
Degree Project in Electrical Engineering and Computer Science

Second cycle, 30 credits

# Trackerless 3D Freehand Ultrasound Reconstruction Using Deep Learning

## Probe Trajectory Prediction for Tibialis Anterior Muscle Ultrasound Using a CNN and Vision Transformer Hybrid Model

**WENRUI ZHAO**

# Trackerless 3D Freehand Ultrasound Reconstruction Using Deep Learning

## Probe Trajectory Prediction for Tibialis Anterior Muscle Ultrasound Using a CNN and Vision Transformer Hybrid Model

WENRUI ZHAO

Master's Programme, Information and Network Engineering, 120 credits
Date: July 21, 2025

Supervisors: Ruoli Wang, Ruoyu Huang
Examiner: Markus Flierl
School of Electrical Engineering and Computer Science
Swedish title: Spårningsfri 3D-frihandsultraljudsrekonstruktion med hjälp av djupinlärning
Swedish subtitle: Prediktering av probbanor för Tibialis Anterior muskulär ultraljudsavbildning med en hybridmodell av CNN och Vision Transformer

# Abstract

This thesis explores the use of deep learning to estimate the spatial trajectory of an ultrasound probe during freehand scanning of the Tibialis Anterior (TA) muscle, aiming to enable 3D ultrasound reconstruction without the need for external tracking devices. Accurate motion estimation of the probe is essential for reconstructing volumetric images from 2D ultrasound slices, but traditional methods require expensive and cumbersome hardware. The reliance on such systems limits their practicality in clinical settings.

To solve this problem, a novel deep learning architecture combining a Residual Network (ResNet18) and a Vision Transformer (ViT) is proposed. This hybrid model leverages CNNs to capture local spatial features and transformers to model long-range temporal dependencies across image sequences. The model was trained and evaluated on a dataset of ultrasound scans from the TA muscle, with groundtruth poses captured via optical tracking.

Experimental results demonstrate that using 50 constructive frames as input yields the best balance between accuracy and computational efficiency. The proposed ResNet18+ViT model significantly outperforms three baseline methods (CNN, ResNet18, DCL-Net) across multiple metrics, including Distance Error, Final Drift, and Hausdorff Distance. Trajectory visualizations confirm superior alignment and reduced drift.

This work establishes a robust, low-cost solution for freehand 3D ultrasound, enabling precise muscle imaging without external tracking. The proposed model opens up new possibilities for portable and accessible musculoskeletal diagnostics and could be extended to other anatomical structures in future studies.

## Keywords

3D Freehand Ultrasound, Deep Learning, Vision Transformer, Trajectory Prediction

# Sammanfattning

Denna avhandling undersöker användningen av djupinlärning för att uppskatta den spatiala banan för ett ultraljudsprob vid frihandsavbildning av den främre skenbensmuskeln (Tibialis Anterior, TA), i syfte att möjliggöra 3D-ultraljudsrekonstruktion utan behov av externa spårningsenheter. Noggrann rörelseuppskattning av proben är avgörande för att rekonstruera volymetriska bilder från tvådimensionella ultraljudssnitt, men traditionella metoder kräver dyr och otymplig hårdvara. Beroendet av sådana system begränsar deras praktiska användbarhet i kliniska miljöer.

För att lösa detta problem föreslås en ny djupinlärningsarkitektur som kombinerar ett Residual Network (ResNet18) och en Vision Transformer (ViT). Denna hybridmodell använder konvolutionella neurala nätverk (CNN) för att fånga lokala spatiala egenskaper och transformatorer för att modellera långsiktiga temporala beroenden mellan bildsekvenser. Modellen tränades och utvärderades på en dataset bestående av ultraljudsskanningar av TA-muskeln, med markpositionsdata insamlade via optisk spårning.

Experimentella resultat visar att användning av 50 konstruktiva bildrutor som indata ger den bästa balansen mellan noggrannhet och beräkningskostnad. Den föreslagna modellen, ResNet18+ViT, överträffar signifikant tre baslinjemetoder (CNN, ResNet18, DCL-Net) över flera mätvärden, inklusive distansfel, slutlig drift och Hausdorff-avstånd. Visualiseringar av banan bekräftar bättre överensstämmelse och minskad avvikelse.

Detta arbete etablerar en robust och kostnadseffektiv lösning för frihands-3D-ultraljud, vilket möjliggör exakt muskelavbildning utan externa spårningssystem. Den föreslagna modellen öppnar nya möjligheter för portabla och tillgängliga muskuloskeletala diagnoser och kan potentiellt tillämpas på andra anatomiska strukturer i framtida studier.

## Nyckelord

Canvas Lärplattform, Dockerbehållare, Prestandajustering 3D Frihandsultraljud, Djupinlärning, Vision Transformer, Banprediktion

# Acknowledgments

I would like to thank to my supervisor, Ruoli Wang and Ruoyu Huang for their guidance, knowledge, and patience. Their expertise and constructive feedback have greatly shaped the direction and quality of this work. I am especially grateful to Ruoyu Huang for her continuous support and assistance, particularly during the data acquisition process. l would also like to thank my examiner Markus Flierl for providing suggestions of how to revise my thesis.

Lastly, I am deeply grateful to my family and friends for their support and encouragement all the time.

In this paper, ChatGPT is used to correct grammatical errors in sentences.

Stockholm, July 2025
Wenrui Zhao

# Contents

# List of Figures

# List of Tables

# List of Acronyms and Abbreviations

| | |
|---|---|
| 3DfUS | three-dimensional freehand ultrasound |
| 6DoF | Six-Degree-of-Freedom |
| CNN | Convolutional Neural Network |
| DCL-Net | Deep contextual learning network |
| DE | Distance Error |
| FD | Final Drift |
| HD | Hausdorff Distance |
| IMU | Inertial Measurement Unit |
| LSTM | Long Short-Term Memory |
| MD | Maximum Drift |
| MSE | Mean Square Error |
| NLP | Natural Language Processing |
| ResNet | Residual Network |
| SD | Sum of Drift |
| TA | Tibialis Anterior |
| ViT | Vision Transformer |

# Chapter 1

# Introduction

This thesis focuses on the prediction of the spatial trajectory of the ultrasound probe during imaging of the skeletal muscle, i.e., Tibialis Anterior (TA) muscle without relying on external tracking devices. Accurate estimation of the ultrasound probe's motion, specifically its relative position and orientation between frames is a critical part for three-dimensional freehand ultrasound (3DfUS) reconstruction [1], which can recover volumetric information by spatially compounding a sequence of 2D ultrasound slices acquired during unconstrained probe movement. However, traditional solutions typically requires specialized and expensive tracking hardware, which limits their practicality in point-of-care settings. To address this limitation, this study proposes a deep learning-based method that predicts the Six-Degree-of-Freedom (6DoF) relative motion between consecutive ultrasound frames directly from image data.

Specifically, a neural network model is trained to infer relative translations and rotations between neighboring frames in a TA muscle scanning sequence. To efficiently capture the spatial features and temporal dependencies across frames, the network architecture integrates a Convolutional Neural Network (CNN) [2] with a Vision Transformer (ViT) [3] module. The proposed method is further compared with several baseline architectures and demonstrated to be effective in accurately estimating frame-to-frame motion for 3D reconstruction of TA muscle structures.

## 1.1 Background

In medical analysis, 3DfUS offers significant advantages over 2D ultrasound imaging. While 2D ultrasound can provide valuable information on specific

anatomical sections within the footprint of the probe, it suffers from inherent limitations, including restricted spatial orientation, the absence of rich 3D contextual information, and a strong reliance on the operator's expertise and subjective interpretation [4]. 3D reconstruction allows for a more complete display of tissues, resulting in a better assessment of morphology, volume changes, and functional characteristics. It also facilitates more accurate surgical planning, diagnosis and longitudinal monitoring of patients [5].

