is decided from the coarse UNet segmentation.

The peak detection output is a segmentation. The result of each unit is decided by the higher probability between foreground and the background. To transform segmentation to a peak location, Equation (6) is used. The probability after Softmax is $O_{prob} = \{p_1, p_2, ..., p_m\}$, which is filtered by 0.5 to get the segment $P_{index} = \{idx_1, idx_2, ..., idx_n\}, n \neq m$. The peak index is the middle $idx_{peak} = (idx_1 + idx_n) \times 0.5$. And each unit length is calculated in the Equation (5).

$$\begin{aligned} d_{peak} =& d_{unit} \times idx_{peak} \\ =& (2 \times \frac{v}{f_s} \times 1000) \times \\ & (idx_1 + idx_n) \times 0.5 \end{aligned} \tag{6}$$

G. Training Strategy

To train our SIRC-UNet, the US signal and labels were first processed to remove outliers (strength greater than 5000) and shuffled into 8:2 training versus testing dataset. To demonstrate the universal and generalized ability, the channels from the same segment of the lower limb are grouped together to form the dataset.

For training loss, dice loss (DL) and cross-entropy loss are used for the segmentation task. Dice loss [30] considers both the recall and precision rates of the target to solve the sparse foreground (target) issue in medical applications. It is the Equation (7).

$$DiceLoss(DL) = 1 - \frac{2 \times \sum_{i=0}^{n} (p_i^{pred} * p_i^{true}) + \varepsilon}{\sum_{i=0}^{n} p_i^{pred} + \sum_{i=0}^{n} p_i^{true} + \varepsilon} \tag{7}$$

For the cross-entropy loss, it is calculated for the binary classification of each unit. In training, Softmax is first applied to the UNet output to have the peak probability O_{prob}, which is compared with the one-hot 2×1 labels to calculate the cross-entropy loss. This loss l_1 and the dice loss l_2 are added together to train only the coarse UNet for 30 epochs, as if the first UNet has not been well trained, sampling-based proposal cannot propose an effective region for the refined UNet.

TABLE I
WINDOW WIDTH AND POSITION FOR US PEAK

LOCATION	CHANNEL	WINDOW POSITION	WINDOW WIDTH
Femur Epicondyle	11	12.427mm	8mm
	12	15.879mm	9mm
	15	21,862mm	10mm
Tibia Epicondyle	16	9.205mm	7mm
	17	7.939mm	5mm
	18	5,383mm	6mm
	19	12.287mm	7mm
	20	8,660mm	6mm
Mid Tibia	24	3.011mm	5mm
	26	6.559mm	5mm
Ankle	28	7.642mm	6mm
	29	5.211mm	5mm

The refined UNet is trained from the 30^{th} epoch to the 50^{th} epoch. Training loss for refined UNet is also the summation of cross-entropy loss l_3 and dice loss l_4. The second training includes both coarse UNet and refined UNet. For the entire training, "RMSprop" is used as the optimization with an initial learning rate 0.00001.

$$loss_{stage1} = l_1 + l_2 \tag{8}$$

$$loss_{stage2} = l_1 + l_2 + l_3 + l_4 \tag{9}$$

III. EXPERIMENTAL RESULTS

To comprehensively evaluate the method, the traditional method was used to compare the performance of peak recognition and detection. Peak recognition represents whether the method can recognize the bone peak, while peak detection represents how precisely the method can locate the peak position.

For peak recognition, in some moments when US waves did not intersect with the bony surface, the bone peaks do not exist. SIRC-UNet needs to distinguish the existing bone peaks from the irrelevant peaks due to noise or other tissues. When the UNet segmentation only contains background, the bone peak is treated as disappeared.

For peak detection, the absolute difference of the index between the predicted peak and the ground truth label is calculated. The difference is transformed into the distance (mm). To show the distribution of distance (or errors), the quartile-based analysis is shown in TABLE II. For a more clear comparison, the boxplot was generated from the table and shown in Fig. 7, where each error distribution has a middle box, a median red line and two extended boundaries. The middle box ranges from the lower quartile (Q1) to the upper quartile (Q3). The extended boundary corresponds to 1.5 IQR (interquartile range = Q3-Q1) outside the middle box range.

A. Results of Traditional Method

For traditional method, the selection of bone peak is based on a channel-specific window size (p_{win}) and window position (p_{win}), which are the locations of femur and tibia under the skin. These parameters are established before the experiment [24] and are presented in TABLE I. To translate the values

TABLE II

THE PEAK DETECTION IN QUARTILE-BASED ACCURACY AND PEAK RECOGNITION RATE FOR THE COMPARISON OF TWO METHODS.

METHOD		TRADITIONAL METHOD						SIRC-UNet						
Area	Channel	Quartile-based Accuracy (mm)					Recogn-ition(%)	Dataset	Quartile-based Accuracy (mm)					Recogn-ition(%)
		mean	std_dev	25%	50%	75%			mean	std_dev	25%	50%	75%	
Femur Epicondyle	11	1.455	1.494	0.245	0.876	2.127	92.16	group1	0.978	7.632	0.0	0.019	0.058	90.65
	12	2.334	2.141	0.327	2.204	3.412	86.27							
	15	3.276	3.183	0.394	2.252	5.135	92.16							
Tibia Epicondyle	16	2.778	2.111	0.886	2.349	4.668	100.0	group2	0.522	0.846	0.02	0.077	0.674	90.79
	17	2.808	1.356	1.838	2.926	3.754	100							
	18	2.033	1.472	0.746	1.858	2.931	100							
	19	4.686	1.477	3.715	4.582	5.852	96.57							
	20	3.314	1.907	1.973	3.350	4.586	100							
Mid Tibia	24	1.302	1.095	0.308	1.059	2.069	100.0	group3	0.562	0.844	0.0	0.019	0.905	90.43
	26	1.380	1.178	0.462	1.030	1.988	100							
Ankle	28	3.482	2.517	1.025	3.619	5.217	94.12	group4	1.108	1.950	0.0	0.019	1.232	90.43
	29	1.167	1.058	0.173	0.963	1.867	94.61							

TABLE III

UPPER QUARTILE (75%) ACCURACY IN PEAK DETECTION, DL: DICE LOSS, SBP: SAMPLING-BASED PROPOSAL, W/O: WITHOUT

SEGMENT	w/o DL (mm)	w/o SBP (mm)	Full Model(mm)	Improvement
Femur Epi.	0.096	0.077	0.058	32.5%
Tibia Epi.	0.693	0.727	0.674	5.03%
Mid Tibia	0.982	0.977	0.905	7.62%
Ankle	1.328	1.328	1.232	7.24%

into the index of signal range, we use the Equations (10) and (11) to determine the window range with the interval length (d_{unit}) calculated from Equation (5).

$$idx_{start} = (p_{win} - w_{win}/2)/d_{unit} \tag{10}$$

$$idx_{end} = (p_{win} + w_{win}/2)/d_{unit} \tag{11}$$

The start and end indices define the candidate area that has the bone peak. Within this region, an automatic peak detection algorithm (Python library *scipy.signal.find_peaks*) is applied. Among all candidate peaks, the bone peak is the one that has the highest strength. This is based on the characteristics that a strong ultrasound response is only created by acoustic bone reflection within the candidate area. Typically, no higher peak should occur in subsequent segments of the signal, except in the case of overlapping secondary echoes from multiple tissues located in the front.

The results of the traditional method, including peak detection accuracy and recognition rate for all channels, are illustrated to the left of TABLE II. In Fig. 7, they are visualized as light blue boxes. The errors predominantly range from 0.5 to 6 mm, as shown by the Q1 to Q3 range.

B. Results of Deep Learning Method

The results of SIRC-UNet in terms of peak recognition and detection are presented on the right side of TABLE II, and is visualized as dark blue boxes in Fig. 7, where a precision threshold of 1 mm is indicated as the red horizontal dashed line. To analyze the effectiveness of dice loss and Sampling-Based Proposal, we trained the SIRC-UNet again without these components. The results are summarized in TABLE III. In addition to assessing peak recognition and detection capabilities, we also tested inference time of SIRC-UNet, particularly for its application in the real-time tracking. The model has a rapid processing speed of 15 ms per signal, while the traditional method requires 1.5 hours for accessing window information [24].

IV. Discussion

For peak detection, the lack of precision in traditional method can be attributed to several factors. Primarily, the bone peak might be obscured by a shift of the peak induced by movements or noise within the signal. In addition, bone peaks may not always have regular shapes for easy detection. Relying only on identifying the highest peaks, without considering peak profiles or the possibility of secondary echos, can lead to inaccuracies.

In SIRC-UNet, for the segment of Femur Epicondyle, Tibia Epicondyle, and Mid Tibia, the precision for over 75% bone peaks does not exceed 1mm, achieving sub-millimeter precision. Even in more challenging areas like the ankle with more curvatures, where bone peak localization is more difficult compared to other bony surfaces with less curvatures, most bone peaks (from Q2 to Q3) meet this precision. The precision of 50% peaks (red line) in all areas approaches zero, demonstrating the exceptional precision. The accuracy can reach sub-millimeter for different segments of the limb. The extended tails above Q3 of the boxplot, representing shifts from Q3 to 1.5 times of the interquartile range, are attributed to occasional processing noise or irrelevant echoes that mimic bone peak profiles. However, from a broader perspective, SIRC-UNet significantly enhances precision over traditional methods, consistently achieving sub-millimeter accuracy in bone peak detection.

Another advantage of SIRC-UNet is that, in each segment of the lower limb, SIRC-UNet is trained on the dataset of all channels. These channels have different peak locations, strengths, and profiles as shown in TABLE I. Despite the advantage of the varied prior knowledge in window size and width, our SIRC-UNet model still performs significantly better than the traditional method. This high level of precision could be attributed to the accurate recognition of the peak profile across different channels. This suggests that the proposed

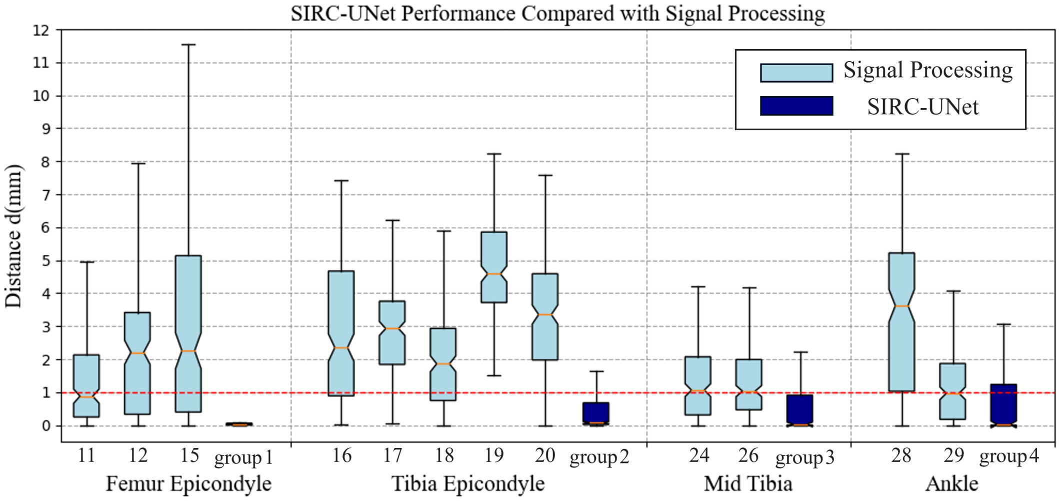


Fig. 7. Boxplot from TABLE II. The traditional method is represented as light blue boxes, while SIRC-UNet is represented as dark blue boxes. The red dashed line refers to the 1 mm precision. For the traditional method, the boxes correspond to the 12 channels. For SIRC-UNet, groups 1-4 refer to the combined channels data in this segment.

method distinguishes the unique bone peak from irregular noise or interference from other soft tissue echo peaks.

For peak recognition, the traditional method performs well. This is because the traditional method can always find a signal spike in the specified region without knowing the bone peak profile. The peak detection from signal processing field can always give a prediction, no matter whether this is the bone peak. Therefore, the relative significance of peak recognition is considered minimal compared to that of the SIRC-UNet.

In the SIRC-UNet result, the performance can achieve around 90% precision to distinguish bone peaks from irrelevant peaks in TABLE II. Several factors attribute to the 10% errors. Firstly, the presence of peaks resulting from processing noise or other soft tissues with similar profiles can interfere with the accurate recognition. In addition, attenuation of bone peaks or their overlap with echoes from irrelevant tissues can complicate the recognition process. These factors often lead to a more irregular peak shape, diverging from the regular features typically found in the training dataset. This increased variability poses a challenge for the deep learning model, especially in the absence of fixed parameters such as window size and width, which are used in the traditional approach.

In TABLE III to compare the variants of SIRC-UNet, the average improvement of dice loss (DL) and Sampling-based Proposal (SBP) for all channels is 13.1%. This is a considerable improvement that underscores the effectiveness of dice loss and sampling-based proposal mechanism.

For the inference time, the speed of 15 ms per signal means that our method can process more than 60 signals per second during real-time data streaming, without the need for additional parameters or pauses. This is due to the architecture that includes fewer 1D convolution kernels and shallower layers compared to the traditional 2D UNet, and a simplified argmax operation in the SBP mechanism. Alongside precision, speed is another key advantage of SIRC-UNet for bone tracking applications.

Based on the analysis, unlike traditional methods that heavily rely on the prior knowledge to achieve millimeter-level precision, our method operates without the need for predefined parameters to obtain sub-millimeter precision in peak detection and recognition. This capability makes it particularly useful for real-time bone tracking. The implementation of Sampling-Based Proposal enhances model's training robustness and performance by focusing on bone peak's nearby region. And the integration of dice loss optimizes the segmentation result by simultaneously considering both precision and recall rates.

Contrasting with the traditional method, our SIRC-UNet eliminates the need for expert knowledge and adapts seamlessly to varying peak profiles across different segments of the lower limb without any change in network structure. Its ability to infer in real time makes it a valuable tool for integration into medical robotics, such as in orthopedic surgery [31] or wearable exoskeletons [32] to analyze the kinematics of human movement. The enhanced precision and automation of bone measurement brought about by SIRC-UNet increases the accuracy of bone registration. This improvement makes the use of A-mode ultrasound in medical applications, particularly in tracking tasks, more feasible and efficient.

V. Conclusion

In this study, we introduced SIRC-UNet, a novel deep learning-based method designed to recognize and detect bone peaks in raw 1D A-mode ultrasound signals. This detection is crucial for measuring bone position in real-time and facilitates high-precision bone tracking tasks. Thanks to its sub-millimeter precision and efficient network structure, our method excels in performing bone tracking with high accuracy and in real-time. These capabilities demonstrate SIRC-UNet's significant potential for widespread application in various medical tasks, particularly in enhancing the utility of A-mode ultrasound.

References

[1] B. Meikle, R. M. Kimble, and Z. Tyack, "Ultrasound measurements of pathological and physiological skin thickness: A scoping review protocol," *BMJ open*, vol. 12, no. 1, p. e056720, 2022.

[2] D. R. Wagner, M. Teramoto, T. Judd, J. Gordon, C. McPherson, and A. Robison, "Comparison of a-mode and b-mode ultrasound for measurement of subcutaneous fat," *Ultrasound in medicine & biology*, vol. 46, no. 4, pp. 944–951, 2020.

[3] L. Guo, Z. Lu, L. Yao, and S. Cai, "A gesture recognition strategy based on a-mode ultrasound for identifying known and unknown gestures," *IEEE Sensors Journal*, vol. 22, no. 11, pp. 10 730–10 739, 2022.

[4] P. Hauff, M. Reinhardt, and S. Foster, "Ultrasound basics," *Molecular Imaging I*, pp. 91–107, 2008.

[5] K. Niu, T. Anijs, V. Sluiter, J. Homminga, A. Sprengers, M. A. Marra, and N. Verdonschot, "In situ comparison of a-mode ultrasound tracking system and skin-mounted markers for measuring kinematics of the lower extremity," *Journal of biomechanics*, vol. 72, pp. 134–143, 2018.

[6] S. Feng, Q.-T.-K. Shea, K.-Y. Ng, C.-N. Tang, E. Kwong, and Y. Zheng, "Automatic hyoid bone tracking in real-time ultrasound swallowing videos using deep learning based and correlation filter based trackers," *Sensors*, vol. 21, no. 11, p. 3712, 2021.

[7] A. Cappozzo, F. Catani, U. Della Croce, and A. Leardini, "Position and orientation in space of bones during movement: anatomical frame definition and determination," *Clinical biomechanics*, vol. 10, no. 4, pp. 171–178, 1995.

[8] S. P. Rana, M. Dey, M. Ghavami, and S. Dudley, "Markerless gait classification employing 3d ir-uwb physiological motion sensing," *IEEE Sensors Journal*, vol. 22, no. 7, pp. 6931–6941, 2022.

[9] M. Akbarshahi, A. G. Schache, J. W. Fernandez, R. Baker, S. Banks, and M. G. Pandy, "Non-invasive assessment of soft-tissue artifact and its effect on knee joint kinematics during functional activity," *Journal of biomechanics*, vol. 43, no. 7, pp. 1292–1301, 2010.

[10] W. Schallig, G. J. Streekstra, C. M. Hulshof, R. P. Kleipool, J. G. Dobbe, M. Maas, J. Harlaar, M. M. van der Krogt, and J. C. van den Noort, "The influence of soft tissue artifacts on multi-segment foot kinematics," *Journal of Biomechanics*, vol. 120, p. 110359, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021929021001391

[11] W. Anderst, R. Zauel, J. Bishop, E. Demps, and S. Tashman, "Validation of three-dimensional model-based tibio-femoral tracking during running," *Medical engineering & physics*, vol. 31, no. 1, pp. 10–16, 2009.

[12] J. Bingham and G. Li, "An optimized image matching method for determining in-vivo tka kinematics with a dual-orthogonal fluoroscopic imaging system," 2006.

[13] J. E. Giphart, C. A. Zirker, C. A. Myers, W. W. Pennington, and R. F. LaPrade, "Accuracy of a contour-based biplane fluoroscopy technique for tracking knee joint kinematics of different speeds," *Journal of biomechanics*, vol. 45, no. 16, pp. 2935–2938, 2012.

[14] S. Banks and P. Flood, "Jointtrack auto: An open-source programme for automatic measurement of 3d implant kinematics from single-or biplane radiographic images," in *Orthopaedic Proceedings*, vol. 98, no. SUPP_1. Bone & Joint, 2016, pp. 38–38.

[15] G. Li, S. K. Van de Velde, and J. T. Bingham, "Validation of a non-invasive fluoroscopic imaging technique for the measurement of dynamic knee joint motion," *Journal of biomechanics*, vol. 41, no. 7, pp. 1616–1622, 2008.

[16] S. Guan, H. A. Gray, F. Keynejad, and M. G. Pandy, "Mobile biplane x-ray imaging system for measuring 3d dynamic joint motion during overground gait," *IEEE transactions on medical imaging*, vol. 35, no. 1, pp. 326–336, 2015.

[17] M. Kozanek, A. Hosseini, F. Liu, S. K. Van de Velde, T. J. Gill, H. E. Rubash, and G. Li, "Tibiofemoral kinematics and condylar motion during the stance phase of gait," *Journal of biomechanics*, vol. 42, no. 12, pp. 1877–1884, 2009.

[18] M. A. Masum, M. Pickering, A. Lambert, J. Scarvell, and P. Smith, "Accuracy assessment of tri-plane b-mode ultrasound for non-invasive 3d kinematic analysis of knee joints," *Biomedical engineering online*, vol. 13, no. 1, pp. 1–16, 2014.

[19] W. A. Hamidzada and E. P. Osuobeni, "Agreement between a-mode and b-mode ultrasonography in the measurement of ocular distances," *Veterinary Radiology & Ultrasound*, vol. 40, no. 5, pp. 502–507, 1999.

[20] G. Barsotti, S. Citi, M. Brovelli, E. Mussi, E. Luchetti, F. Carlucci, and M. Sgorbini, "Equine ocular ultrasonography: Evaluation of some biometric measurements," *Ippologia*, vol. 21, pp. 39–43, 09 2010.

[21] K. Niu, J. Homminga, V. Sluiter, A. Sprengers, and N. Verdonschot, "Measuring relative positions and orientations of the tibia with respect to the femur using one-channel 3d-tracked a-mode ultrasound tracking system: A cadaveric study," *Medical engineering & physics*, pp. 61–68, 2018.

[22] A. K. Sahani, J. Joseph, and M. Sivaprakasam, "Automatic measurement of lumen diameter of carotid artery in a-mode ultrasound," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2013, pp. 3873–3876.

[23] A. Ekizos, F. Papatzika, G. Charcharis, S. Bohm, F. Mersmann, and A. Arampatzis, "Ultrasound does not provide reliable results for the measurement of the patellar tendon cross sectional area," *Journal of Electromyography and Kinesiology*, vol. 23, no. 6, pp. 1278–1282, 2013.

[24] K. Niu, V. Sluiter, J. Homminga, A. Sprengers, and N. Verdonschot, "A novel ultrasound-based lower extremity motion tracking system," *Intelligent Orthopaedics: Artificial Intelligence and Smart Image-guided Technology for Orthopaedics*, pp. 131–142, 2018.

[25] V. Moskalenko, N. Zolotykh, and G. Osipov, "Deep learning for ecg segmentation," in *Advances in Neural Computation, Machine Learning, and Cognitive Research III: Selected Papers from the XXI International Conference on Neuroinformatics*. Springer, 2020, pp. 246–254.

[26] K. Niu, J. Homminga, V. I. Sluiter, A. Sprengers, and N. Verdonschot, "Feasibility of a-mode ultrasound based intraoperative registration in computer-aided orthopedic surgery: A simulation and experimental study," *Plos One*, vol. 13, no. 6, p. e0199136, 2018.

[27] W. Schroeder, K. Martin, and B. Lorensen, *The Visualization Toolkit (4th ed.)*. Kitware, 2006.

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

[29] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "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.

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

[31] C. Li, Z. Zhang, G. Wang, C. Rong, W. Zhu, X. Lu, Y. Liu, and H. Zhang, "Accuracies of bone resection, implant position, and limb alignment in robotic-arm-assisted total knee arthroplasty: a prospective single-centre study," *Journal of Orthopaedic Surgery and Research*, vol. 17, no. 1, p. 61, 2022.

[32] T. B. Meier, N. A. Goldfarb, C. J. Nycz, and G. S. Fischer, "Evaluating knee exoskeleton design based on movement with respect to underlying bone structure using mri," *IEEE Transactions on Medical Robotics and Bionics*, 2023.

4 Anatomical Region Perception Bone Tracking Methods

Building on the previous chapters, the third chapter explores a simplified and efficient deep learning structure for the real-time bone tracking method that have accurate anatomical region perception. This approach utilized the end-to-end cascaded U-Nets to effectively find the bone peaks and classify anatomical regions under the bone dynamic motions. The high accuracy to identify knee joint areas and measure the bone demonstrate the improvement over the traditional methods and methods in previous chapters. In the end of this chapter, the good and worse recognition cases have been analyzed to explore the source of prediction errors.

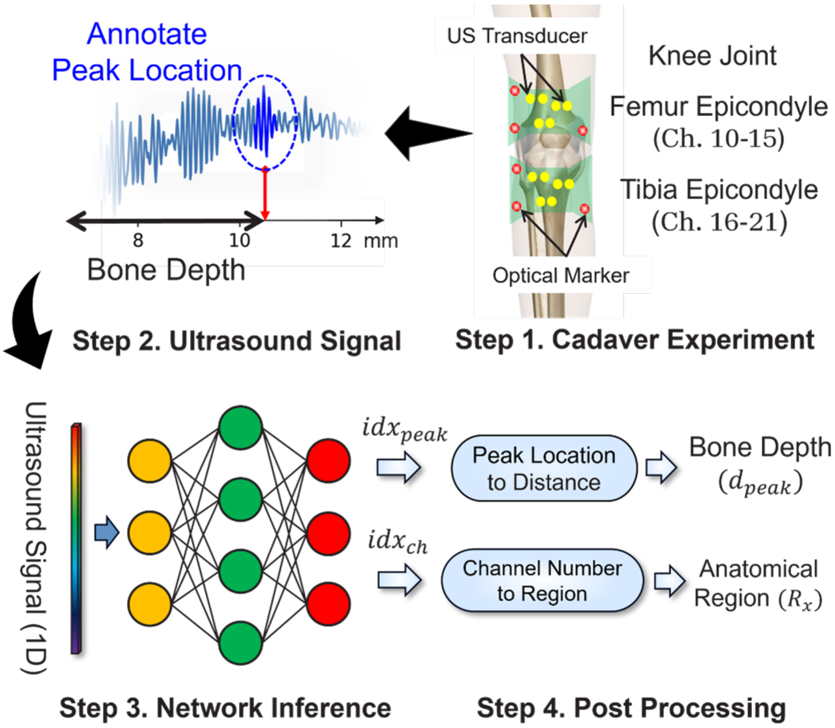


Figure 3: Graphical Abstraction of the Chapter 4

Anatomical Region Perception and Real-time Bone Tracking Methods by Dynamically Decoding A-Mode Ultrasound Signals

Bangyu Lan[1], Stefano Stramigioli[1] and Kenan Niu[1]

***Abstract*— Accurate bone tracking is crucial for kinematic analysis in orthopedic surgery and prosthetic robotics. Traditional methods, such as skin markers, are subject to soft tissue artifacts, and the bone pins used in surgery introduce the risk of additional trauma and infection. For electromyography (EMG), its inability to directly measure joint angles requires complex algorithms for kinematic estimation. To address these issues, A-mode ultrasound-based tracking has been proposed as a non-invasive and safe alternative. However, this approach suffers from limited accuracy in peak detection when processing received ultrasound signals. To build a precise and real-time bone tracking approach, this paper introduces a deep learning-based anatomical region perception tracking method for A-mode ultrasound signals. Simultaneously, it is capable of identifying the corresponding anatomical region to which the A-mode ultrasound is attached. This model contains the fully connection between all encoding and decoding layers of the cascaded U-Nets to decode only the signal region that is most likely to have the bone peak, thus pinpointing the exact location of the peak and classifying the anatomical region of the signal. The experiment showed a 97% accuracy in the classification of the anatomical regions and a precision of around 0.5±1 mm under dynamic tracking conditions for various anatomical areas surrounding the knee joint. This method shows great potential beyond the traditional method, in terms of the accuracy achieved and the perception of the anatomical region where the ultrasound has been attached as an additional functionality.**

I. INTRODUCTION

Bone tracking technology is essential for the kinematic analysis of the human body. The highly precise tracking produces accurate kinematics data, vital for surgical procedures [1], prosthetic robotics [2], and wearable exoskeletons [3]. Typically, the gold standard of tracking is achieved by using bone pins with optical markers [4], but it introduces invasive procedures and infection risks to subjects. Another method is electromyography (EMG)-based techniques [5], [6], [7], [8], but indirect measurement based on muscle activation patterns requires complex algorithms to analyze kinematics. In this context, a more accurate and convenient approach is preferable to obtain the knee kinematics in a non-invasive manner.

Recently, an A-mode ultrasound (US) based tracking method has been introduced as a solution [9]. Compared to B-mode US, A-mode US can perform bone tracking in real-time, without the receiving and processing time of 2D images, and the need to analyze medical images by experts. Compared with other tracking techniques, A-mode US is safe, noninvasive, and cost-effective. However, its accuracy and robustness are compromised due to the reliance on traditional peak detection to analyze one-dimensional raw US signals. The automatic peak detection algorithm from traditional signal processing theory falls short in extracting comprehensive information from the raw A-mode ultrasound signals.

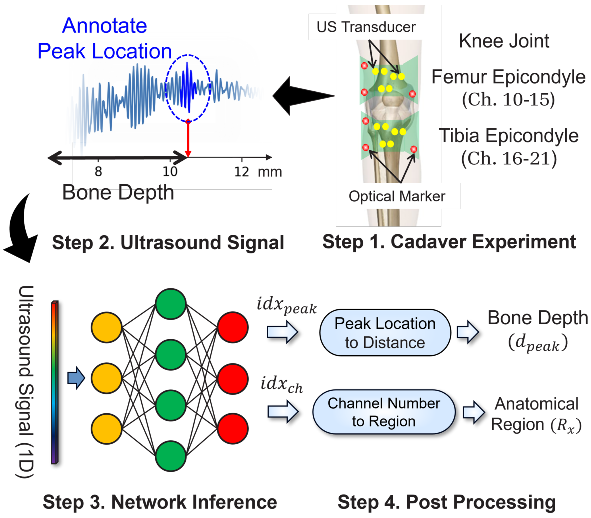


Fig. 1. Steps to build our method: A cadaver experiment to get the dataset and annotate the bone peak location. Our network is trained by the dataset and infers the actual bone depth and anatomical region.

In related research fields, deep learning has been employed for signal peak recognition. For example, in the diagnosis of heart disease, a U-Net framework was developed to segment and identify meaningful peaks in raw EEG signals [10]. However, the A-mode US signal presents a unique challenge: the meaningful bone peaks are actually sparse and ambiguous due to acoustic strength attenuation or the unclear interface between the soft tissues (e.g. tendon and muscles) and the bony surface. For this reason, to our knowledge, few studies have reported using deep learning for A-mode ultrasound diagnosis in knee kinematics tracking or real-time bone tracking. A novel method that considers both the local features (sparse bone peak) and the general features (entire ultrasound echo signal) of the A-mode US signal is needed to address the challenge.

In this study, we focused on tracking the knee joint, a challenging area for traditional tracking methods, especially around the femur epicondyle and the tibia epicondyle, where curved skin and bone surfaces are associated. Around the knee joint, curvatures of the skin and bone surfaces lead

[1] Robotics and Mechatronics, University of Twente, Enschede, AE, The Netherlands

to challenges of distance measurements and inaccuracies. The powerful capability of deep learning to extract abstract features can probably enhance the precision of peak recognition in this area. This is because bone peaks, despite their subtle and complex characteristics in geometry, may share underlying similar features that deep learning algorithms can identify and analyze.

Specifically, we employed fully connected cascaded U-Nets with the interconnections determined by the proposed Sampling-based Proposal (SBP), enhancing efficiency and accuracy. Furthermore, an anatomical region classification network in the first U-Net bottleneck layer facilitated the extraction of comprehensive signal knowledge, enabling anatomical region perception. In general, this approach not only achieved high precision in tracking knee joint movement but also extracted anatomical knowledge from the US signal simultaneously. The perceived anatomical region is especially useful after tracking the location of the bone. To have a complete bone registration, accurate predefined landmarks on the bones are required for the correct alignment between 3D scans of bony surface and US transducer locations[4], [11]. The anatomical region detected by our method can provide position calibration for the tracking robot.

In summary, our method simultaneously identifies anatomical regions and performs real-time tracking of the bony surface. It shows great potential for usage in prosthetic robot control and bone or tissue tracking.

II. Method

A. Motion Tracking System

Our motion tracking method began with collecting the knee joint motion dataset from a cadaver specimen [12]. This dataset included the positions of optical markers from bone pins and US holders. The optical markers transformed the actual 3D positions of the US transducers (frame Ψ_1) and bone pins (frame Ψ_2) into the experimental coordinate frame Ψ_3. After rendering their 3D positions relative to the bony surface in the same coordinate frame, the intersections between the bony surface and the directions of the US waveform could be found. The depth of bone under the skin, denoted d, was obtained by calculating the distance between the intersections and the origin of the ultrasound transducers [13]. For each transducer, the corresponding depth distance corresponded to the location of a specific bone peak in the A-mode US signal. This conversion from depth distance to index of bone peak (idx_{peak}) was listed in Equation (1), where $v = 1540m/s$ (the speed of US in soft tissues) and $f_s = 40 \times 10^6$Hz (the US sampling rate). The d_{unit} is the actual length of one unit in the US signal. All bone peak locations in the signals were annotated as the dataset labels to train our fully connected cascaded U-Nets. During testing, the network detected the bone peak position in each signal, which was later converted to the actual depth of the bone for evaluation using the calculated ground truth distance.

$$idx_{peak} = \frac{d}{d_{\text{unit}}} = \frac{d}{\left(2 \times \frac{v}{f_s} \times 1000\right)} \tag{1}$$

B. Overall Structure of the Network

Our method was designed to perceive the anatomical region and perform bone tracking with high accuracy. To this end, a novel structure of fully connected cascaded U-Nets was proposed, which is depicted in Fig. 2. The input of Coarse U-Net was a 1D signal. The output features of the Coarse U-Net bottleneck layer were used to classify the signal channel. The output features from all decoder layers of the Coarse U-Net were linked to the encoder of the Refined U-Net through the Sampling-based Proposal. This strategy pinpointed the dynamic region that was most likely to contain bone peaks. The Refined U-Net output yielded a more precise peak detection (existing as a segment form). The combination of two segments ultimately determined the existence and location of the bone peak.