Conventionally, 3DfUS requires external tracking devices to acquire the spatial position of each frame. However, these systems are expensive and can limit the portability of 3D ultrasound examinations. This project aims to utilize deep learning networks to predict the relative spatial transformations between consecutive ultrasound frames directly from the image data, thus eliminating the need for an external tracker.

The primary objective of this study is to accurately estimate the spatial trajectory of the ultrasound probe during freehand scanning of the TA muscles. The predicted trajectory provides essential spatial information needed to align and integrate 2D frames into a common 3D coordinate system. As such, trajectory estimation serves as a fundamental prerequisite for subsequent 3D reconstruction. Accurate estimation of the probe's motion is particularly valuable for musculoskeletal imaging, where 3D visualization of structures like the TA muscles can support early diagnosis of muscle atrophy, monitor rehabilitation progress, and facilitate biomechanical analysis. By improving the precision and accessibility of trajectory estimation, this work contributes a critical step towards robust trackerless 3DfUS reconstruction.

## 1.2 Problem

### 1.2.1 Original Problem and Definition

The original problem in this thesis is to enable 3D reconstruction of TA muscles from freehand ultrasound sequences using deep learning models. Specifically, the goal is to accurately predict the 6DoF transformation between consecutive ultrasound frames based solely on the image content. This leads to the key research question: To what extent can a deep learning model accurately estimate the spatial trajectory of the ultrasound probe during freehand scanning of the TA muscle?

### 1.2.2 Scientific and Engineering Issues

The scientific challenges in this thesis involves learning spatial and temporal correlations from sequential ultrasound images, which are often noisy, low-contrast, and user-dependent. The problem requires a deep understanding of both medical imaging and pose estimation using data-driven approaches.

From an engineering perspective, the design of an appropriate deep learning model that can infer relative probe motion based only on ultrasound image features is a non-trivial task. Choosing the right model architecture, designing suitable loss functions to reflect both geometric and structural consistency, and ensuring generalization across subjects and scanning styles are major challenges. Additionally, constructing a reliable and synchronized dataset that links ultrasound images to accurate spatial poses without relying on expensive additional hardware is another practical engineering issue.

The solution to these issues must balance computational efficiency, medical relevance, and practical usability in clinical environments.

## 1.3 Purpose

The primary purpose of this thesis is to develop a deep learning based method for predicting the spatial trajectory of the ultrasound probe during freehand scanning of the TA muscle. By estimating the frame to frame motion directly from ultrasound sequences, the proposed approach aims to eliminate the need for external tracking devices and lay the foundation for accurate and low-cost 3D ultrasound reconstruction.

## 1.4 Research Methodology

A quantitative approach was adopted in this project, focusing on the design, implementation and evaluation of a deep learning model for predicting the spatial trajectory of the ultrasound probe during freehand scanning of the TA muscles. The quantitative approach was chosen because the performance of the model needs to be objectively measured by numerical metrics such as Mean Square Error (MSE) and distance error between the predicted trajectory and the groundtruth trajectory.

A hybrid architecture that combines Residual Network (ResNet)[6], which is a common CNN model and ViT was chosen because the two architectures have complementary strengths in processing medical image sequences.

CNNs are very effective in extracting local spatial features from images, especially for medical ultrasound data, which often contain fine-grained texture, noise, and local anatomical patterns. However, since the receptive field of a CNN is localized, its ability to capture long-range dependencies is limited. To address this limitation, the ViT module is incorporated, which excels at modeling global contextual relationships between input features. By using a self-attention mechanism, the ViT can learn dependencies across frames and across spatial regions, making it suitable for understanding temporal motion patterns in ultrasound sequences.

## 1.5 Delimitations

Several important delimitations have been set to define the boundaries of the work:

Region of Interest: The study is limited to the TA muscles. Other muscle groups or body regions, such as the upper limb or abdominal muscles, are not considered in this project.

Model Scope: Only deep learning based methods are explored. Classical optimization or geometry-based reconstruction methods, are not developed in this thesis.

Evaluation Scope: The evaluation focuses on quantitative error metrics such as distance error in pose estimation and point reconstruction accuracy. Broader clinical validation, such as assessing reconstructed muscle quality by medical professionals, is left for future work.

## 1.6 Structure of the Thesis

Chapter 2 presents relevant background information about TA muscle, different 3D freehand ultrasound tracking methods and structure of several deep learning networks. Chapter 3 presents the methodology and method used to solve the problem. Chapter 4 shows the acquired results, and Chapter 5 discusses the achievements and limitations. Finally, Chapter 6 draws conclusions and proposes the future work.

# Chapter 2

# Background

This chapter establishes the foundational background for 3DfUS reconstruction of the TA muscle. It begins with an anatomical overview of the TA muscle, highlighting its location, function, and the clinical necessity of accurate 3D imaging for diagnosis and treatment. The discussion then transitions to an introduction of 3DfUS methods, covering various tracking devices and speckle decorrelation techniques, along with their underlying principles and practical limitations.

Subsequently, the chapter explores the advancements brought by deep learning in this field. The narrative continues by reviewing various deep learning-based 3DfUS approaches that leverage temporal and spatial information for improved accuracy. Finally, introductions of key architectures such as ResNet and ViT are given to explain how these models enhance feature extraction and capture global context in ultrasound images.

## 2.1 Tibialis Anterior Muscle

The lower leg, or crural region, contains several muscle groups that are essential for mobility, balance, and postural control. These muscles work in coordination to enable walking, running, jumping, and standing.

The TA is the largest muscle within the anterior compartment of the lower leg, accounting for over 60% of the total volume of the ankle dorsiflexor muscle group, the group responsible for lifting the foot upward [7]. Its shape and anatomical location are shown in figure 2.1. Anatomically, it originates from the lateral condyle and the upper two-thirds of the lateral surface of the tibia, the anterior surface of the interosseous membrane, and the deep surface of the fascia cruris [7]. Distally, it typically inserts at the medial cuneiform and the

base of the first metatarsal bone [8].

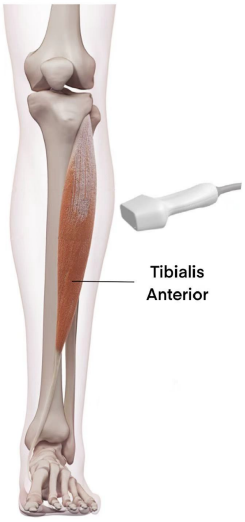



Figure 2.1: The shape and position of the Tibialis Anterior muscle, and its size compared to the ultrasound probe. Adapted from [9].

TA works in conjunction with the extensor hallucis longus and extensor digitorum longus to facilitate dorsiflexion and inversion of the foot [7]. It plays an essential role in maintaining upright posture [10] and contributes significantly to energy absorption during the gait cycle, particularly in the stance phase of walking (when the foot is in contact with ground) [11]. Furthermore, age-related declines in TA muscle strength have been linked to an increased risk of falls in older adults [12].

During the gait cycle, the TA muscle is particularly active during the swing phase (when the foot is lifting off the ground and in the air) and heel strike, preventing foot drop by lifting the toes and allowing correct foot position. Foot drop is one of the most common post-stroke manifestations, with patients often exhibiting muscle weakness and movement control disorders in the affected limb, mainly due to weakness or activation disorders of the dorsiflexor muscles such as the TA [13].On the other hand, dysfunction of the TA muscle pathology is also associated with systemic diseases such as diabetes . In most cases, a careful history and thorough physical examination are sufficient to make the diagnosis. However, clinical evaluation alone may not be sufficient to differentiate between conditions such as tendinopathy, tears, and bursitis [14].

Therefore, imaging of the muscle is valuable as it provides quantitative assessments that can improve diagnostic accuracy and support more informed

treatment planning aimed at functional recovery. Radiography can be used for ossifying myositis or to evaluate arthropathy [15]. Magnetic resonance imaging, due to its excellent tissue contrast, allows simultaneous evaluation of muscle, joint and bone planes. However, due to high costs and the need of specialized equipments, its limited accessibility significantly restricts its widespread use in routine clinical settings.

Whereas ultrasonography is relatively low cost, radiation free, fast imaging and can be used in dynamical condition. It can also be used to study muscle dynamics during contraction and relaxation. The scan can be easily obtained repeatably including the comparison with contralateral muscles. These qualities make it the preferred option for muscle scanning today.

## 2.2 3D Freehand Ultrasound

With the deepening of scientific research and the increasing clinical needs, the shortcomings of traditional 2D ultrasound imaging have been gradually exposed: 2D ultrasound scanning can only show one plane of muscle tissue, and cannot give complete 3D information (see figure 2.2). This requires clinicians to utilize their professional knowledge and experience to infer information in the third dimension. However, the human muscle structure, the variety of morphology, and the differences between individuals are considerable, it can be challenge. If the two-dimensional ultrasound image can be reconstructed and visualized in three-dimension, the doctor can obtain a full picture of the muscle tissue and observe any cross-section plane from any angle, thus improving the diagnostic efficiency and accuracy.

3DfUS is a technique based on 2D ultrasound, in which the spatial trajectory of the probe is tracked and registered to enable 3D reconstruction. The illustration of 3DfUS is shown in figure 2.3. The operator manually moves a conventional 2D ultrasound probe over the target area, while an external tracking device or a computational algorithms estimate the spatial position and orientation of each image frame. Subsequently, a set of 2D images are fused into a 3D volume using interpolation and reconstruction algorithms [5].

Compared with fixed 3D probes or mechanically swept imaging, the freehand technique, though more challenging in terms of registration and reconstruction, offers significant advantages in flexibility, cost-effectiveness, and field of view [4, 16].

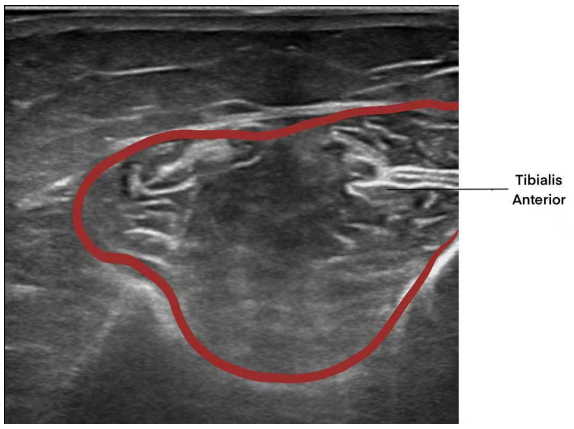



Figure 2.2: A 2D ultrasound slice of the TA muscle.

### 2.2.1 3DfUS Based on External Tracking Device

Currently, there are many ways to track the spatial trajectory of the ultrasound probe, among which the utilization of external tracking devices is a conventional and effective method. Common external position tracking methods are optical tracking, magnetic field sensor,etc. A ultrasound transducer with optical tracking system consists of passive or active targets fixed on the transducer and at least two cameras used to track targets. By observing targets from 2D images, the position and orientation can be calculated with knowledge of relative positions of targets [18].

On the other hand, A magnetic field sensor consists of a time-varying magnetic transmitter placed near the patient and a receiver containing three orthogonal coins attached on the transducer. The receiver measures the strength of magnetic field in three orthogonal directions; then the position and orientation of the transducer can be calculated, which is needed for 3D reconstruction. Magnetic field sensors are relatively small and more flexible without a need for unobstructed sight [19].

However, the shortcomings of these systems are also evident: the introduction of external tracking devices restricts the environment in which they can be used and, together with their high cost, leads to limitations in their clinical application.

### 2.2.2 3DfUS Based on Speckle Decorrelation

In order to eliminate the reliance on external tracking devices, attempts have been made to utilize images to estimate the trajectory of the ultrasonic probe

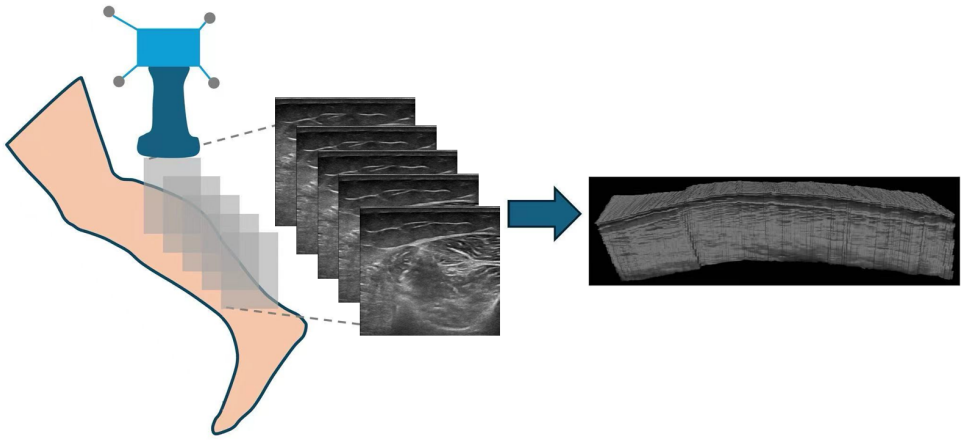


Figure 2.3: llustration of 3D freehand ultrasound based on optical tracking system. Adapted from [17].

for 3D reconstruction, and one of the most representative approaches is to utilize scatter decorrelation.

A pioneering work by Chen et al. [20] in 1997 proposed to estimate the trajectory of an ultrasound scan by utilizing scatter decorrelation. The core of this method is to decompose the inter-frame motion of two consecutive ultrasound frames into two parts: one is the in-plane motion, i.e., the displacement of the part in the plane of the ultrasound image. Since no out-of-plane information is involved this part is less difficult to evaluate. Although the situation in real scans is much more complex than in theory, there are well-established estimation methods, such as the block matching method [21] and the optical flow method [22].

The second component is out-of-plane motion, i.e., elevational distance. The conventional methods use the scattering noise in the image to construct a management model of the correlation and elevation distances between the scattering of images in neighboring frames to estimate the out-of-plane displacements. The specific flow of the scatter decorrelation algorithm is shown in figure 2.4, where two neighboring frame images are firstly segmented into small image blocks, and the image blocks that do not satisfy the assumptions of the scatter model are ignored. After that, for each image block of the first frame image, the normalized cross-correlation with a set of image blocks in the second frame image in its neighborhood is computed, storing the maximum correlation as well as the two-dimensional displacement map in the plane it produces. Next, the elevation distances between frames are obtained

using the decorrelation model, and finally a rigid transformation is performed to obtain the final positional parameters.

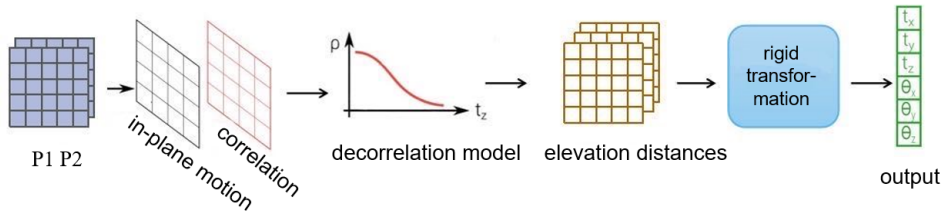



Figure 2.4: Flowchart of scatter decorrelation algorithm.

In the following decade, this method has been improved continuously, and Rivaz et al. [23] introduced the beam tuning method to improve the estimation accuracy. Since the number of patches that satisfy the scatter condition is very small in real ultrasound images, a large number of image blocks are discarded in the scatter decorrelation algorithm. In 2014, Afsham et al. [24] introduced a statistical model based on the inverse Gaussian distribution, which utilizes the image blocks that do not fully satisfy the condition of fully divergent scatter and reduces the number of discarded images. However, these methods, although effective in phantoms datasets, have poor accuracy in real clinical images, where various errors and noise are present.