C. Details of the Structure

1) Cascaded U-Nets: Inspired by [14], our cascaded U-Nets structure was designed to utilize the underlying hierarchical structure within the US signal. The two U-Nets shared a similar structure: each comprised an encoder and a decoder with five layers of dual 1D convolutions. The number of kernels for each convolution was indicated next to the blue boxes in Fig. 2. In the first four layers, the encoder's output was filtered by the feature from the deeper layer through the 1D feature attention block [15] before proceeding to the decoder. This filtering suppressed irrelevant features (peaks resulting from other tissues or noise) and highlighted the salient bone peak-related features. For the Refined U-Net, the encoder's input at each layer was a concatenation between the outputs of Region Cropping and the MaxPooling. The advantage of cascaded U-Nets instead of only one U-Net was the augmented signal scale, which helps to improve the method's perceptual resolution and improve the detection accuracy.

2) Sampling-based Proposal: To determine the region that is most likely to have the bone peak, a Sampling-based Proposal (SBP) inspired by [16] was established between the layers of the coarse U-Net encoder and the refined U-Net encoder. The steps are as follows: Initially, the Coarse U-Net output was converted into the probability of bone peak (p_i^{peak} at the i^{th} location) using SoftMax, serving as a preliminary guess of the location of the bone peak. Subsequently, a candidate region (a sequence of indexes $\{idx_{start}, idx_{start+1}, ..., idx_{end}\}$) was identified around the point of highest probability, the size being three times the width of the final sampled region. Within this candidate area, a Gaussian distribution (GaussianDist($mean, std$)) was generated for each probability point. The cumulative effect of all these Gaussian distributions formed the final sampling distributionSamplingDist, which is found in Equation (2). The final signal region was then sampled using this distribution as the input of the Refined U-Net. Compared to [16], the region of the segment sampled in SBP was also used to crop the features of each output layer of the Coarse U-Net decoder. Note that before cropping, the region was down-sampled first to match the feature resolution in

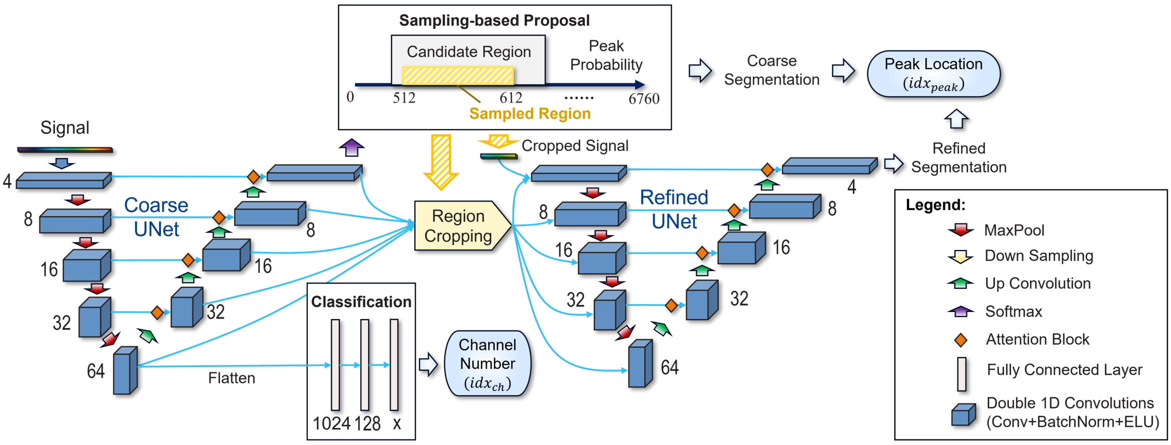


Fig. 2. Our network structure had two U-Nets with different perception scales for bone peaks detection and input signal classification. The input was a 1D ultrasound signal. The outputs were the signal classification results and the prediction of the peak location, which was a segment after thresholding the predicted probability sequence. The result Two U-Nets were connected by Sampling-based Proposal.

the corresponding layer. The outputs of Region Cropping were directly concatenated with the inputs of each layer in the Refined U-Net. Overall, this strategy captured only the essential regions and increased the resolution. Compared to the normal network, this probabilistic approach offered a better perception of the dynamic peak region, facilitating further investigation by the following network.

$$\text{Sampling Dist} = \sum_{i=start}^{end} p_i^{peak} * \text{Gaussian Dist}(idx_i, 1) \tag{2}$$

3) Classification Network: The classification network had three fully connected layers to reduce the dimension. The number of neurons in each layer was specified under the gray rectangle (Fig. 2), where 'x' denoted the number of categories (3 for the femur and 5 for the tibia). LeakyReLU, with a negative slope of 0.1, was used as the activation function between layers. The final classification result was determined by the last layer with a Softmax function. The advantage of linkage between the classification network and the bottleneck layer was that the Coarse U-Net encoder had the capacity to capture the entire signal. This meant that the encoded features from the bottleneck contain comprehensive information, not just bone peaks related but also various soft tissue characteristics, which are crucial for anatomical area identification.

4) Network Output: The network generated two outputs, which are represented as geometric symbols on the light blue background in Fig. 2. The upper right light blue part pinpointed the location of the bone peak, using both coarse and refined segmentation results from two U-Nets. This segmentation was obtained by applying a threshold to the probability of bone peak. The rule for peak determination was based on the priorities of two segmentations: the Coarse segmentation confirmed the existence of a peak, while the Refined segmentation ascertained its precise location. Therefore, a bone peak was considered to exist only if it was indicated by the Coarse U-Net output. Once a bone peak was confirmed to exist, the Refined segmentation was used to determine the exact position. Regarding the bottom-middle light blue box, it gave the signal classification by Argmax on the output of the classification network. This was illustrated in Equation (3), where n is the total number of channels, p_i^{ch} is the probability of the i^{th} channel. Each channel was correlated with the anatomical regions beneath the corresponding US transducer. The anatomical region was characterized by unique subcutaneous tissues, creating distinctive signal characteristics that are useful for classification.

$$R_x \leftarrow idx_{ch} = \arg\max(p_1^{ch}, p_2^{ch}, ..., p_n^{ch}) \tag{3}$$

D. Training Strategy and Post-Processing

To train the network for accurate segmentation, dice loss and cross-entropy loss were used for both the Coarse U-Net l_{dice}, l_{ce} and Refined U-Net l'_{dice}, l'_{ce}. Dice loss [17] can mitigate the problem of sparse foregrounds. This was crucial as over 6760-unit signal length, the peak region spanned merely 10 units, a dimension easily overlooked when relying solely on cross-entropy loss. Equation (4) detailed the dice loss formula. Cross-entropy loss was used as a binary classification for the network to identify the foreground (peak region) or background at each unit. For classification, the training loss l_{cls} was also the cross-entropy loss. The final training loss was in Equation (5). The network was trained by RMSprop optimization [18] with a learning rate of 1e-5, a batch size of 10 [10] and a duration of 50 epoches.

$$DiceLoss(DL) = 1 - \frac{2 \times \sum_{i=0}^{n}(p_i^{pred} * p_i^{true}) + \varepsilon}{\sum_{i=0}^{n} p_i^{pred} + \sum_{i=0}^{n} p_i^{true} + \varepsilon} \tag{4}$$

$$loss = l_{dice} + l_{ce} + l'_{dice} + l'_{ce} + l_{cls} \tag{5}$$

To construct the training dataset, two distinct movements of the knee joint were collected. They were merged and segmented into 2033 samples for all transducer channels. Within the femur epicondyle channels No. 10 to No. 15, only

TABLE I
PEAK DETECTION ACCURACY AND PROPORTION OF THE SUB-MILLIMETER BIAS FOR EACH CHANNEL. THE UNIT IS MILLIMETER.

Area	Channel (Region)	% CLS	Mean	STD	% sub-mm	RMSE
Femur Epi.	11 (R_α)	97.04%	0.434	0.843	84.7%	0.945
	12 (R_β)		0.453	1.201	89.7%	1.279
	15 (R_γ)		0.551	1.322	84.2%	1.428
Tibia Epi.	16 (R_δ)	97.48%	0.582	1.205	87.8%	1.336
	17 (R_ε)		0.604	0.750	87.1%	0.961
	18 (R_ζ)		0.666	0.852	77.7%	1.079
	19 (R_η)		0.312	0.662	89.7%	0.730
	20 (R_θ)		0.683	0.959	77.9%	1.175

channels No. 11, No. 12, and No. 15 exhibited discernible bone peaks; the others were excluded. The signals from the viable channels were truncated if the strengths exceeded 5000. These truncated signals were augmented tenfold by shifting the units on the x-axis. Subsequently, the dataset was divided into training and testing parts in an 8:2 ratio. An identical process was also applied to the tibia epicondyle channels. We shuffled US signals from all channels in the same epicondyle for training and testing, as we assumed that the bone peak in the same area (femur or tibia) exhibited similar profiles.

During post-processing, Equation (6) was used to convert the peak location to the actual depth of the bone. We also verified the anatomical region of the classified channel.

$$d = d_{peak} = idx_{peak} \times d_{\text{unit}} \tag{6}$$

E. Evaluation

To demonstrate the improvement in accuracy, we introduced the traditional method in [12] for comparative analysis. The conventional method of detecting bone peaks involves using expert knowledge to pinpoint the general vicinity of the peak. Within this localized area, a classical peak detection was used to identify the highest peak as the bone peak.

To evaluate our approach, we first collected the bias distance between the positions of the predicted peaks (the peak position was regarded as the middle position of the segmentation) and the ground truth peaks. Then the mean and standard deviation of bias were calculated. Outliers that were much divergent from most biases were analyzed by examining the corresponding 3D position of the knee joint and the US waveform, which is shown in Fig. 4.

We also recorded the network inference time to determine the speed of our method.

III. RESULT

TABLE I is the quantitative results, where CLS is the classification accuracy, STD is the standard deviation, Epi. refers to Epicondyle. Mean, STD, and RMSE are the statistic results of the bias distance. It showed that our method achieved an approximate classification accuracy 97% in both the femur and the tibia epicondyles. For the bias distance, the accuracy was approximately 0.5 ± 1 mm, with the RMSE at around 1.1 mm. In contrast, the traditional method, referenced in [12], showed the best accuracy of only 2.835 ± 1.893 mm. To have a visual overview of the bias, the bias histogram was plotted for all channels in Fig. 3. A single continuous movement sequence was used for the analysis to ensure consistency. The proportion of samples as the submillimeter accuracy is between 80% and 90% for most channels.

For an in-depth examination of the outliers, Fig. 4 presents two situations that contain both a large and a small bias in channel 12. In the top row, an outlier was carefully analyzed in which the prediction deviated by 5.87mm from the ground truth location. The corresponding 3D position and waveform at this moment were also presented. Similarly, plots representing low-error scenarios are provided in the bottom row, offering a balanced view to investigate the method's performance.

Except for these experiments, the network inference speed is recorded, which is 15 ms per batch using the normal laptop (Intel i7-10875H CPU, GeForce RTX 2080 Super Max-Q designed GPU, and 32GB RAM, 2TB SATA SSD). This means that our method has a rapid peak detection speed of 15 ms per signal.

IV. DISCUSSION

In this paper, we proposed a deep learning based method that utilized A-mode ultrasound (US) for measuring bone depth and identifying the anatomical regions during bone tracking. The network could process data in 15ms per signal, which meant that our method is capable of processing ultrasound signal in real-time while obtaining the anatomical regions as extra information.

Our method has been improved upon the CasAtt-UNet [16] by fully integrating cascaded dual U-Nets in each layer and modifying Sampling-based Proposal (SBP) structure for an end-to-end training. This enhancement streamlined the learning process and simplified training. Additionally, the integration of a classification network offered deeper insights into the 1D signal and validates the discriminating features learned by the Coarse U-Net encoder. With the ability to identify anatomical regions, our method offers an option for position calibration, which is critical for precise landmark registration during post-processing in bone tracking [19], [20]. This feature is particularly beneficial in computer-assisted and robotic-assisted orthopedic surgeries, such as Total Knee Arthroplasty (TKA) [21], where accurate bone tracking is vital for kinematic analysis and disease diagnosis. Furthermore, in applications like exoskeletons [22] and Human-Robot Interaction [23], accurate position calibration is essential to ensure correct location sensing, thus facilitating task completion.

In addition, our study had a limitation due to the use of a single cadaver specimen. Inclusion of specimens with varied human characteristics such as gender and age would enhance the robustness of our results. To mitigate this limitation, we gathered two datasets that were recorded at different postures and times of the day. In data augmentation, we pre-processed

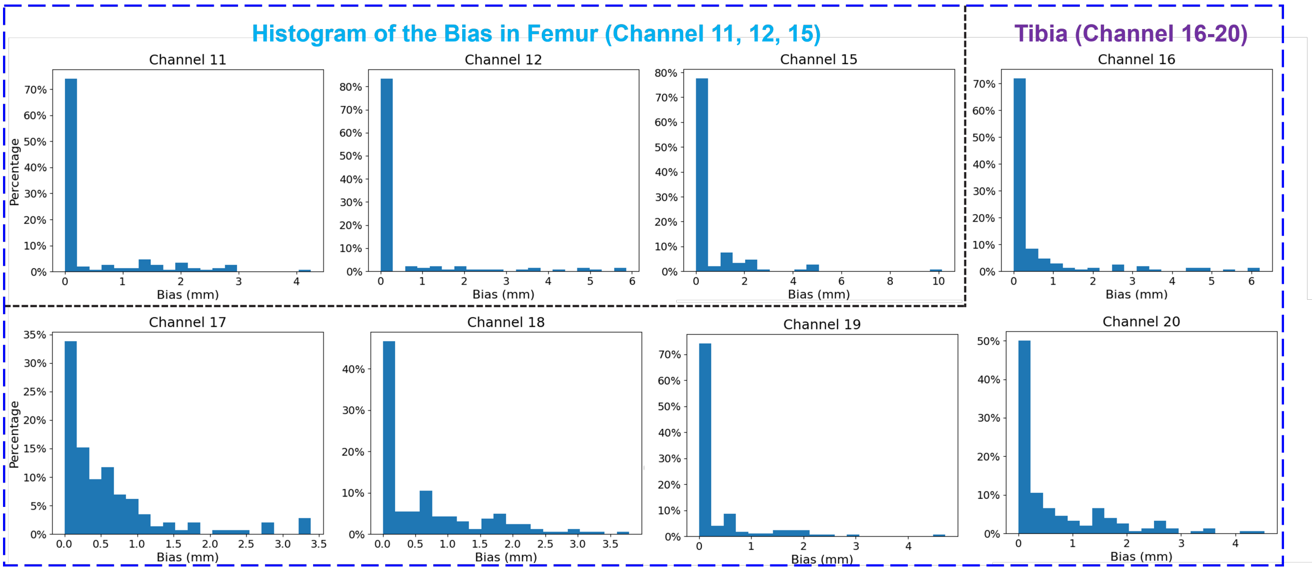


Fig. 3. The histogram of bias between the predicted and the ground truth bone peaks for each channel. They are all in the same dataset, a continuous movement sequence.

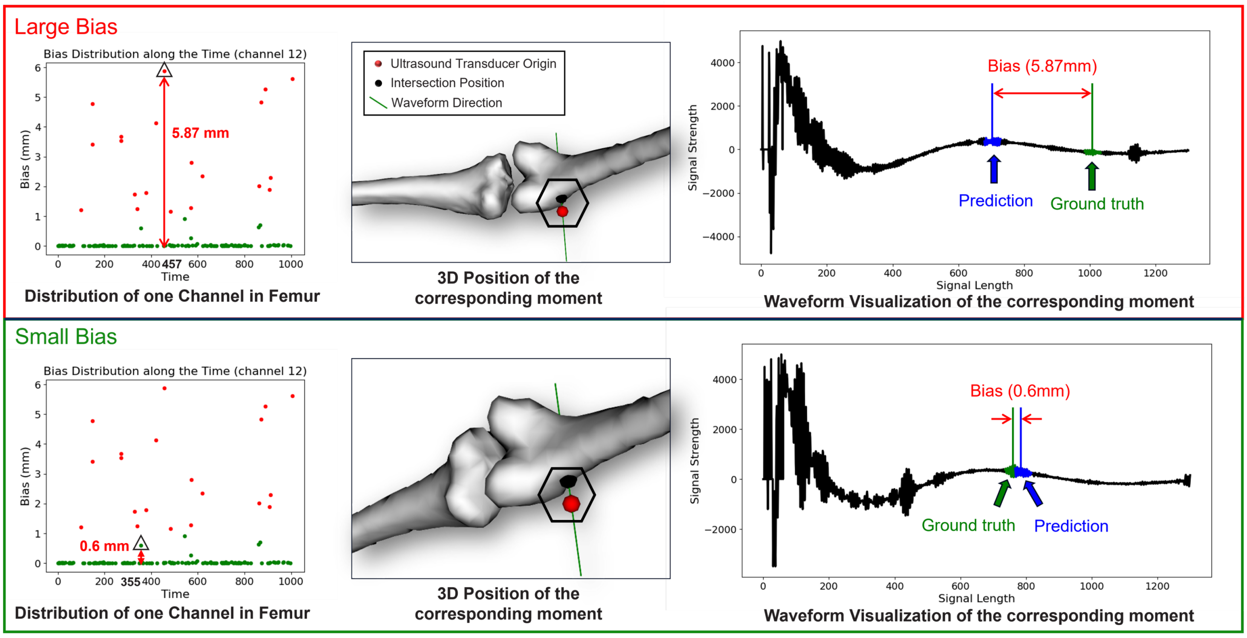


Fig. 4. Closer look at the large and small bias. The distribution of bias along the time (across the entire 1017 samples) is plotted on the left. The 3D position of knee joint in the middle was the same moment of the specified bias. In the right waveform figure, the bias was visually showed.

these datasets by slicing and shuffling them to demonstrate the validity and generalizability of our approach.

When looking at the result, the bias between the predicted and ground truth peaks was approximately 0.5 ± 1 mm in the TABLE I, indicating that our method achieved a sub-millimeter accuracy for most cases, significantly surpassing the accuracy of the traditional method of 2.835 ± 1.893 mm, as derived from [12]. In addition, there are slight variations across channels, possibly due to the different characteristics of the soft tissue surrounding the knee joint. However, these variations also provided unique signal characteristics beneficial for classification, resulting in a high accuracy rate of 97%. This highlights the efficacy of our anatomical-aware bone tracking approach, finely classifying anatomical regions during bone peak detection.

To have a closer look at the large and small errors that occurred in the femur epicondyle, we conducted a detailed analysis of the scenarios in channel 12, with the findings presented in Fig. 4. In the 3D position of the top row, we observed that at the 457^{th} moment, the right leg was transitioning from extension to flexion. During this phase, the labeled ground truth was situated in the green segment of the waveform. Probably the peak profile in the predicted location resembling the actual bone peaks in the training dataset, the network mistakenly identifies the bone peak. It is important to note that there are several possible reasons for the attenuated bone peaks: (1) the curvature of the skin at that moment could lead to loss of skin contact with the transducers, leading to incorrect ground truth calculation and labeling, and (2) the specific posture of the specimen (fixed on the surgical table) at that moment might cause an unusual distribution of the soft tissues, attenuating the bone peaks. However, these potential causes are required to be investigated later. In contrast, the bottom row of this figure illustrates a case of small error, where the leg was in an extension position. In the vicinity of the waveform, the bone peak has an apparent shape without other similar-strength peaks presented nearby to interfere with peak recognition. This makes it easier to accurately identify the bone peak in the signal.

V. CONCLUSIONS

In this study, we introduced an anatomical region perception tracking method capable of performing real-time bone tracking for knee kinematics measurement. This method significantly exceeds traditional A-mode ultrasound techniques in bone measurement accuracy. Additionally, it demonstrates a high rate of anatomical classification accuracy, thereby enabling precise identification of anatomical regions. Our approach makes the A-mode ultrasound a safe and non-invasive alternative which can be used not only in accurately identifying the anatomical region from the signal, but also in precisely tracking bone movements. Potentially, our approach not only enhances the current capabilities of A-mode ultrasound but also paves the way for its future integration into robotics and prosthetic systems, promising advancements in accurate kinematics measurements that provide real-time feedback for precise robot control.

REFERENCES

[1] C. Li, Z. Zhang, G. Wang, C. Rong, W. Zhu, X. Lu, Y. Liu, and H. Zhang, “Accuracies of bone resection, implant position, and limb alignment in robotic-arm-assisted total knee arthroplasty: a prospective single-centre study,” *Journal of Orthopaedic Surgery and Research*, vol. 17, no. 1, p. 61, 2022.

[2] K. R. Embry and R. D. Gregg, “Analysis of continuously varying kinematics for prosthetic leg control applications,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 262–272, 2020.

[3] T. B. Meier, N. A. Goldfarb, C. J. Nycz, and G. S. Fischer, “Evaluating knee exoskeleton design based on movement with respect to underlying bone structure using mri,” *IEEE Transactions on Medical Robotics and Bionics*, 2023.

[4] K. Niu, J. Homminga, V. I. Sluiter, A. Sprengers, and N. Verdonschot, “Feasibility of a-mode ultrasound based intraoperative registration in computer-aided orthopedic surgery: A simulation and experimental study,” *Plos One*, vol. 13, no. 6, p. e0199136, 2018.

[5] N. Lotti, M. Xiloyannis, G. Durandau, E. Galofaro, V. Sanguineti, L. Masia, and M. Sartori, “Adaptive model-based myoelectric control for a soft wearable arm exosuit: A new generation of wearable robot control,” *IEEE Robotics & Automation Magazine*, vol. 27, no. 1, pp. 43–53, 2020.

[6] K. C. Tse, P. Capsi-Morales, T. S. Castaneda, and C. Piazza, “Exploring muscle synergies for performance enhancement and learning in myoelectric control maps,” in *2023 International Conference on Rehabilitation Robotics (ICORR)*, 2023, pp. 1–6.

[7] A. J. Young, L. H. Smith, E. J. Rouse, and L. J. Hargrove, “Classification of simultaneous movements using surface emg pattern recognition,” *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 5, pp. 1250–1258, 2013.

[8] M. A. Díaz, S. De Bock, P. Beckerle, J. Babič, T. Verstraten, and K. De Pauw, “An emg-based objective function for human-in-the-loop optimization,” in *2023 International Conference on Rehabilitation Robotics (ICORR)*, 2023, pp. 1–6.

[9] K. Niu, V. Sluiter, J. Homminga, A. Sprengers, and N. Verdonschot, “A novel ultrasound-based lower extremity motion tracking system,” *Intelligent Orthopaedics: Artificial Intelligence and Smart Image-guided Technology for Orthopaedics*, pp. 131–142, 2018.

[10] V. Moskalenko, N. Zolotykh, and G. Osipov, “Deep learning for ecg segmentation,” in *Advances in Neural Computation, Machine Learning, and Cognitive Research III: Selected Papers from the XXI International Conference on Neuroinformatics, October 7-11, 2019, Dolgoprudny, Moscow Region, Russia*. Springer, 2020, pp. 246–254.

[11] P. M. B. Torres, J. M. Sanches, P. J. S. Gonçalves, and J. M. M. Martins, “3d femur reconstruction using a robotized ultrasound probe,” in *2012 4th IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, 2012, pp. 884–888.

[12] K. Niu, T. Anijs, V. Sluiter, J. Homminga, A. Sprengers, M. A. Marra, and N. Verdonschot, “In situ comparison of a-mode ultrasound tracking system and skin-mounted markers for measuring kinematics of the lower extremity,” *Journal of biomechanics*, vol. 72, pp. 134–143, 2018.

[13] K. Niu, J. Homminga, V. Sluiter, A. Sprengers, and N. Verdonschot, “Measuring relative positions and orientations of the tibia with respect to the femur using one-channel 3d-tracked a-mode ultrasound tracking system: A cadaveric study,” *Medical engineering & physics*, vol. 57, pp. 61–68, 2018.

[14] H. Liu, X. Shen, F. Shang, F. Ge, and F. Wang, “Cu-net: Cascaded u-net with loss weighted sampling for brain tumor segmentation,” in *Multimodal Brain Image Analysis and Mathematical Foundations of Computational Anatomy: 4th International Workshop, MBIA 2019, and 7th International Workshop, MFCA 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Proceedings 4*. Springer, 2019, pp. 102–111.

[15] P.-H. Chen, C.-H. Huang, W.-T. Chiu, C.-M. Liao, Y.-R. Lin, S.-K. Hung, L.-C. Chen, H.-L. Hsieh, W.-Y. Chiou, M.-S. Lee, *et al.*, “A multiple organ segmentation system for ct image series using attention-lstm fused u-net,” *Multimedia Tools and Applications*, vol. 81, no. 9, pp. 11 881–11 895, 2022.

[16] B. Lan, M. Abayazid, N. Verdonschot, S. Stramigioli, and K. Niu, “Deep learning based acoustic measurement approach for robotic applications on orthopedics,” in *2024 International Conference on Robotics and Automation (ICRA)*. IEEE, in press.

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

[18] R. Elshamy, O. Abu-Elnasr, M. Elhoseny, and S. Elmougy, “Improving the efficiency of rmsprop optimizer by utilizing nestrove in deep learning,” *Scientific Reports*, vol. 13, no. 1, p. 8814, 2023.

[19] M. A. Price, P. Beckerle, and F. C. Sup, “Design optimization in lower limb prostheses: A review,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 8, pp. 1574–1588, 2019.

[20] P. Henrich, B. Gyenes, P. M. Scheikl, G. Neumann, and F. Mathis-Ullrich, “Registered and segmented deformable object reconstruction from a single view point cloud,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 3129–3138.

[21] A. Mannan, J. Vun, C. Lodge, A. Eyre-Brook, and S. Jones, “Increased precision of coronal plane outcomes in robotic-assisted total knee arthroplasty: A systematic review and meta-analysis,” *the surgeon*, vol. 16, no. 4, pp. 237–244, 2018.

[22] G. Rinaldi, L. Tiseni, M. Xiloyannis, L. Masia, A. Frisoli, and D. Chiaradia, “Flexos: A portable, sea-based shoulder exoskeleton with hyper-redundant kinematics for weight lifting assistance,” in *2023 IEEE World Haptics Conference (WHC)*. IEEE, 2023, pp. 252–258.

[23] E. Galofaro, E. D'Antonio, N. Lotti, F. Patané, M. Casadio, and L. Masia, “Bimanual motor strategies and handedness role in human-robot haptic interaction,” *IEEE Transactions on Haptics*, vol. 16, no. 2, pp. 296–310, 2023.

5 A Dual-Attention Framework to Decipher Muscle Dynamics

The final chapter transfer the focus from bone to muscle, as the distance between the skin and bone actually reflect the contraction degree of muscle in a quantitative way. This chapter introduce a different deep learning framework based on hierarchical attentions that predicts the muscle thickness deformations merely from sEMG (surface electromyography) data. By integrating the self-attention and cross-attention mechanisms, this approach eliminates the need for the heavy and complex ultrasound imaging or magnetic resonance imaging (MRI), facilitating the real-time, wearable, and portable daily monitoring of the muscle health. Combining with the muscle activation measurement from the sEMG device itself, this technique shows great potential to bring clinical diagnostics or rehabilitation evaluation to the user's daily life.

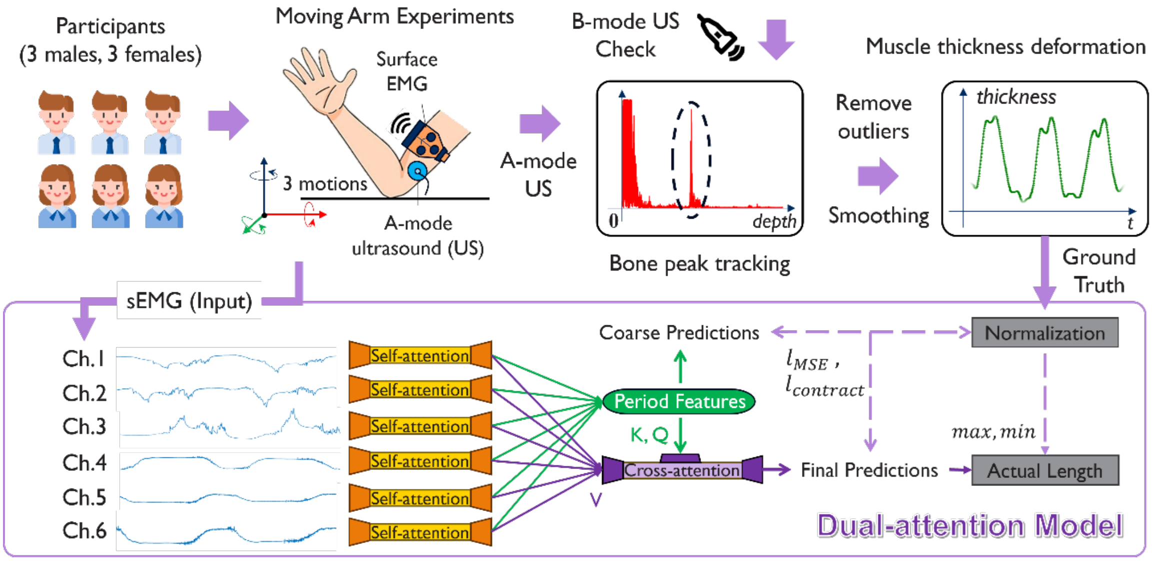


Figure 4: Graphical Abstraction of the Chapter 5

Deciphering Muscular Dynamics: A Dual-Attention Framework for Predicting Muscle Contraction from Activation Patterns

Bangyu Lan, *Student Member, IEEE*, Stefano Stramigioli, *Fellow, IEEE*, and Kenan Niu, *Member, IEEE*

***Abstract*—Quantitatively deciphering the relationship between muscle activation and thickness deformation is essential for diagnosing muscle-related diseases and monitoring muscle health (e.g., Facioscapulohumeral Dystrophy). Despite the potential of ultrasound (US) imaging and sensing could measure the changes in muscle thickness during movements, it remains challenging to make a fully portable device, considering the wires and data collection. On the other hand, surface electromyography (sEMG) is capable of recording muscle bioelectronical signals and measuring muscle activations, which offers a unique perspective to correlate with the underlying muscle thickness changes. This paper introduced a deep learning-based approach that leveraged sEMG signals to infer muscle thickness deformation. Utilizing a hierarchical combination of self-attention and cross-attention mechanisms, this method predicted muscle deformation directly from sEMG data, eliminating the dependency on applying ultrasound imaging techniques. Experimental results on six healthy subjects indicated that our approach could accurately predict muscle thickness deformation with an average precision of 0.923±0.900 mm, showcasing substantial benefits in measuring muscle thickness deformation only by sEMG device. This technique facilitates real-time portable muscle health monitoring by sEMG to provide not only bioelectronical signals but also biomechanical information. It indicates the great potential of utilizing such a technique in clinical diagnostics, sports science, and rehabilitation.**

***Index Terms*—Ultrasound, Surface EMG, Muscle deformation, Muscle activation, Muscular dynamics, Dual-Attention, Deep learning**

I. Introduction

QUANTITATIVE relationship between muscle activation and thickness deformation depicts muscular dynamics and reveals muscle health status, which is prominent for disease diagnosis. In Facioscapulohumeral Dystrophy (FSHD), loss of muscle strength and deformation are due to the fatty infiltration and fibrosis of the histopathological changes [1]. Quantitatively measuring muscle thickness variations and tracking muscle activation is pivotal for understanding fundamental aspects of the disease and muscle mechanics [2]. It also helps to diagnose and monitor muscle pathological conditions in an early stage [3]. However, a portable and wearable device is lacking, which could measure muscle activation and thickness deformation simultaneously [4]. Thus, it fails to offer a grand view of muscle status in a convenient and low-cost way.