## 2.3 3D Freehand Ultrasound Based on Deep Learning Networks

With the continuous development of artificial intelligence, deep learning has become an important tool for automatic feature extraction [25]. This has enabled signiffcant progress in overcoming the intrinsic difffculties of 3DfUS [1].

### 2.3.1 Simple CNN

In 2018, Prevost et al. [26] introduced for the first time a CNN into the field of 3D reconstruction of ultrasound images, building a neural network capable of estimating motion between neighboring ultrasound frames. In this work, the authors used a convolutional filter to replace the local inter-correlation

computation in the scatter decorrelation algorithm and used an activation layer to realize the selection of scatter-like regions. Figure 2.5 shows the structure of the CNN developed by Prevost et al. Two consecutive ultrasound frames are the input, and after a series of convolutional and pooling layers, the 6DoF parameters are the output. Also, the optical flow and the Inertial Measurement Unit (IMU) can be optionally input to improve the training accuracy [27, 22]. The training results demonstrate the feasibility of using deep learning for sensorless 3DfUS and point the way for subsequent research work.

To evaluate the reconstruction accuracy and generalization capability of the method under realistic clinical conditions, experiments were conducted on five datasets comprising a total of 800 ultrasound sweeps acquired from the forearm, lower leg, and carotid artery. On the forearm dataset with IMU input, the CNN-based approach achieved a median final drift of 10.4 mm, and maintained consistent performance across different motion types. The predicted sweep lengths showed a strong correlation with groundtruth, with an overall median length error of 6.84 mm. Further generalization experiments on carotid data revealed a slight drop in accuracy when transferring across anatomies; however, performance was significantly improved through fine-tuning or joint training.

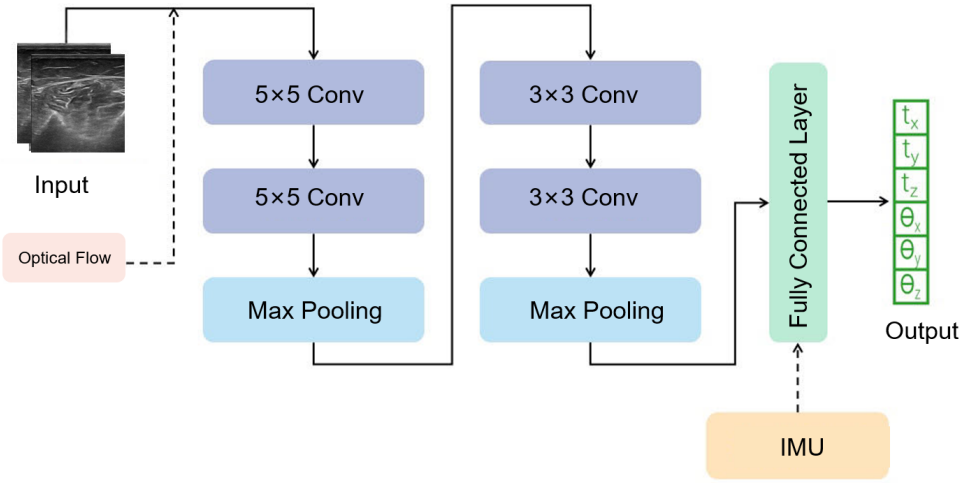



Figure 2.5: Structure of the CNN developed by Prevost et al.[26].

However, Prevost's research results also have a number of drawbacks, such as the optical flow focuses on describing the in-plane motion with limited help for the out-of-plane motion, the limited information that can be obtained from

inputting two frames of images, which makes it difficult to globally estimate the trajectory of the motion and the large dependence on the training data and so on. On the other hand, the neural network structure introduced by Prevost can be considered as advanced and pioneering at that time, but such a neural network structure is considered overly simplistic now. It mainly relied on simply stacking multiple convolutional layers to build a model, which makes it difficult to accurately extract the feature information in the image, and also fails to focus the attention of the deep learning network on the feature-rich region, so there is still a lot of space left to improve.

### 2.3.2 Deep Contextual Learning Network

Guo et al. [28] proposed to use multiple frames instead of two neighboring frames as inputs to fully utilize the contextual information in the ultrasound scanning sequence, and designed Deep contextual learning network (DCL-Net) for inter-frame relative position prediction. Figure 2.6 shows the basic structure of DCL-Net, which is based on the ResNext [29]. The rich contextual information is fully utilized by taking N consecutive ultrasound image frames as inputs, and the embedded self-attention module makes the network model focus more on scatter-rich regions, and finally the relative 6DoF position between every two adjacent frames is output through the pooling layer and the fully connected layer.

Their study was conducted on a large-scale clinical dataset consisting of 640 transrectal ultrasound scanning videos with each video corresponding to a unique patient. It achieved a median distance error of 9.15 mm and a median final drift of 17.40 mm, These results indicate the model's effectiveness in improving 3D reconstruction accuracy from freehand ultrasound sequences, representing a substantial step forward for sensorless ultrasound navigation in real clinical settings.

### 2.3.3 Other Related Works

In the last few years, several different deep learning networks have been used in 3D ultrasound reconstruction.

Miura et al. [30] develop a two-branch network based on ResNet and Flownet [31], incorporating a flipped input consistency loss to stabilize motion estimation across adjacent frames. The model was trained and evaluated on a dataset comprising forearms and phantoms. Quantitative results were evaluated using mean absolute error (MAE). The results showed that the model

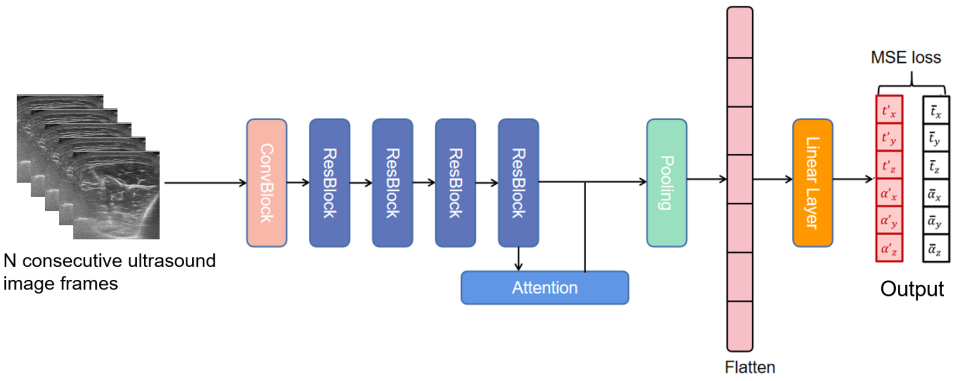



Figure 2.6: The structure of DCL-Net.

achieved high accuracy in most motion parameters, with MAE values of 0.53 mm (x), 1.21 mm (y), and 0.47 mm (z) for translation directions, and 0.64°, 0.15°, and 0.80° for rotation directions.

Xie et al. [32] present Pyramid Warping Network, which reconstructs ultrasound volumes by integrating features from adjacent B-mode images and optical flow. A key component of the network is the Pyramid Warping Layer, which performs multi-scale feature warping based on optical flow to capture motion information at different resolutions. Additionally, the method incorporates channel and spatial attention mechanisms to enhance informative feature regions. A dataset containing frames of forearm B-mode ultrasound images were used in this study. the model's performance was evaluated primarily through mean distance error and final drift error metrics. The results demonstrate that the proposed method achieving an average distance error of 10.52 mm and a final drift error of 4.73 mm.

Luo et al. [33]propose a novel deep motion network (MoNet) for freehand 3D ultrasound reconstruction by integrating image data with IMU signals from a velocity perspective. MoNet employs a temporal and multi-branch architecture based on ResNet and Long Short-Term Memory (LSTM) to extract meaningful features from low signal-to-noise IMU acceleration, enhancing the estimation of elevational displacement. Additionally, a multi-modal online self-supervised strategy is introduced, using IMU data as weak supervision for adaptive optimization during inference. They later proposed OSCNet [34], which further improved trajectory consistency by introducing path-level and multi-IMU consistency learning. Both papers constructed high-quality arm and carotid datasets for evaluation. The results show that under

the same evaluation metrics, OSCNet outperforms MoNet across all metrics, particularly achieving significant improvements in the three core metrics of final drift rate(FDR), average drift rate(ADR), and error of angle(EA): arm dataset improvements of 13.56%/7.32%/30.65%, and 7.62%/4.00%/29.16% improvement on the carotid dataset.