In traditional methods, ultrasound (US) images directly visualize muscle thickness [5], while surface electromyography (sEMG) device demonstrates the ability to capture muscle activation patterns [6]. Interpreting the two sources to summarize the muscle health status requires expert knowledge, difficult for users to have a portable daily tracking tool [7]–[9]. Moreover, integrating the A-mode US transducers and sEMG into a compact and wearable device remains challenging. The alternatives solutions like magnetic resonance imaging (MRI) are impractical for real-time and daily applications due to their size, cost, and operational complexity [10]. Therefore, it is hard to reach a wide spectrum of accessibility for users to track and monitor normal muscle health status in daily life.

Aiming at extending the wider understanding of muscle health status by a portable and convenient device, an alternative solution is to infer muscle thickness deformation quantitatively from the muscle activation pattern, derived by sEMG signals. The sEMG device records the muscle activities during muscle contraction and extension [11]. In repeated motion periods, the muscle electrode signals show specific patterns that can be quantitatively measured and analyzed [12]. Therefore, the patterns from the surface electromyography signals strongly correlate with the mechanical behavior of muscles [13], suggesting a viable pathway to infer muscle thickness changes [14]. This capability, combined with the portability and convenience of the sEMG device, makes it a potentially promising alternative for continuous and daily tracking of muscle thickness deformation (MTD).

To infer the MTD from sEMG signal, we introduce a novel dual-attention based approach to quantitatively correlate muscle activation and thickness deformation to decipher the muscle dynamics. By utilizing both self-attention [15] and cross-attention [16] mechanisms hierarchically, this framework aims to precisely predict muscle contraction solely from the sEMG data, thereby eliminating the requirement for the ultrasound imaging during muscle activity assessment, and enable portable and real-time muscle health tracking.

This paragraph of the first footnote will contain the date on which you submitted your paper for review. It will also contain support information, including sponsor and financial support acknowledgment. For example, “This work was supported in part by the U.S. Department of Commerce under Grant 123456.”

Bangyu Lan, Stefano Stramigioli and Kenan Niu are with the Robotics and Mechatronics group, the Faculty of EEMCS in University of Twente, P.O. Box 217, 7500 AE Enschede, Netherlands. (e-mail: bangyu@ieee.org, S.Stramigioli@ieee.org, and k.niu@utwente.nl)

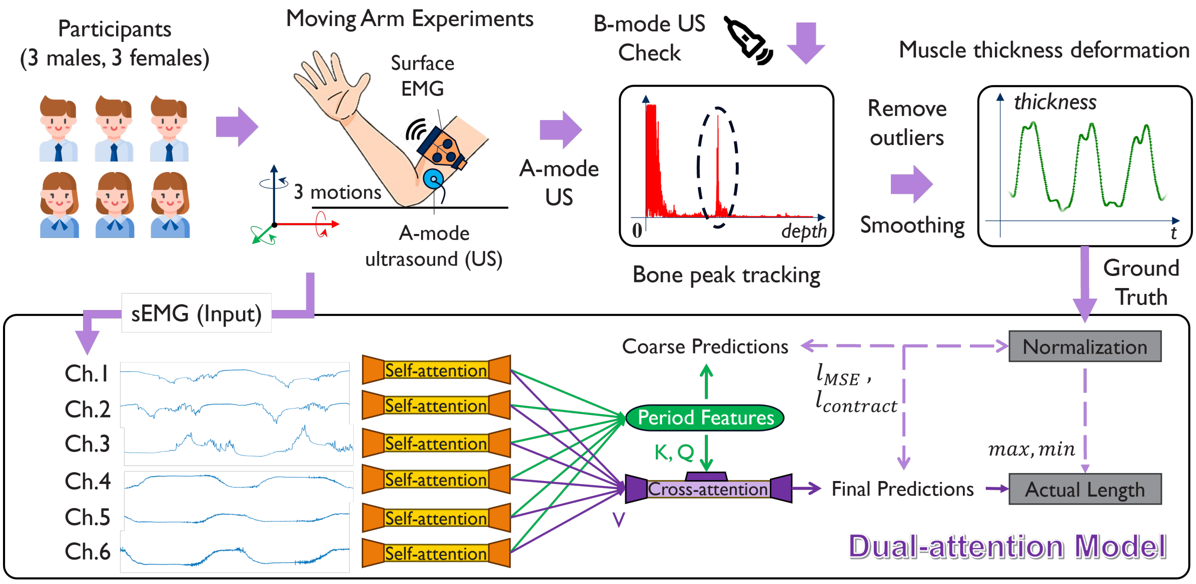


Fig. 1. The overview of the proposed system from data acquisition at the experiments to the network structure of the dual-attention model. The experiment subjects were six persons having different Body Mass Index (BMI). They were required to perform forearm movements in several motion patterns for sEMG and US signals recording. The sEMG device recorded muscle activation as the sole input of the dual-attention model, while the US signals were preprocessed to recover the actual muscle thickness deformation as the model's target. The dual-attention network extracts periodic features and gives accurate muscle thickness deformation measurement.

TABLE I
EXPERIMENTS PARTICIPANTS INFORMATION (MTC: MUSCLE THICKNESS CONTRACTION)

Subject	Gender	BMI	MTC range (mm)	MTC distribution (mm)
A	Male	20,76	[1.94, 11.35]	6.45 ± 2.54
B	Female	20.96	[3.07, 11.08]	6.33 ± 1.51
C	Male	22,28	[1.36, 12.48]	7.09 ± 2.87
D	Female	25,22	[2.06, 12.84]	6.12 ± 2.81
E	Male	24.62	[3.26, 16.27]	9.82 ± 3.49
F	Female	26.44	[2.05, 14.11]	7.46 ± 2.83

In our experiments, the A-mode ultrasound signals provided the quantitative ground truth record of the muscle contraction, and the sEMG signal worked as the sole caused signal that sparks the muscle thickness deformation. A self-attention structure is leveraged for each sEMG channel signal to extract muscle movement periodic patterns. A cross-attention mechanism was built upon on the six channels features for the considerable and more accurate thickness deformation prediction. The experiments were performed on six subjects to validate model's universality, generalizability, and robustness. The ablation study was performed to research the effects of proposed losses and network structure designs on the method's performance.

To the best of our knowledge, few study inferred muscle thickness deformation solely from sEMG signals. The proposed method demonstrates the substantial potential to enhance muscle contraction analysis and provide new insights for how muscle health can be monitored in the clinical settings, sports science, and rehabilitation. Further developments could explore the integration of additional biometric data [17], enhancing the model's applicability and accuracy in the real-world scenarios.

II. MATERIAL AND METHOD

In this section, the experiment to collect the representative dataset and the dual-attention framework to predict the muscle thickness deformation were described.

A. Overview of the system

The whole pipeline of the experiment and the prediction model was shown in Fig. 1. Six participants having different BMI indexes (see TABLE I) were invited to perform three types of arm motions representing the daily activities. Each motion was performed repeatedly to obtain different muscle contraction levels measured by the A-mode US. Meanwhile, the sEMG device collected muscle electrode signals from the nearby muscle positions as the muscle activation patterns. The sEMG and A-mode US devices in the experiment were demonstrated in Fig. 2. The sEMG was a 3D-printed, eight channels, and dry electrodes measurement device based on the custom-developed amplifier Octopus [18]. Of the eight channels of the Octopus, six channels were connected to the biceps and triceps electrodes, while the other two were used for measuring the average voltage and the ground level voltage [19]. After negating the ground level voltage, the six channels voltage were used together to record subjects' biceps and triceps muscle electronics signals. The A-mode ultrasound (US) used a ultrasonic testing device, OPBOX ver 2.1 (OPTEL Ultrasonic Technology, Wrocław, Poland). The device was set to 100 MHz sampling frequency, 4 to 25 MHz bandwidth analog filters, +30dB constant gain, and +24 dB pre-amplifier to increase the raw signal strength. The x-axis

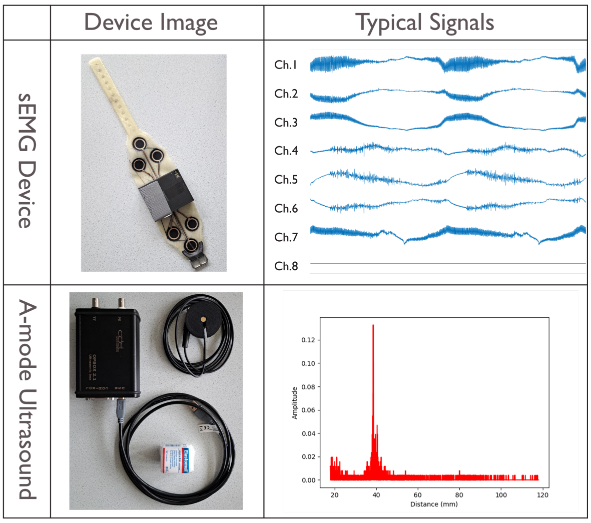


Fig. 2. The setup and typical signals of the sEMG device and A-mode ultrasound device. The sEMG device contained three electrodes for biceps and three electrodes for triceps. In the recorded signals, the channel 1 to 6 were used for network training. The ultrasound data specified the bone location and indicated the movement of bone, which reflected the muscle thickness deformation

(time-domain) of the US signal has been shifted to a personal-related amount of time to focus only on the area that the bone peaks most likely to appear. The recorded A-mode US signals were compared with the B-mode US images to verify the actual muscle contraction range, which was then used as the prediction targets of the network. The sEMG signals, on the other hand, worked as the input of the model.

The model to predict the muscle thickness deformation apply a dual-attention structure (see bottom of Fig. 1), inspired by the foundational principles of transformer architecture [20], [21]. It contained attention mechanisms operated at different hierarchical levels. The self-attention mechanism encoded six channels of sEMG signals as separate modality features. Each channel feature encoded local muscle contraction patterns. The combination of all features encoded the entangled and correlated biceps and triceps muscle contractions in the upper arm, which was then decoded by the cross-attention mechanism to predict the muscle thickness deformation more accurately.

B. Human-Related Muscle Contraction Experiment

To validate the universality and generalizability of the approach thoroughly, a representative dataset, including diverse personal physical characteristics (e.g., BMI) and different muscle movements, was collected. The details were illustrated in TABLE I. Three male and three female participants with different BMI and muscle thickness contraction (MTC) joined the experiment. This experiment had been approved by Ethics Committee Information & Computer Science of the University of Twente. All participants had signed the consent forms. Before the experiment setup, the A-mode US signals from each subject had been checked with the B-mode US images to locate the general range of bone peak positions in the upper arm humerus among all visible peaks (Fig. 3). The bone peak location reflects the muscle thickness between the skin and bone [22], [23], indicative of the quantitative degree of the muscle thickness deformation. Before performing arm movements, each subject wear A-mode US and sEMG devices on the upper arm with the help of the researchers (Fig. 4). The measuring position of the A-mode US was above the elbow on the inside of the right upper arm. This was because when the forearm moved, the muscle contraction near the elbow displayed large bone peak movement in the A-mode US, which facilitated the easy measurement and the muscle movement tracking. The sEMG device used to trace muscle activation was fixed in the nearby middle of the upper arm to collect muscle electrode signals from the right biceps and triceps brachii. As the muscle activation and thickness deformation of the upper arm was highly correlated, the sEMG signal could be used to predict the muscle deformation on specific position.

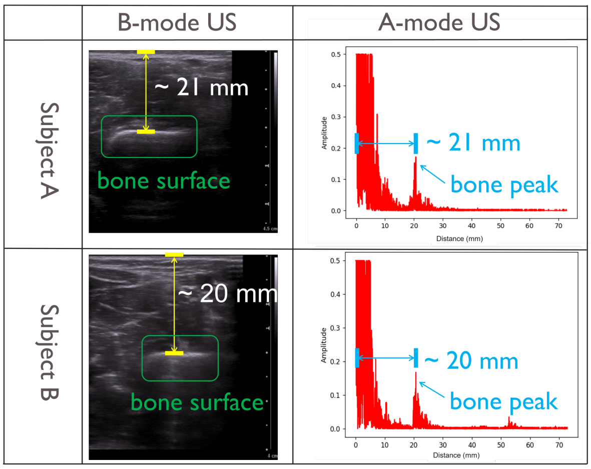


Fig. 3. The A-mode US bone peak verification using the B-mode ultrasound. This verification was performed on the same position of the A-mode ultrasound. The B-mode ultrasound visualized the relative positions between bone surface and the skin, indicative of the muscle thickness. This distance is the same as the bone peak location.

After the preparation was finished, the subjects started to perform three types of forearm motions. There are totally three stages recorded in different daily time. The first two stages are the same, while the last stage involved a 500 grams weight carrying in the right hand. Each forearm motion is performed repeatedly for five minutes. Each single motion period was five seconds, except for the ten seconds of the motion C. They were all performed by rotating the forearms around the touching position of the elbow (see Fig. 5), where the motion A rotated around the y-axis, the motion B around x-axis, and the motion C rotated the forearm around the 135° between x-axis and z-axis. The three motions were performed continuously, with the order from motion A, B, to C. After one stage had been finished, the volunteer was forced to take a rest for 5 to 30 minutes and prepare the next stage. In the end, both muscle electrode signals (sEMG) and muscle thickness deformation (US) were collected from all the three stages of the six people at different daily time.

C. Ultrasound and sEMG signal processing

To have a proper setting for the training-effective dataset collecting, the US signals were recorded at 30Hz, while sEMG

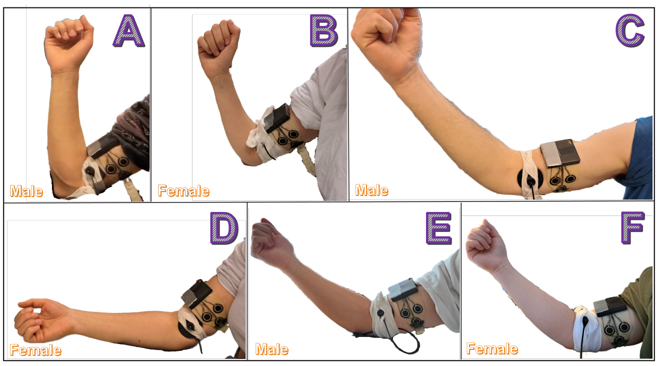


Fig. 4. The experiment setups for six participants. The ultrasound probe was attached above the elbow, while the sEMG device was attached on the middle of the upper arm. The background had been removed for a clear visualization.

signals were recorded at 1000Hz. They were synchronized using the timestamps recorded at the same time. Each moment of the US signal was visualized as a 2D plot (see Fig. 1), so that the bone peak locations in the signals could be tracked. The outliers had been removed from the tracking curves, which continuously to be smoothed for the ground truth labels of the muscle thickness deformation (MTD). The muscle thickness contraction (MTC) was the largest range of muscle thickness deformation (MTD) in a single motion period, which was calculated by Equation (1). In each single motion period, MTC equals to the distance between the largest muscle thickness deformation (H_{MTD}) and the smallest muscle thickness deformation (L_{MTD}).

$$MTC = H_{MTD} - L_{MTD} \tag{1}$$

After the muscle thickness deformation being obtained from the US, the sEMG data began to be pre-processed. For each single data-point of the recorded US signals, the surrounding 30 sEMG data-point were concatenated as the corresponding synchronized muscle activation pattern. This aimed to keep as much as raw sEMG signals to remain more useful information. Then the strength of sEMG was increased 500 times to reach the 0.1 unit magnitude. No further complex pre-processing steps were done on the signals, as we believed that even tiny features in the local frame of one period could be considered for the more accurate prediction.

As the motion was performed repeatedly, some regular and repeatable patterns can be visually noticeable from the raw sEMG signals and the pre-processed MTD. Noticed that as there was no requirement for the starting and end positions of subjects' forearms, they could freely moved forearms to different positions in different motion periods. This created the differences of MTD and MTC among each single motion period. For each subject's motion, the MTC range and distribution (mean±standard deviation) was summarized on the last two columns of TABLE I. Comparing the BMI and MTC, we noticed that there was no linear relationship between the MTC and BMI index, although higher BMI index prone to have a higher MTC.

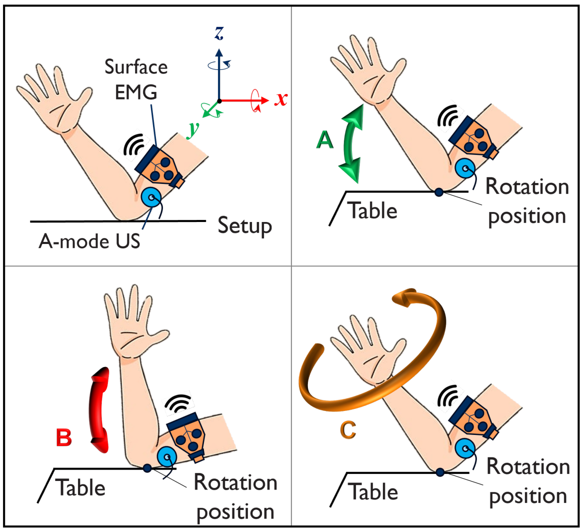


Fig. 5. The three arm motions performed by the subjects. Each motion required forearm rotation around at the elbow. The motion A rotated around the y-axis, motion B around the x-axis, while motion C draw a large circle anti-clockwise in the air. After the hand touched the table, the forearm returned to the original positions then repeatedly performed the same motion again. Each single period of the motion A or B was performed for five seconds, while the single period of motion C was performed for ten seconds.

D. Dual-attention network structure

To extract periodic patterns from sEMG signals to accurately decode muscle thickness deformation, a regression model able to consider all bio-electronics signals from different muscle positions was created. The overview of the dual-attention model was in Fig. 6. Inspired by the transformer structure, a self-attention mechanism was designed as a self-adaptive signal encoder for each sEMG channel signal. The detail was in the middle of the Fig. 6. The assumption was that the raw sEMG signal could be regarded as a special transformed pattern from the pure muscle contraction periods, as it was recorded during the repeated motions. Inside a single period motion, the muscle activation signal within a small range was assume to have some recognizable patterns corresponding to the muscle periodic tiny deformation, critical for recognizing current period position. To capture the local patterns for the accurate periodic position recognition, the self-attention structure adjusts the original sEMG signals to de-noise and de-transform for the motion period recovery, which was achieved by the two-head attention mechanism. This process produced the self-adaptive and periodic features, which shared the same encoding backbone but with different linear operations and output activation functions. The periodic feature was the coarse prediction directly supervised by the ground truth labels.

Based on all sEMG channels features, a cross-attention mechanism was built upon (bottom pipeline of Fig. 6). Since on the forearm all channels signals record the synergy of an entangled muscle (biceps and triceps muscles) movement, each self-attention structure was applied on a single channel as a separate modality encoder. To prevent the situation when one of channel signals contained purely electronic noise, the cross-

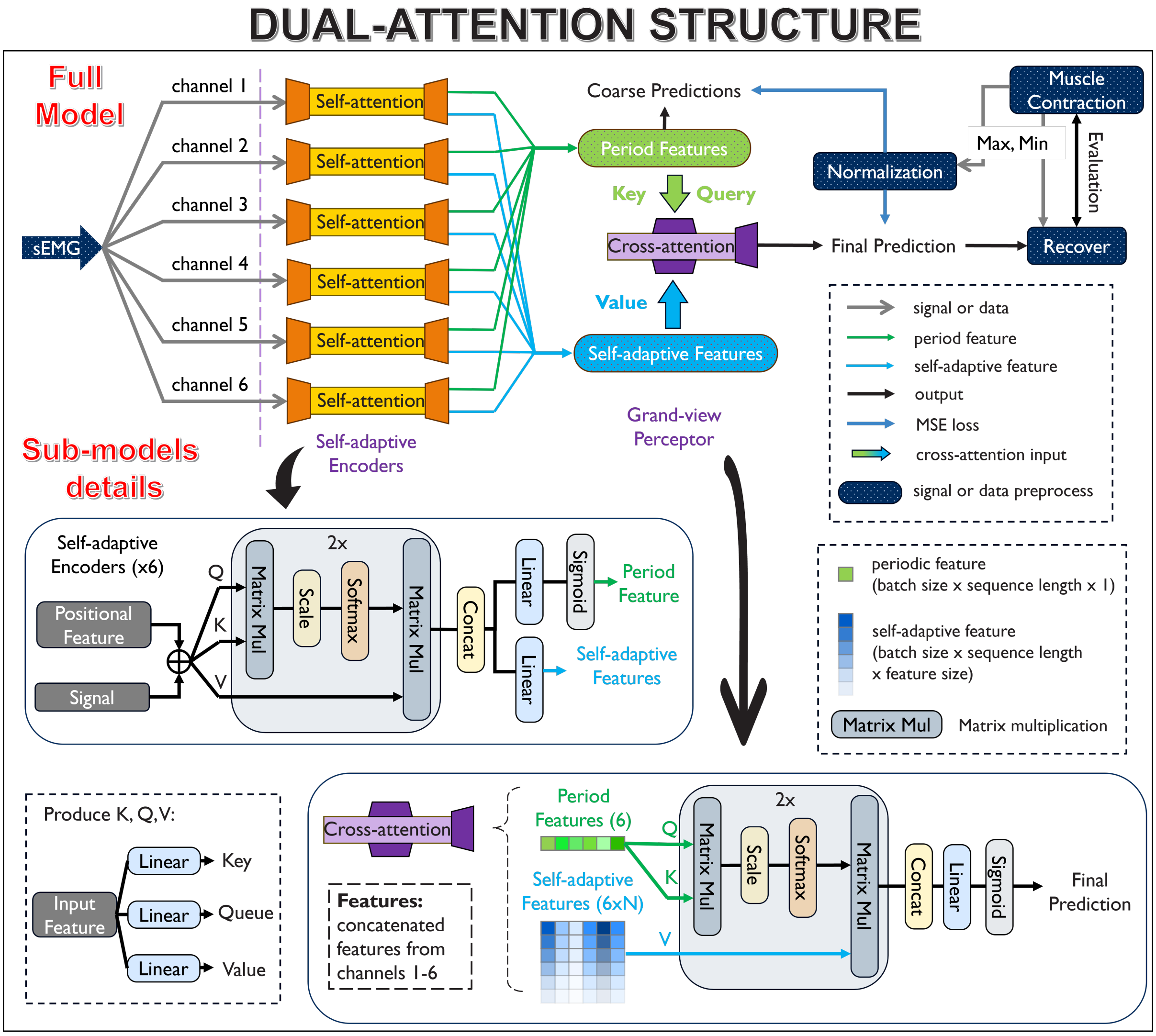


Fig. 6. Overview of the proposed dual-attention network and sub-structures in details. Each sEMG signal channel connected with a self-adaptive encoder (self-attention structure), the encoded periods and self-adaptive features worked as the Key (K), Queue (Q) and Value (V) respectively. Based on the period features, the grand-view preceptor (cross-attention) weighted averaged all self-adaptive features to decode the final prediction. This prediction was de-normalized to recover the actual muscle thickness deformation in millimeters. The bottom two pipelines showed the details of self-attention and cross-attention structures. The input features of cross-attention module were the concatenation of all encoded features from the self-attention structures. Noticed that for all keys, queues and values, they were produced from separated linear operations.

attention perceive all channel features to supplement potential lacking information for a grand-view prediction. This helped to decode the complete muscle contraction movement. To decide the importance in each channel and pay more attention to the specific ones, the periodic features worked as the K (key) and Q (query) of the cross-attention input to guide the attention. While the self-adaptive features, adjusted and de-transformed from the sEMG raw signals, worked as the V (values) to be filtered and decoded. The regression output from the cross-attention was the final prediction of the muscle thickness deformation. The training labels for the coarse and final predictions were the same, but the learning targets were different.

In general, this dual-attention structure predict the muscle deformation by learning the correlations between the changing patterns of sEMG signals and the muscle contraction. To facilitate network's training, the prediction target had been normalized and recovered in the end based on the personal MTC range.

E. Training and validation

To effectively train the network's capability for accurate periodic muscle deformation regression, the input and the ground truth label were set as one period (5 seconds, 150 data-points). Within each period, the MSE loss was adopted to minimize the errors of MTD prediction. The muscle contraction loss ($Loss_c$) was proposed to minimize the muscle thickness contraction, which was defined by Equation (5): In each single motion period, the top 5 and bottom 5 muscle thickness deformation values ($\mathrm{Top}_{5,MTD}$ and $\mathrm{Bottom}_{5,MTD}$) were selected. For each 5 values the averages were calculated

TABLE II
GENERALIZATION ABILITY EVALUATION IN MTC ORDER, WITHOUT CARRYING 500 GRAMS (F: FULL LENGTH, P: SINGLE PERIOD)

SUBJECTS				TESTING			DOMAIN ADAPTATION		
Train	Test	BMI	MTC	Distance(F, mm)	Distance(P, mm)	Percentage (%)	Distance(F, mm)	Distance(P, mm)	Percentage (%)
BCEF	AD	22.99	6.29	1.625 ± 1.456	2.771 ± 1.813	56.23 ± 25.06	0.810 ± 0.826	0.743 ± 0.740	13.22 ± 14.59
CDEF	AB	20.86	6.39	1.998 ± 1.703	3.083 ± 1.861	55.52 ± 24.42	0.832 ± 0.859	0.806 ± 0.776	11.70 ± 10.82
ABEF	CD	23.75	6.61	2.018 ± 1.557	1.880 ± 1.385	43.35 ± 41.38	0.927 ± 0.888	0.745 ± 0.646	14.00 ± 14.30
ADEF	CB	21.62	6.71	2.060 ± 1.497	2.523 ± 1.335	53.40 ± 50.31	0.966 ± 0.983	0.824 ± 0.858	13.28 ± 15.94
BCDE	AF	23.60	6.96	2.597 ± 2.086	4.165 ± 2.346	61.80 ± 26.78	1.082 ± 1.309	0.961 ± 1.070	13.96 ± 16.93
ABDE	CF	24.36	7.28	2.319 ± 1.788	2.504 ± 1.903	46.63 ± 46.21	1.138 ± 1.342	0.946 ± 1.169	14.36 ± 19.82
ABCF	ED	24.92	7.97	2.511 ± 1.934	3.594 ± 2.869	43.58 ± 22.62	1.270 ± 1.235	0.974 ± 0.809	15.19 ± 13.69
ACDF	EB	22.79	8.08	2.724 ± 2.095	4.395 ± 2.746	52.47 ± 20.94	1.266 ± 1.227	1.124 ± 1.012	15.22 ± 16.42
ABCD	EF	25.53	8.64	2.834 ± 1.979	4.719 ± 3.376	48.59 ± 22.91	1.508 ± 1.558	1.182 ± 1.017	15.36 ± 13.81
AVERAGE		23.38	7.21	2.298 ± 1.788	3.293 ± 2.182	51.29 ± 31.18	**1.089 ± 1.136**	**0.923 ± 0.900**	**14.03 ± 15.15**

TABLE III
GENERALIZATION ABILITY EVALUATION IN MTC ORDER, WITH CARRYING 500 GRAMS (F: FULL LENGTH, P: SINGLE PERIOD)

SUBJECTS				TESTING			DOMAIN ADAPTATION		
Train	Test	BMI	MTC	Distance(F, mm)	Distance(P)	Percent	Distance(F)	Distance(P)	Percent
BCEF	AD	22.99	6.29	2.311 ± 1.859	3.264 ± 2.182	62.57 ± 37.55	0.878 ± 0.909	0.766 ± 0.813	13.19 ± 16.11
CDEF	AB	20.86	6.39	2.039 ± 1.547	3.203 ± 2.039	56.37 ± 26.69	0.866 ± 0.884	0.836 ± 0.903	13.33 ± 15.01
ABEF	CD	23.75	6.61	2.975 ± 2.162	3.014 ± 1.962	63.05 ± 56.67	1.125 ± 1.125	0.901 ± 0.913	15.44 ± 16.56
ADEF	CB	21.62	6.71	2.352 ± 1.781	2.617 ± 1.733	52.65 ± 53.76	1.061 ± 1.069	0.874 ± 0.842	15.10 ± 20.71
BCDE	AF	23.6	6.96	2.321 ± 1.775	3.467 ± 2.126	51.50 ± 25.65	1.227 ± 1.436	1.158 ± 1.265	16.87 ± 18.87
ABDE	CF	24.36	7.28	2.579 ± 1.939	2.465 ± 1.636	42.53 ± 40.14	1.244 ± 1.365	1.000 ± 1.100	14.63 ± 19.40
ABCF	ED	24.92	7.97	3.020 ± 2.434	2.979 ± 2.053	38.03 ± 25.68	1.151 ± 1.334	1.010 ± 1.054	13.56 ± 14.60
ACDF	EB	22.79	8.08	3.093 ± 2.428	2.583 ± 1.603	32.86 ± 21.53	1.299 ± 1.333	1.077 ± 0.928	13.78 ± 12.11
ABCD	EF	25.53	8.64	3.422 ± 2.605	5.451 ± 3.640	52.99 ± 23.03	1.519 ± 1.652	1.178 ± 1.098	14.21 ± 13.64
AVERAGE		23.38	7.21	2.679 ± 2.059	3.227 ± 2.108	50.28 ± 34.52	**1.152 ± 1.234**	**0.978 ± 0.991**	**14.46 ± 16.33**

to represent the largest MTD (H_{MTD}) and smallest MTD (L_{MTD}). The two averages were subtracted to calculate the muscle thickness contraction (MTC) in that period. The average absolute difference of the ground truth MTC (MTC_{gt}) and the prediction MTC (MTC_{pred}) over all periods was defined as the muscle contraction loss. The effect of the contraction loss was demonstrated in the ablation study.

$$H_{MTD} = \text{mean}(\text{Top}_{5,MTD}), \tag{2}$$

$$L_{MTD} = \text{mean}(\text{Bottom}_{5,MTD}), \tag{3}$$

$$MTC = H_{MTD} - L_{MTD}, \tag{4}$$

$$Loss_c = abs(MTC_{pred} - MTC_{gt}) \tag{5}$$

To verify model's generalizability to different people in daily applications, the model was trained only on two pairs of females and males and tested on the rest one pair of female and male. As the model has not seen the rest two subjects before, a domain adaptation method was directly applied on the rest subjects: the well-trained model on the four subjects was continually trained by the 20% (1 minute) of the rest subjects using a smaller learning rate, and tested on the other 80% (4 minutes) of the rest subjects' data. This evaluated model's generalization and adaptation to a new person. In both the two training stages, the network was trained for 100 epoches using the RMSProps optimization in Pytorch framework, with the initial learning rate 1e-4 in the main training and 1e-5 in domain adaptation stage.

III. Experimental Results

As different people have different MTC range due to the individual variation in muscle tissue physiology, model's direct testing performance on the rest two subjects were bad, which was illustrated in the "**TESTING**" column of TABLE II. Therefore, a followed domain adaptation process was applied. The model's performance after domain adaptation had been summarized in the right "**DOMAIN ADAPTATION**" column. As the model's prediction was a normalized result and needed to be recovered to the actual distance, the accuracy of the inferred final MTC was highly depended on personalized MTC range. The models' performance was compared with a descend order of the MTC.

To evaluate model's robustness when the movement remain the same but with a greater arm strength, we evaluated model's performance using the same metrics but with subjects' carrying 500 grams weight in hand. The performance was shown in another TABLE III. The comparison between TABLE II and TABLE II directly demonstrated model's robustness.