Li et al. [35] explored long-range temporal modeling with an EfficientNet backbone that captures long-term dependencies over hundreds of frames. This approach significantly suppressed cumulative error propagation during reconstruction. Dataset of this study include ultrasound scanning trajectories of both left and right forearms and the result shows that the proposed methods achieved an 82.4% reduction in accumulated tracking error compared to baseline methods.

## 2.4 ResNet

Since there are many different deep learning networks used for ultrasound scan trajectory prediction, in the following sections I will introduce the structure and principles of the models involved in this paper.

The more layers of the neural network, the stronger the ability to extract features, and theoretically better training results can be achieved. However, in practical experiments, with the increase of the number of network layers, the accuracy rate appears to be saturated or even decreases, because the more complex model trained is not necessarily closer to the real value. To solve this problem, He et al. [6] proposed ResNet, so that each additional layer contains the original layer as one of its elements, which ensures that the performance of the deep network is not worse than that of the shallow network.

### 2.4.1 Residual Block

The residual block is the core component of the ResNet architecture. Instead of learning the desired underlying mapping $H(x)$ directly—where $H(x)$ represents the true target output that the network aims to approximate given the input $x$—the residual block reformulates the problem as learning a residual function:

$$F(x) = H(x) - x \tag{2.1}$$

This leads to the final output of the residual block being:

$$y = F(x) + x \tag{2.2}$$

In this formulation, $x$ is the input to the residual block, $F(x)$ is the residual mapping learned by a sequence of layers (typically two or three convolutional layers with batch normalization and non-linear activation), and $y$ is the output that aims to approximate $H(x)$. The shortcut connection that adds $x$ directly to the output facilitates gradient flow, which helps to mitigate the vanishing gradient problem and enables the successful training of very deep neural networks.

### 2.4.2 Network Structure

The basic structure of ResNet18 is shown in Figure 2.7. It starts with a 7×7 convolutional layer, followed by a pooling layer, and then stacks 4 residual structures. Each residual structure uses 2 residual blocks with the same output channel dimension. The first block has the same channel dimension as the input channel dimension. Each subsequent module doubles the channel dimension of the previous module in the first residual block and halves the height and width. The final output is obtained by averaging pooling and fully connected layers. Different variants of ResNet—such as ResNet-34, ResNet-50, and ResNet-101—are constructed by adjusting the number and type of residual blocks.

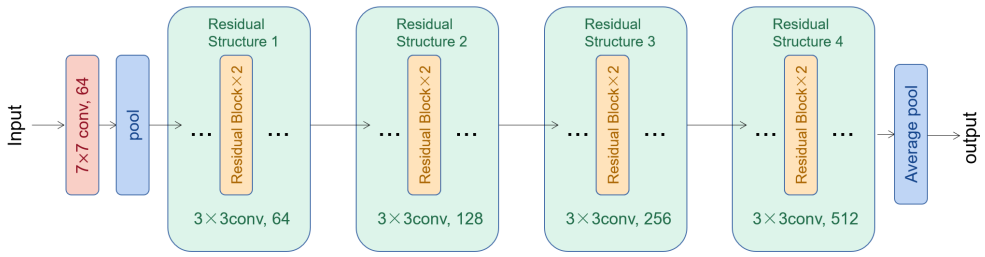



Figure 2.7: Structure of ResNet18

### 2.4.3 ResNext

ResNext [29] is an enhanced CNN architecture based on ResNet. Its key innovation lies in introducing grouped convolutions within the residual blocks, thereby adding a new dimension called cardinality, which refers to the number

of parallel paths within a block. Each path shares the same topology but maintains separate weights. The outputs of all paths are aggregated and added to the identity mapping, forming the final residual output (see figure 2.8). While ResNeXt retains the core idea of residual connections from ResNet, it differs structurally by employing a multi-branch design that increases feature diversity.

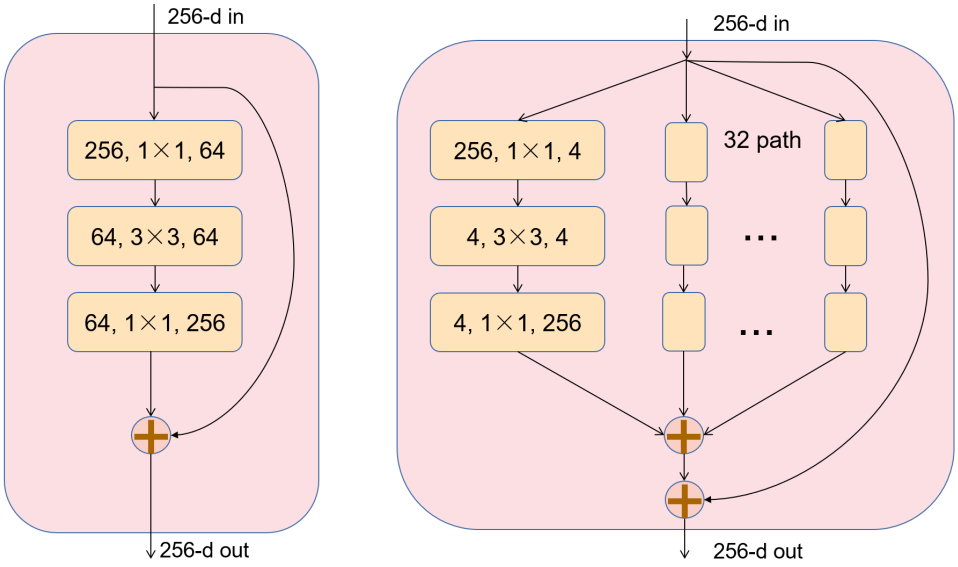



Figure 2.8: **Left**: A residual block of ResNet. **Right**: A residual block of ResNeXt with cardinality = 32.

## 2.5 Vision Transformer

Transformer was initially introduced for Natural Language Processing (NLP) by Vaswani et al. [36]. Its key innovation lies in replacing recurrence with self-attention, allowing each token in a sequence to directly attend to all others, regardless of their position.

The self-attention mechanism allows each token in a sequence to weigh its relationship to every other token, enabling the model to dynamically focus on relevant features. For an input sequence represented by matrices $Q$(Query), $K$(Key), $V$(value), self-attention is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V \tag{2.3}$$

Here, $d$ denotes the dimensionality of the keys and queries. The dot product $QK^\top$ computes the pairwise similarity between tokens, reflecting how much attention each token should pay to the others. The result is then scaled by $\sqrt{d}$ and passed through a softmax function, which normalizes the weights across each row so that they sum to 1. This allows each token to adaptively aggregate contextual information from all other tokens in the sequence, making self-attention particularly effective for modeling global interactions in both language and vision tasks.

To enhance the representational capacity of the model, multi-head attention splits the input into multiple subspaces and applies self-attention independently in each. The results are concatenated and projected back to the original dimension:

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \text{Concat}(H_1, H_2, \ldots, H_h)W_O \\ H_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{2.4}$$

Here, $h$ represent the number of attention heads, $W_i^Q, W_i^K, W_i^V$ are the learnable projection matrices for head $i$ and $W_O$ is the final projection matrix to combine heads.