As different subjects had different MTC range, the prediction accuracy should not only include the actual distance (mm) but also the percentage accuracy (%) of the muscle contraction. They were defined by the Equation (6) and Equation (7) in a single period. The average accuracy for all motion periods were shown under the columns **Distance(P)** and **Percentage(%)**. In addition, the prediction bias of muscle thickness deformation for the full length motions (instead of single period) had been calculated between the whole ground truth sequence and whole prediction sequence, denoted as **Distance(F)**. This checked if the model tracked good on the

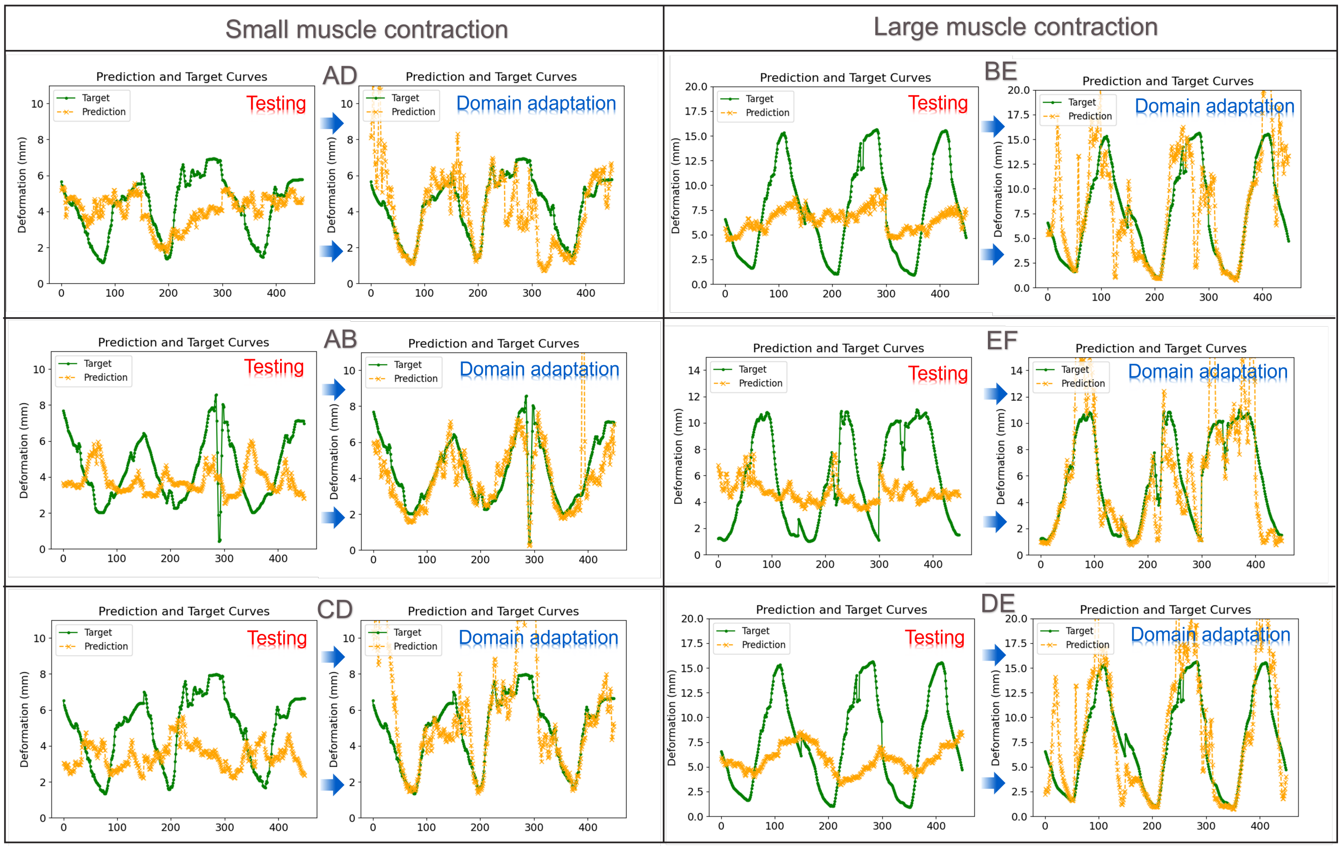


Fig. 7. The MTD ground truth and prediction curves for the test subjects who have the smallest or the largest MTC ranges. In each experiment group, the left figure showed the testing performance, while the right figure showed the performance after domain adaptation.

ground truth MTD.

$$Acc_{MTC}(mm) = abs(MTC_{pred} - MTC_{GT}) \tag{6}$$

$$Acc_{MTC}(\%) = \frac{abs(MTC_{pred} - MTC_{GT})}{MTC_{GT}} \tag{7}$$

Finally, the effects of contraction loss and the cross-attention structure were evaluated on the ablation study.

IV. Discussion

The TABLE II showed all experiment results after training on different combinations of males and females from six subjects. In each experiment, after training on the two pairs of males and females, the model directly tested on another pair of male and female that had never been seen before. Although the initial prediction is bad, the model quickly improve the performance merely using 20% of dataset. In the domain adaptation results, from a grand-view perspective, our approach achieved an average 1.089 ± 1.136 mm accuracy for all six subjects in a point-to-point muscle thickness deformation (MTD) curves tracking accuracy. From a single period contraction perspective, our approach could achieve the average sub-millimeter accuracy 0.923 ± 0.900 mm among the six subjects for the muscle thickness contraction (MTC) prediction, which corresponded to 14.03 ± 15.15 % errors of the ground truth MTC, which proved that this approach could achieve millimeter-level precision in both MTD and MTC. In the following sections, more analysis was done on the TABLE II (without carrying a weight) to prove the generalizability, robustness, and the designs of the approach.

A. Model's Generalizability to different subjects

The prediction performance change from the testing to the domain adaptation can show whether the approach can quickly adapt to a new subject. This could be visualized in several random periods of the predictions from Fig. 7. Two situations (low and high MTC) were shown to explore the extreme cases. In each experiment group performance, from left to right are the performances before and after domain adaptation. Noticed that the model had a large bias of MTC range before adaptation, as different people had different relationship between the strength of sEMG signals and MTC range. After merely 20% subjects' data have been used, the large MTC bias range was significantly decreased.

From all visualized curves, the low MTC groups were adapted better, while the high MTC group had a slightly worse result. This may due to the fact that subject E had a much larger MTC over 9 mm, at least 2mm larger than the others. The training result using a different MTC range (from other subjects) was hard to quickly adapt to the subject E data domain. This problem also existed in the BE experiment group, where the subject B had almost the smallest MC on

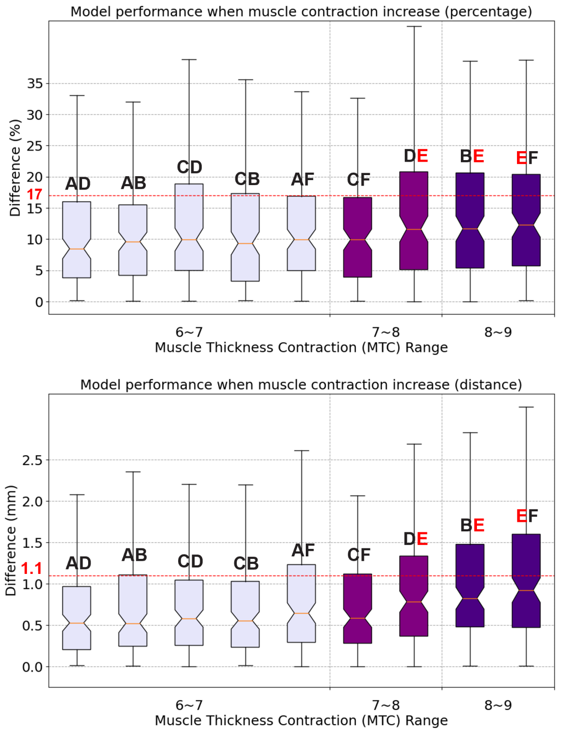


Fig. 8. The histograms for both the accuracy of distance and MTC percentage when subjects did not carry 500 grams weights. The subject E has been highlighted as its average MTC is 2 mm above all other subjects' average MTC.

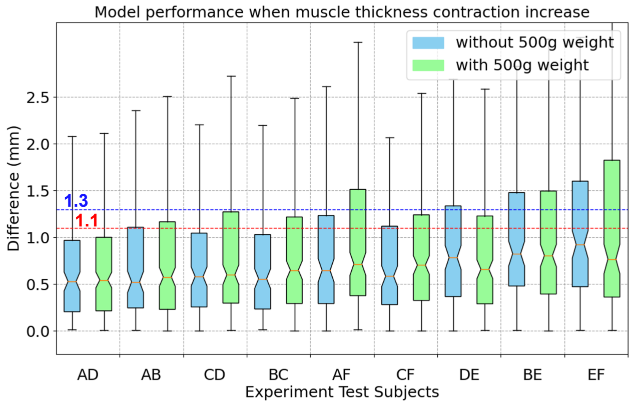


Fig. 9. The combined histograms for both subjects' carrying and not carrying 500 grams by hands. The light purple was without carrying weights, while the dark purple referred to the carrying weights situations.

the extreme contrary of the subject E.

To explore the impact of subject E on the performance, the boxplots for accuracy distribution of all experiment groups were drawn for clear visualization. The results was in Fig. 8. The top figure was based on the MTC percentage accuracy, while the bottom figure used the distance accuracy. The red dash lines corresponded to the 17% and 1.1 mm. In each figure the experiment groups were divided into three larger groups using average MTC of the test subjects. From left to right the test subjects MTC got increased. Except for the groups that contained subject E, mostly other groups' accuracy were higher than the red dash line, which represents the good generalizability for the model when the subjects shared similar MTC, no matter how the personal sEMG signals, BMI or the muscle thickness deformation looked like. For the performance accuracy measured by the distance, although the accuracy decreased when the MTC increased (see DE, BE, and EF cases), the performance by percentage kept almost the same. This was because the training labels have been normalized, the network learn to approach the ground truth curves in a percentage way, helping to better control the training accuracy when the subjects had a very diverse MTC range.

B. Model's Robustness Evaluation

To better visualize the model's robustness when facing with different muscle strength but had the same movement, the distributions of the two situations (carrying or not carrying 500 grams) in TABLE II and TABLE III had been visualized in Fig 9. When subjects did not carry weight, most of performance accuracy were higher than the 1.1 to 1.2 mm. After carrying the 500 grams weight, the model's performance drop to the accuracy higher than 1.3 to 1.4 mm, which was almost the same. Noticed that when comparing the groups AB, CD with DE and BE, carrying a weight did not always bring a performance drop, sometimes increased (see the last three groups). This may due to that carrying weights can lead to a more diverse muscle thickness changes, highly related to personal muscle deformation dynamics (activation and deformation). As the US muscle measurement was only on one position, the muscle dynamics on that specific position brought much more uncertainty and complexity to the problem. Therefore, except for the robustness analysis, the evaluation and accuracy analysis in the paper were summarized from the situations when subjects did not carry the 500 grams weight.

C. Ablation study for the Model's designs

To understand the impacts of the proposed contraction loss and cross-attention structure, the ablation study was done to check model's performance before and after the designs. This comparison was based on the distance (mm) as it showed more precise accuracy changes. For the contraction loss design, the comparison experiments were performed using merely the MSE loss and using both the MSE loss and contraction loss together. The mean and standard deviation before and after applying the loss were shown in TABLE V as a mean±standard deviation format. Except for the last two experiment groups that had the largest MTC (EB and EF), other experiment groups had around 5% to 10% accuracy improvement, which created a general 4.2% increased accuracy performance. This demonstrated that the contraction loss helped the network grasp the muscle thickness contraction range from the sEMG signals. For the two worse cases, the contraction loss cannot perform well due to that subject E brought a different and diverse MTC range, and there was no higher and similar MTC range in the training subjects.

For the cross-attention structure, both the single period MTC (**Distance(P)**) and the prediction accuracy over the full length sequence (**Distance(F)**) had been summarized. The coarse

TABLE IV
ABLATION STUDY FOR CONTRACTION LOSS (CLOSS), AFTER DOMAIN ADAPTATION (P: SINGLE PERIOD)

SUBJECTS		Distance(P)	
Test	MTC	w/o CLoss	Full model
AD	6.29	0.817±0.650	0.743±0.740
AB	6.39	0.854±0.975	0.806±0.776
CD	6.61	0.772±0.706	0.745±0.646
CB	6.71	0.902±0.939	0.824±0.858
AF	6.96	1.027±1.236	0.961±1.070
CF	7.28	0.957±1.019	0.946±1.169
ED	7.97	1.121±1.059	0.974±0.809
EB	8.08	1.074±1.032	1.124±1.012
EF	8.64	1.157±1.330	1.182±1.017
AVERAGE		0.964 ± 0.994	**0.923 ± 0.900**

TABLE V
ABLATION STUDY FOR CROSS ATTENTION STRUCTURE (C-ATT), AFTER DOMAIN ADAPTATION (P: SINGLE PERIOD, F: FULL LENGTH)

SUBJECTS		Distance(P)		Distance(F)	
Test	MTC	w/o C-Att	Full model	w/o C-Att	Full model
AD	6.29	0.768±0.748	0.743±0.740	0.817±0.824	0.810±0.826
AB	6.39	0.762±0.786	0.806±0.776	0.850±0.879	0.832±0.859
CD	6.61	0.714±0.609	0.745±0.646	0.952±0.892	0.927±0.888
CB	6.71	0.845±0.901	0.824±0.858	0.994±0.958	0.966±0.983
AF	6.96	0.931±1.045	0.961±1.070	1.113±1.313	1.082±1.309
CF	7.28	0.984±1.090	0.946±1.169	1.173±1.316	1.138±1.342
ED	7.97	1.047±0.909	0.974±0.809	1.255±1.199	1.270±1.235
EB	8.08	1.042±0.982	1.124±1.012	1.258±1.192	1.266±1.227
EF	8.64	1.137±1.060	1.182±1.017	1.508±1.528	1.508±1.558
AVERAGE		**0.914 ± 0.903**	0.923 ± 0.900	1.102 ± 1.122	**1.089 ± 1.136**

prediction from the period features was used to analyze the performance accuracy that was without the cross-attention. Noticed that for the single period MTC, the performance behave similar with or without the cross-attention. However for the full length sequence accuracy, the performance using cross-attention was better than the model without the structure in most of experiment groups. It meant that the cross-attention structure could not bring much benefits (even sacrifice) for the signal period MTC range prediction, but instead it increases the full length sequence prediction accuracy to improve the general muscle thickness deformation (MTD) tracking performance, which was also very important to decrease the phase shifts and muscle position prediction shifts.

D. Limitation and future perspectives

Based on the above analysis, this dual-attention based approach could be verified to some extent as a universal, generalizable, and robustness for muscle thickness contraction prediction. It was universal to different subjects as long as they share similar MTC range, or the MTC range has been included in the pre-training datasets. It could easily generalize and efficiently adapt to different people with limited dataset (20%) and perform in millimeter-level accuracy. It is robust to some extent when subjects exerting different muscle activities and having different strength movements. We assume that when the training datasets increased to have more diverse personalized MTC datasets, the pre-trained model can well-adapt to more people and have a better performance.

In addition, this approach also has its limitations. Firstly, this method only used six subjects datasets, more diverse BMI / MTC subjects could be included for better model generalization evaluation. Secondly, the ultrasound holder was attached on the elbow using medical cotton cloth, the attach phase may have shift that probably brought some system errors. Thirdly, there was only one A-mode ultrasound probe, could not truly reflect the whole muscle contraction movement. fourthly, as the model required one period length (150 datapoints) as input, if the motion has the indefinite length in time, this dual-attention model could not be used. Lastly, the movements and variants were only limited to the three arm motions and 500 grams weights. More arm motions and weights could be considered to increase the variants and enriched experiment results. However, although there existed the weakness regarding to the experiment setups and approach designs, the highlight of our method was to decipher the correlations between sEMG signals and muscle thickness deformation. The millimeter-level prediction accuracy for the muscle thickness contraction after domain adaption suggest the potential to use sEMG to replace the ultrasound. This demonstrates the possibility to combine muscular mechanical and energy features to create a portable and wearable medical devices (like sEMG device) for the disease and rehabilitation tracking. In the future, based on these weakness, the variety of experiment subjects and movements will be increased for a more complete evaluation of the technique. In addition, more ultrasound probes will be installed on the rigid ultrasound holders for better reflecting the whole muscle movements. The model structure will also be modified so that the input could be available for the diverse periods of motions data. We hope this finding can bring better development in the current medical device and increase the efficiency and convenience for a user-friendly usage in the daily scenarios.

V. CONCLUSION

This paper presents a dual-attention based structure to predict muscle thickness deformation (MTD) and muscle thickness contraction (MTC) only from the muscle activation electrode signals (sEMG). The experiment was performed on the three pairs of males and females with the different BMI indexes and muscle thickness contraction (MTC). Thanks to the self-attention and cross-attention hierarchical structures, the method was demonstrated as universal, generalizable and robust for different subjects. The ability that the model can quickly adapt to new person using merely limited data can help for more widespread application, particularly for those that required convenient and portable devices to replace the ultrasound. This finding can inspire the portable and wearable medical device development, and increase the real-time and daily tracking of the patient's disease.

Acknowledgment

References

[1] M. An, "Specific muscle strength is reduced in facioscapulohumeral dystrophy," *Personalized musculoskeletal modeling of the knee joint*, vol. 26, p. 131, 2018.