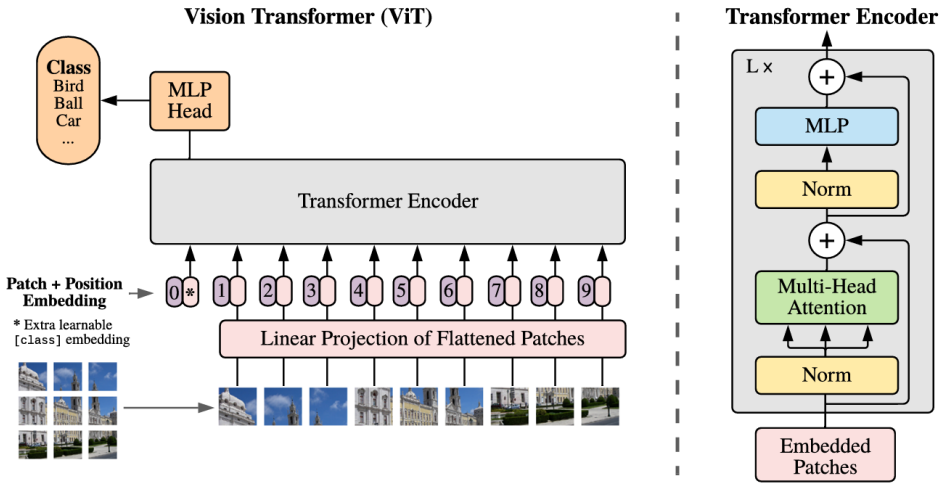



Figure 2.9: The architecture of the ViT from [3].

ViT [3] adapts this architecture to visual data by treating image patches as input tokens, thereby extending the benefits of global attention mechanisms to image recognition tasks. The structure of ViT is shown in Figure 2.9.

Given an image $x$, it is divided into several non-overlapping patches, each of which is flattened and mapped to a D-dimensional embedding via a trainable matrix. A learnable class token and positional embeddings are then added to this sequence, forming the input to the Transformer encoder. Then the output corresponding to the class token is passed through a classification head to produce the final prediction.

# Chapter 3

# Methods

## 3.1 Research Process

The research process of this project has been systematically divided into five essential steps to ensure a structured and rigorous approach. The process includes the following stages:

**Step 1** **literature review.** The research began with an extensive literature review of existing methods in 3D ultrasound reconstruction, pose estimation techniques, and deep learning models.

**Step 2** **data collection and preprocessing.** A dataset was constructed using freehand ultrasound scans of the lower leg, specifically targeting the TA muscle. Each scan consisted of hundreds of frames, along with corresponding 4×4 transformation matrices representing each frame's spatial pose. Images were resized to prepare for neural network training.

**Step 3** **model training.** Train the proposed ResNet+ViT model.

**Step 4** **evaluation and optimization.** The performance of the model is evaluated using metrics such as MSE loss and correlation loss. The estimation performance is evaluated and compared to other baseline models using metrics such as final drift and distance error.

**Step 5** **discussion** Finally, the results are analyzed to assess the effectiveness of the model in predicting accurate 6DoF poses. The estimated 3D trajectories are compared visually and numerically with the ground truth to verify spatial consistency. The strengths and weaknesses of the model are discussed in detail and insights are provided for future improvements and possible extensions.

## 3.2 Data Collection

### 3.2.1 Experimental Setup for Data Acquisition

In this study, conventional 2D US system combined with a motion tracking system was employed to acquire volumetric images of the TA muscle in the lower leg. The system combined a Mindray M9 diagnostic ultrasound machine with a linear transducer (38 mm field of view), a 3D-printed probe holder mounted with four reflective markers(see figure 3.1), and a Vicon[37] optical motion capture system (10 cameras, 100 Hz sampling rate, 1 mm accuracy) for real-time tracking of the probe's position and orientation. Participants were positioned supine and the foot stabilized on a custom footplate. Scans were repeated if motion artifacts or involuntary contractions occurred.

Figure 3.1: The US probe with 4 reflective markers.

### 3.2.2 Data Description and Preprocessing

A total of 12 healthy adult subjects participated in the data acquisition. From these participants, 30 valid scans were collected. Each scan comprises a freehand ultrasound scanning sequence of the TA muscle, consisting of 400 to 700 high-quality 2D ultrasound frames. Each frame is associated with a

corresponding 4×4 transformation matrix that records the spatial pose of the ultrasound probe. Ethical approval was obtained prior data collection.

To facilitate model training and evaluation, the dataset was randomly divided into three subsets: 24 scans were used as the training set, 3 scans as the validation set, and the remaining 3 scans as the test set. Each ultrasound image was resized to 224×224 pixels from 858×790 pixels to be compatible with the input requirements of ResNet used in this study.

All data were anonymized prior to processing to protect participant privacy. Data handling, storage, and usage were carried out in strict accordance with data protection regulations, and no personally identifiable information was retained. This ensured that the research met both scientific rigor and ethical responsibility.

## 3.3 Model

### 3.3.1 Model Design

We propose a hybrid neural network architecture that combines ResNet18 and ViT to leverage the complementary strengths of CNN and transformer-based model for accurate and robust pose estimation. We choose the ResNet architecture as the backbone for feature extraction primarily because its residual connections effectively mitigate the vanishing gradient problem in deep networks and exhibit strong generalization ability. Among the ResNet variants, we specifically adopt ResNet18 due to its smaller parameter size and faster inference speed compared to deeper versions like ResNet50 or ResNet101. This makes ResNet18 more suitable for small datasets, reducing the risk of overfitting, while still preserving CNN's ability to model local spatial structures and effectively capture textures and edge features in ultrasound images.

However, ResNet models inherently have a limited receptive field due to their hierarchical convolutional structure, which makes it challenging to model long-range dependencies or capture global contextual information across sequences of frames. To address this limitation, we incorporate a ViT module, which excels at modeling global context and long-term dependencies using self-attention mechanisms. ViT processes input features as a sequence of patches and models interactions between them regardless of their spatial distance, making it highly effective for capturing temporal correlations and structural continuity in freehand ultrasound scans.

The structure of the model is shown in Figure 3.2. The inputs are multi-frame grayscale image sequences with a shape of $(N, 224, 224)$ where N represents the number of frames. Each frame is independently passed through a ResNet18 network to extract spatial features. Then, each frame is encoded into a 1024-dimensional feature vector, producing a feature sequence of shape $(N, 1024)$. After that, the sequence is passed through a ViT where a learnable class token was added. Therefore the output has a shape of $(N + 1, 1024)$. The sequence is finally passed through a linear layer which maps the output dimension to 6, representing the relative translation and rotation.

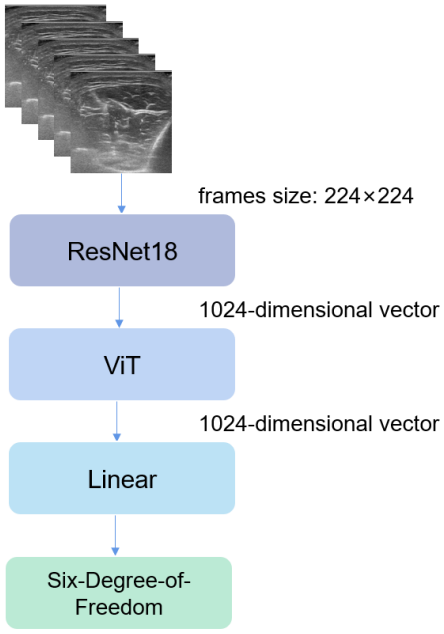



Figure 3.2: The structure of the proposed ResNet18+ViT model.

### 3.3.2 Model Comparison

To validate the effectiveness of the proposed ResNet18+ViT architecture, three other models are chosen as baseline models, including the CNN model proposed by Prevost et al. [26], ResNet18 and DCL-Net [28].

## 3.4 Training Details

The core task of the model is to estimate the relative spatial transformation between two or more consecutive ultrasound frames. Given two adjacent ultrasound frames $I_i$ and $I_{i+1}$, with corresponding transformation matrices $M_i$ and $M_{i+1}$, the relative transformation matrix can be computed as:

$$M_i' = M_{i+1} M_i^{-1}$$

This relative transformation $M_i'$ is then decomposed into 6DoF pose parameters

$$\theta_i = (t_x, t_y, t_z, \alpha_x, \alpha_y, \alpha_z)$$

where $t$ represents translation in millimeters, and $\alpha$ denotes rotation angles.

The model was trained for 200 epochs using the Adam optimizer[38]. The initial learning rate was set to $5 \times 10^{-6}$, and the batch size was set to 8. During each epoch, N constructive ultrasound frames were randomly selected from each sequence to form the training inputs. The predicted output $\theta'$ was then compared to the ground truth $\bar{\theta}$, which is the mean value of relative transformation parameters for N neighboring frames acquired from the dataset. The reason for using the mean as the true value is that the motion between two frames is very small, and using the mean value is effective in smoothing out the noise in the detected motion.

The implementation was done using Pytorch [39] and ran on a Nvidia RTX 4060 GPU.

## 3.5 Data Analysis

### 3.5.1 Loss Function

To accurately learn the 6DoF pose transformation, we design a composite loss function combining the MSE loss and a correlation loss, defined as:

$$L_{\text{total}} = L_{\text{MSE}} + L_{\text{corr}} \tag{3.1}$$

#### 3.5.1.1 MSE Loss

MSE is a widely used loss function particularly when the goal is to measure the difference between predicted and ground truth continuous values. Given a set of $n$ predictions $\hat{y}_i$ and their corresponding ground truth values $y_i$, MSE is

defined as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.2}$$

This formula calculates the average of the squared differences between predicted values and actual values.

#### 3.5.1.2 Correlation Loss

Using only MSE loss results in a smooth estimate of motion, so the trained network tends to memorize the average trajectory of the ultrasound probe in the way clinicians commonly move the probe [40]. To solve this problem, we introduced a case-wise loss function based on the Pearson correlation coefficient [28]. This loss calculates the correlation coefficient between the estimated relative positional pose and the true value for each degree of freedom. It is calculated as:

$$L_{\text{corr}} = 1 - \frac{1}{6}\sum_{d=1}^{6}\frac{\text{Cov}(\theta_d^{\text{GT}}, \theta_d^{\text{OUT}})}{\sigma(\theta_d^{\text{GT}}) \cdot \sigma(\theta_d^{\text{OUT}})} \tag{3.3}$$

Here, $\theta^{\text{GT}}$ and $\theta^{\text{OUT}}$ represent the ground truth and predicted pose vectors, respectively. $\text{Cov}(\cdot,\cdot)$ denotes the covariance and $\sigma(\cdot)$ denotes the standard deviation.

### 3.5.2 Performance Evaluation

#### 3.5.2.1 Performance Evaluation Process

During the training process, the model predicts the 6DoF pose between adjacent frames. Although the 6DoF pose provides a compact representation of the frame-to-frame transformation, the associated errors do not directly reflect the actual spatial deviation in the image plane—especially since even small rotational errors can cause significant shifts in positions far from the rotation center. To address this issue, during testing and evaluation, we use accumulated pose transformations to convert corner points (i.e., the four vertices of each image plane) into a unified world coordinate system. This transformation converts pose errors into spatial point errors, enabling evaluation metrics to more accurately reflect reconstruction accuracy while possessing clear physical significance and intuitive interpretability.

#### 3.5.2.2 Performance Evaluation Metrics

Several metrics are used to measure the reconstruction performance.

**1. Distance Error (DE) [28]:** The average distance between all the corresponding frame corner-points throughout a sequence. This distance error reveals the difference in speed and orientation variations across the entire video.

**2. Final Drift (FD) [26]:** The positional drift at the last frame of the sequence, calculated as the Euclidean distance between the estimated and the true position of the last frame.

$$\text{Final Drift} = \|\hat{p}_N - p_N\|_2 \tag{3.4}$$

where $\hat{p}_N$ is the estimated center coordinate of frame $N$, and $p_N$ is the ground truth center coordinate of frame $N$.

**3. Maximum Drift (MD) [41]:** The largest drift error among all frames in the sequence.

$$\text{MD} = \max_{1 \le i \le N} \|\hat{p}_i - p_i\|_2 \tag{3.5}$$

A smaller MD suggests no severe deviations in any frame.

**4. Sum of Drift (SD):** The total accumulated drift error across all frames in the sequence.

$$\text{SD} = \sum_{i=1}^{N} \|\hat{p}_i - p_i\|_2 \tag{3.6}$$

**5. Hausdorff Distance (HD):** The bidirectional HD highlights the maximum discrepancies between the predicted positions—accumulated through the relative transformation parameters—and the actual positions across all frames in the sequence. This metric is computed using the Euclidean distance between 3D point positions, effectively measuring the greatest spatial deviation between the predicted and ground-truth trajectories.

$$HD(A, B) = \max \left\{ \sup_{a \in A} \inf_{b \in B} \|a - b\|, \quad \sup_{b \in B} \inf_{a \in A} \|b - a\| \right\} \tag{3.7}$$

# Chapter 4

# Results

This chapter presents a comprehensive evaluation of the proposed ResNet18+ViT model. A series of experiments were conducted to determine the optimal number of input frames and assess model performance against established baselines. Quantitative analysis revealed that increasing the input frame count improves reconstruction accuracy up to a certain point, with 50 frames providing the best trade-off between temporal context and computational efficiency. Further comparisons demonstrated that the ResNet18+ViT model consistently outperforms three baseline across all key metrics. Visual trajectory analysis confirmed these results, showing that the proposed model produces the most accurate and stable reconstructions, with minimal drift and high structural fidelity. These findings validate the effectiveness of combining convolutional encoders with Transformer-based attention mechanisms for capturing both local spatial textures and long-range temporal dependencies in complex ultrasound motion sequences.

## 4.1 Optimal Number of Input Frames

Determining the optimal number of input frames is one of the key steps in achieving high-precision trajectory estimation. This step is particularly important for ensuring that the network maintains lightweight while obtaining sufficient temporal information.

The input frame number was set to 2, 5, 15, 25, 50, and 70, respectively, to train the model. It is worth noting that the number of input frames has a significant impact on the training time of the model. Specifically, when using 2 frames as input, the training time is approximately 3 minutes. However, when the number of frames is increased to 50, the training time is approximately 4.5

hours. Further increasing the number of input frames to 70 frames extends the training time to approximately 7 hours.

Three test scans were used for testing and the average of each metric was evaluated. The outputs are shown in table 4.1.

The results show that under our current framework, using 50 input frames offers the best trade-off between prediction accuracy and computational efficiency.

Table 4.1: Evaluation metrics across different input frame numbers.

| Frame num\Metric | DE (mm) | FD (mm) | MD (mm) | SD (mm) | HD (mm) |
|---|---|---|---|---|---|
| 2 | 12.91 | 19.81 | 21.31 | 5713.14 | 21.31 |
| 5 | 10.73 | 15.63 | 17.22 | 4725.35 | 17.18 |
| 15 | 9.51 | 12.98 | 15.17 | 4173.97 | 15.09 |
| 25 | 9.19 | 12.47 | 14.68 | 4026.55 | 14.59 |
| 50 | 8.69 | 11.61 | 13.76 | 3803.40 | 13.73 |
| 70 | 9.05 | 12.28 | 14.35 | 3962.88 | 14.33 |

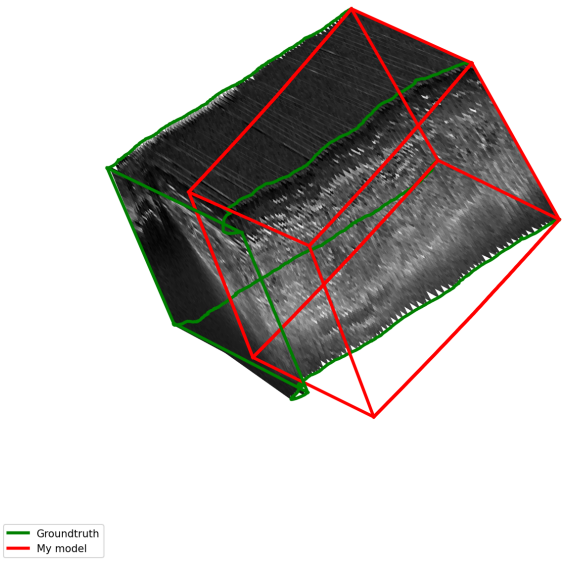
Next, the predicted ultrasound trajectories were visualized and qualitatively analyzed. Figure 4.1a and 4.1b show the ultrasound trajectories predicted by the model for the same scan with input frame numbers of 2 and 50, respectively. The green lines represent the groundtruth of spatial displacement, illustrated on the stacked 2D ultrasound images, and the red lines represent the predicted trajectories. By comparison, it is evident that when the input frame number is 2, the deviation between the predicted trajectory and the groundtruth is greater, not only showing a significant rightward skew but also having a shorter trajectory length.