[2] K. Engelke, O. Museyko, L. Wang, and J.-D. Laredo, "Quantitative analysis of skeletal muscle by computed tomography imaging—state of the art," *Journal of orthopaedic translation*, vol. 15, pp. 91–103, 2018.

[3] I. Campanini, C. Disselhorst-Klug, W. Z. Rymer, and R. Merletti, "Surface emg in clinical assessment and neurorehabilitation: barriers limiting its use," *Frontiers in neurology*, vol. 11, p. 556522, 2020.

[4] X. Xue, B. Zhang, S. Moon, G.-X. Xu, C.-C. Huang, N. Sharma, and X. Jiang, "Development of a wearable ultrasound transducer for sensing muscle activities in assistive robotics applications," *Biosensors*, vol. 13, no. 1, p. 134, 2023.

[5] S. Ling, Y. Zhou, Y. Chen, Y.-Q. Zhao, L. Wang, and Y.-P. Zheng, "Automatic tracking of aponeuroses and estimation of muscle thickness in ultrasonography: A feasibility study," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 6, pp. 1031–1038, 2013.

[6] X. Chen, Y. Li, R. Hu, X. Zhang, and X. Chen, "Hand gesture recognition based on surface electromyography using convolutional neural network with transfer learning method," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 4, pp. 1292–1304, 2020.

[7] A. Molina-Molina, E. J. Ruiz-Malagón, F. Carrillo-Pérez, L. E. Roche-Seruendo, M. Damas, O. Banos, and F. García-Pinillos, "Validation of mdurance, a wearable surface electromyography system for muscle activity assessment," *Frontiers in Physiology*, vol. 11, p. 606287, 2020.

[8] K. Niu, V. Sluiter, B. Lan, J. Homminga, A. Sprengers, and N. Verdonschot, "A method to track 3d knee kinematics by multi-channel 3d-tracked a-mode ultrasound," *Sensors*, vol. 24, no. 8, p. 2439, 2024.

[9] K. Niu, V. Sluiter, J. Homminga, A. Sprengers, and N. Verdonschot, "A novel ultrasound-based lower extremity motion tracking system," *Intelligent Orthopaedics: Artificial Intelligence and Smart Image-guided Technology for Orthopaedics*, pp. 131–142, 2018.

[10] S. Corradini, F. Alongi, N. Andratschke, C. Belka, L. Boldrini, F. Cellini, J. Debus, M. Guckenberger, J. Hörner-Rieber, F. Lagerwaard, *et al.*, "Mr-guidance in clinical reality: current treatment challenges and future perspectives," *Radiation Oncology*, vol. 14, pp. 1–12, 2019.

[11] A. Waris, I. K. Niazi, M. Jamil, K. Englehart, W. Jensen, and E. N. Kamavuako, "Multiday evaluation of techniques for emg-based classification of hand motions," *IEEE journal of biomedical and health informatics*, vol. 23, no. 4, pp. 1526–1534, 2018.

[12] Y. Xue, X. Ji, D. Zhou, J. Li, and Z. Ju, "Semg-based human in-hand motion recognition using nonlinear time series analysis and random forest," *IEEE Access*, vol. 7, pp. 176 448–176 457, 2019.

[13] C. Chen, Y. Yu, X. Sheng, and X. Zhu, "Non-invasive analysis of motor unit activation during simultaneous and continuous wrist movements," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 5, pp. 2106–2115, 2021.

[14] H. Cui, X. Chi, L. Wang, and X. Chen, "A high-rate hybrid bci system based on high-frequency ssvep and semg," *IEEE Journal of Biomedical and Health Informatics*, 2023.

[15] X. Jia, Y. Liu, Z. Yang, and D. Yang, "Multi-modality self-attention aware deep network for 3d biomedical segmentation," *BMC Medical Informatics and Decision Making*, vol. 20, pp. 1–7, 2020.

[16] H. Pang, L. Zheng, and H. Fang, "Cross-attention enhanced pyramid multi-scale networks for sensor-based human activity recognition," *IEEE Journal of Biomedical and Health Informatics*, 2024.

[17] W. Zhang, Z. Bai, P. Yan, H. Liu, and L. Shao, "Recognition of human lower limb motion and muscle fatigue status using a wearable fes-semg system," *Sensors*, vol. 24, no. 7, p. 2377, 2024.

[18] M. Schouten, P. van de Maat, K. Nizamis, and G. Krijnen, "Evaluating 3d printed semg electrodes with silver ink traces using in-situ impedance measurements," in *2022 IEEE Sensors*. IEEE, 2022, pp. 1–4.

[19] M. Schouten, "3d printed sensors and bio-electronics for robotic applications," 2023.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[21] Y. Li, G. Zeng, Y. Zhang, J. Wang, Q. Jin, L. Sun, Q. Zhang, Q. Lian, G. Qian, N. Xia, *et al.*, "Agmb-transformer: Anatomy-guided multi-branch transformer network for automated evaluation of root canal therapy," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 4, pp. 1684–1695, 2021.

[22] B. Lan, M. Abayazid, N. Verdonschot, S. Stramigioli, and K. Niu, "Deep learning based acoustic measurement approach for robotic applications on orthopedics," *arXiv preprint arXiv:2403.05879*, 2024.

[23] K. Niu, J. Homminga, V. Sluiter, A. Sprengers, and N. Verdonschot, "Measuring relative positions and orientations of the tibia with respect to the femur using one-channel 3d-tracked a-mode ultrasound tracking system: A cadaveric study," *Medical engineering & physics*, vol. 57, pp. 61–68, 2018.

6 Discussion

Based on the orthopedics and muscle activity monitoring background, this master thesis try to develop an intelligent algorithm to classify and recognize the objective world using merely the one-dimensional signal, the A-mode ultrasound. The proposed approach widely adopted one-dimensional convolution neural network (CNN) and its variants for the high-precision perception and tracking of the motion dynamics and movement activities. In addition, the attached Multi-Layer Perceptron structure further decoded category knowledge, and the combination of both sEMG and US signals could unveil the information conversion from energy to momentum in muscle activity. In a more in-depth analysis, the performance (good and bad) of method had been closely checked, and the generalizability and robustness had been demonstrated for better defining the technical application scope from the engineering perspective.

6.1 Answers to the Questions

(1) Can this signal reflect the actual positions and track the movement of the subjects?

From chapter 2 and chapter 3, the CasAtt-UNet and SIRC-UNet were developed to automatically track the bone movement in a high accuracy using merely the one-dimensional ultrasound raw data. From chapter 4, a more simplified model easier to train was proposed to classify the type of channel information. These work demonstrate that through the one-dimensional signal, it is possible to track the actual positions and dynamics movement automatically. Therefore, the A-mode US transducer can be used to track body movement, or to be installed on the robotics arm for the surgery navigation.

(2) What is the range and scope of this perception when it possess a high accuracy, and what components in this method really take effects?

From the TABLE II of chapter 2, the proposed SIRC-UNet can recognize the bone peaks in the local bone area. The effective recognition regions in the joints and the surface of middle bone are around 30 to 70 mm [1]. From the ablation study of chapter 2, both the dice loss and the sampling-based proposal contributed to the high accuracy. The dice loss balance the numbers of positive and negative labels, while the sampling-based proposal connect two UNet to increase the signal perception field, paying more attention to the signal areas that most likely to appear the bone peaks. Therefore, a hierarchical UNet structure was successfully constructed for the accurate detection of the sparse and less evident bone peaks.

(3) When doing the (1) and (2), can it still classify and recognize different types of signals so that it can have a grand view of the subjects for a complete perception?

Chapter 3 proposed a similar but simplified network structure that can be trained end-to-end. Different channels of signals can be automatically classified through decoding the encoded features from the first Coarse UNet. Therefore, the model is able to distinguish different signals types even in the small local areas. The situations of the large bias and small bias have been analyzed to identify the error sources. The capability to classify signal after the high precision position tracking is useful especially when the robotics need to calibrate the position during movement. Also it helps to correct bone registration when the relative shifts between the skin and ultrasound probes occur.

(4) Can this capability of interpreting one-dimensional signal be integrated with other forms of signals for a more complex but useful system development?

Chapter 4 propose such possibility to use the US detection results as the ground truth labels, while the sEMG signals as model's input to predict the muscle thickness deformation solely from muscle activation pattern. The experiment was performed in different subjects under different motions, so that the results with a millimeter accuracy were actually in a generalizable and robust way. Although the two types of signals made the system more complex, it proved on the other hand the feasibility to use convenient sEMG device for replacing the ultrasound, increasing the portable and wearable of the device and making the daily usage of the measurement and patient's care possible.

6.2 Limitation and future perspectives

The proposed methods also have limitations. Firstly, the methods from the first three chapters were only tested on the one cadaver, while the method in the last chapter was only tested on the six subjects and three motions. More subjects and diverse situations should be included for the methods' robustness and generalization evaluations. Secondly, the collected data may not reflect the actual situations. There is a gap between the cadaver setting and the real person's movement. To transform the experiment setting to the in-vivo situations using A-mode ultrasound, the subjects' body movement and the relative shifts between tissue layers will bring much diverse changes of bone peaks in the signals [2, 3]. Thirdly, model's prediction and the recognition results were lack of explanation and transparency. It is unknown for us when and why the models behave good or bad, and the potential ethical challenges may arise [4]. Although two specific cases had been discussed in chapter 4, a more general calibration on the prediction confidence [5] and more in-depth understanding of the performance are lacking. These hinder the further step to be adopted in the clinical decisions. Therefore, in the future, the proposed techniques will be applied on more people and scenarios to validate its performance in a more complex situations. In addition, the transparency and the explainable aspect will be researched and considered to increase the practical values of the approach.

A-mode ultrasound is only one type of the one-dimensional signal, and there exist many other types of one-dimensional signals. To integrate more compact devices on robotics for sensing more traces and signals from the surroundings is another promising directions for the future work [6]. Based on this, a starting point of building the intelligent robotics can be built, which shed the light towards the future of building the robust, transferable and powerful robotics intelligence.

References

[1] Susan Standring, Harold Ellis, J Healy, D Johnson, A Williams, P Collins, and C Wigley. Gray's anatomy: the anatomical basis of clinical practice. *American journal of neuroradiology*, 26(10):2703, 2005.

[2] Kenan Niu, Thomas Anijs, Victor Sluiter, Jasper Homminga, André Sprengers, Marco A Marra, and Nico Verdonschot. In situ comparison of a-mode ultrasound tracking system and skin-

mounted markers for measuring kinematics of the lower extremity. *Journal of biomechanics*, 72:134–143, 2018.

[3] Kenan Niu, Victor Sluiter, Jasper Homminga, André Sprengers, and Nico Verdonschot. A novel ultrasound-based lower extremity motion tracking system. *Intelligent Orthopaedics: Artificial Intelligence and Smart Image-guided Technology for Orthopaedics*, pages 131–142, 2018.

[4] Alvaro Fernandez-Quilez. Deep learning in radiology: ethics of data and on the value of algorithm transparency, interpretability and explainability. *AI and Ethics*, 3(1):257–265, 2023.

[5] Yingxiang Huang, Wentao Li, Fima Macheret, Rodney A Gabriel, and Lucila Ohno-Machado. A tutorial on calibration measurements and calibration models for clinical prediction models. *Journal of the American Medical Informatics Association*, 27(4):621–633, 2020.

[6] Chan Wang, Tianyiyi He, Hong Zhou, Zixuan Zhang, and Chengkuo Lee. Artificial intelligence enhanced sensors-enabling technologies to next-generation healthcare and biomedical platform. *Bioelectronic Medicine*, 9(1):17, 2023.

7 Summary

This thesis has demonstrated that the A-mode ultrasound, when combined with the proposed deep learning techniques, holds significant potential for the high accuracy bone measurement and various types of robotic applications. With the accurate interpretation of the A-mode US raw signals, the methods have been successfully developed that it not only increase the precision of bone tracking in orthopedic surgery, but also enable the real-time anatomical region classification and dynamic muscle monitoring.

Each chapter has contributed to a comprehensive understanding of how the enhanced capabilities be integrated into the practical, non-invasive tools for the clinical use. The CasAtt-UNet and SIRC-UNet models have shown high accuracy in the bone tracking, while the dual-attention framework for the muscle contraction prediction lead to an accurate muscle health monitoring. In the end, these findings not only pave the way for its broader application in healthcare settings, but also inspire the construction of an intelligent robotics that can perceive the surrounding environment.

Acknowledgments

Life seldom sails smoothly; yet, after braving tempests manifold, it unfurls sails of unexpected joy. The boy, once lost in the wilderness of life, could scarcely imagine that after three years, his one part of journey could be illuminated by his efforts on the four submitted papers, which help him complete the master's degree. The talented professors in his life shed the intelligence on him to illuminate his mind and forge his temper, so that he could be more ready to welcome the future academic challenges. Here, the main subject is the author, who is now presenting the thesis and sharing the happiness to you.

To start with, I need to first express my sincere respect and gratefulness to my current master degree supervisor (probably my future Ph.D. supervisor), Dr. ir. Kenan Niu. He chose me among all other master students in 2022. Before that, I was just forced to stop my Ph.D. journey in the U.S. due to the VISA rejection, and had to start a new master degree in another totally different country. Even before that, the COVID-19 brought me a long gap year when I was totally lost in my life direction. I am so lucky that Kenan's invitation could take me to a new life stage, where I receive more formal research and academic thinking training. During one years of getting along, Kenan was good and cultivate me. His endlessly eagerness and the passions to the new things as well as his persistence always inspire me to be more professional in my academic career. Also I need to thank Dr. ir. Momen Abayazid and Prof.dr.ir. N.J.J. Verdonschot for your kind and patient guidance on my publications.

In addition, I need to thank all my surrounding friends in the RAM group who have helped me in this journey. Marjon, Sander and Marcel, thank you for your assistance to magically show any devices or tools that I want in no time! I was amazed by your clear logic and well-organized management for all the different stuffs, and admired for your serious attitudes towards the work. Lennard and Ana, you are always nice to me for helping me and discussing anything about the life and probably my future Ph.D. period. I wish to work with you in the future to enjoy this chill environment. Thank you Toon for sharing your device and suggest me to test and measure using the B-mode ultrasound. I also need to thank Celia and Adithya that were willing to join my experiment! The discussion with you was interesting and full of happiness.

Later, I want to thank for all my closed friends in the daily life, who have encourage me a lot during all my master thesis period, and accompanied with me when I was tired and exhausted. I enjoyed spending the happy time with you! And I feel sorry and sad as I know I cannot see most of you later as you will graduate and go to different countries for your future careers, but just keep in mind that you are always the best and wish all the good blessings always with you! You are Bolin Huang, Yixiang Lu, Ally Lin, Yu-Chian Huang, Akash Ramakrishnan, Jiancheng Hou, Jinze Wang, Sri Saran, Sawan, Pratyush Vaslas, Brook, Tao Huang, Kai Wang, Zhipeng Li, etc. All of you made me a more interesting and relaxed person to enjoy the moment, enjoy life.

In the end, I want to thank for my parents to keep supporting me to walk in here. You paid all my intuition fees and living cost so that I can focus on my study and research with nothing to worry. You always assured me when I was tired, and keep urging, advising and supporting me. I cannot imagine what I will be without you in my life. I wish you always have the happy marriage and enjoy life.

List of Publications

Journal articles

Bangyu Lan, Momen Abayazid, Nico Verdonschot, Stefano Stramigioli, Kenan Niu. SIRC-UNet: Improving Bone Tracking Precision of A-mode Ultrasound Signals by Decoding Hierarchical Resolution Features. *Submitted to IEEE Sensors Journal*

Bangyu Lan, Stefano Stramigioli, Kenan Niu. Deciphering Muscular Dynamics: A Dual-Attention Framework for Predicting Muscle Contraction from Activation Patterns. *Submitted to IEEE Journal of Biomedical and Health Informatics*

Conference contributions

Bangyu Lan, Momen Abayazid, Nico Verdonschot, Stefano Stramigioli, Kenan Niu. Deep Learning based acoustic measurement approach for robotic applications on orthopedics. *Accepted by International Conference on Robotics and Automation (ICRA) 2024*

Bangyu Lan, Stefano Stramigioli, Kenan Niu. Anatomical Region Perception and Real-time Bone Tracking Methods by Dynamically Decoding A-Mode Ultrasound Signals. *Accepted by IEEE RAS EMBS 10th International Conference on Biomedical Robotics and Biomechatronics (BioRob 2024)*