## 4.2 Model Comparison

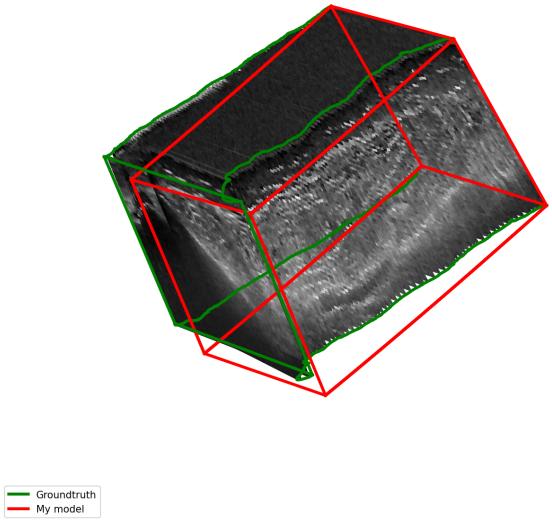
To comprehensively evaluate the proposed ResNet18+ViT model, performance was compared with the baseline methods. The results are shown in table 4.2. The proposed ResNet18+ViT model clearly outperforms the baseline methods

Figure 4.2a, 4.2b, 4.2c and 4.2d show the probe trajectory prediction results of four different models.

In comparison to these three baseline methods, the proposed ResNet18+ViT model achieves significantly more accurate and stable trajectory prediction. The predicted ultrasound trajectory closely matches the groundtruth in terms




(a) Trajectory prediction result when input frame numbers=2.



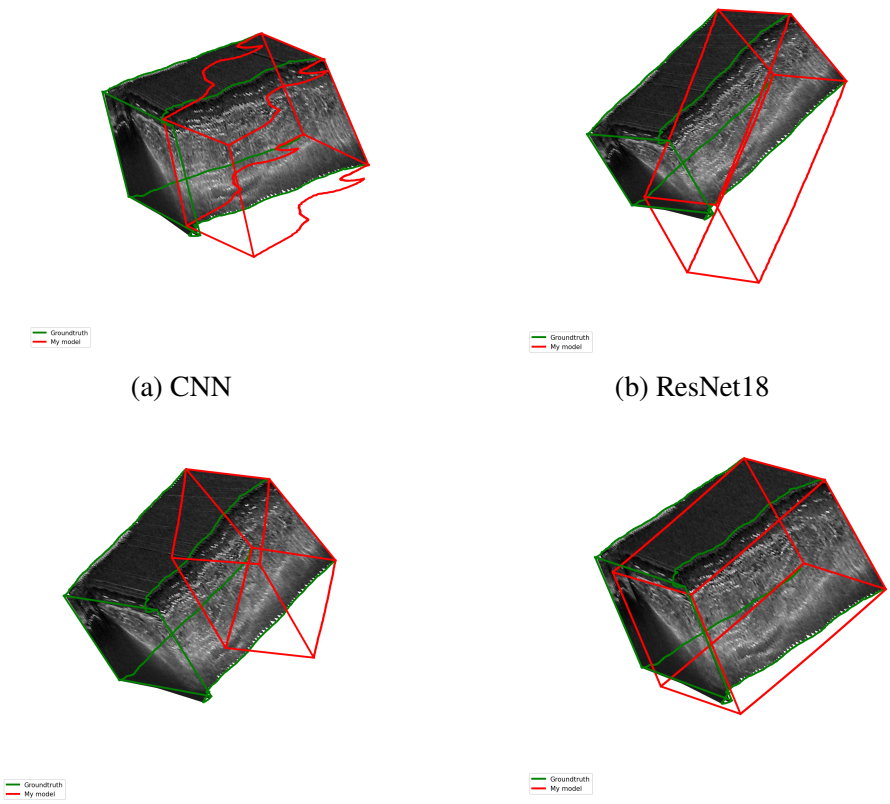

(b) Trajectory prediction result when input frame numbers=50.

Figure 4.1: Trajectory prediction results for different input frame numbers.

Table 4.2: Evaluation metrics across different models.

| Model\Metric | DE (mm) | FD (mm) | MD (mm) | SD (mm) | HD (mm) |
|---|---|---|---|---|---|
| CNN[26] | 36.91 | 55.24 | 58.77 | 16257.12 | 41.68 |
| Resnet18[6] | 20.57 | 37.97 | 37.98 | 9201.03 | 37.58 |
| DCL-Net[28] | 17.15 | 29.59 | 29.87 | 7612.57 | 29.86 |
| ResNet18+ViT | 8.69 | 11.61 | 13.76 | 3803.40 | 13.73 |

of overall shape, directional alignment, and spatial scale, while exhibiting minimal positional drift or jitter across the entire scanning sequence. These findings demonstrate that the integration of convolutional feature encoders with Transformer-based attention mechanisms enables the model to capture both local spatial features and long-range temporal dependencies more effectively. As a result, the proposed architecture shows superior performance in freehand 3D ultrasound reconstruction, particularly in tasks requiring precise and consistent pose estimation across entire sequences.




(a) CNN (b) ResNet18

(c) DCL-Net (d) ResNet18+ViT

Figure 4.2: Trajectory prediction results using different models.

## 4.3 Scan Comparison

Three ultrasound scanning sequences are used for testing the performance of the model. The result of each scan are shown in Table 4.3. It is obvious that compare to Scan 1 and Scan 2, Scan 3 shows a higher FD and MD, meaning its predicted trajectory has a higher drift.

Table 4.3: Evaluation metrics across different scans.

| Scan\Metric | DE (mm) | FD (mm) | MD (mm) | SD (mm) | HD (mm) |
|---|---|---|---|---|---|
| Scan 1 | 8.28 | 8.19 | 12.56 | 3771.04 | 12.52 |
| Scan 2 | 8.91 | 9.76 | 11.86 | 4325.87 | 11.79 |
| Scan 3 | 8.89 | 16.88 | 16.88 | 3313.28 | 16.88 |

# Chapter 5

# Discussion

## 5.1 Key Findings

This research proposes ResNet18+VIT model to address the challenges of sensorless freehand 3D ultrasound reconstruction, specifically targeting the TA muscle. The results demonstrate that the proposed model achieves substantial improvements over existing baseline models across all key evaluation metrics.

### 5.1.1 Optimal Number of Input Frames

The study first focuses on the identification of the optimal number of input frames. Since 3D ultrasound reconstruction tasks rely on the model's ability to perceive spatial changes between consecutive frames, the choice of input frame count directly impacts the model's ability to model motion patterns and the final pose prediction accuracy. Too few input frames may result in insufficient contextual information, making it difficult for the model to capture effective spatial changes; conversely, too many frames may introduce redundant information, increase computational burden, and potentially obscure critical local pose change features. Therefore, systematically comparing experiments with different frame counts can help clarify the model's dependence on temporal information, thereby determining the optimal frame count configuration to achieve stable and efficient training and prediction performance.

After training the model using different input frame numbers, the results show that increasing the number of input frames from 2 to 50 significantly improves model performance. Specifically, DE decreased from 12.92 mm to

8.70 mm (a 32.7% reduction), FD dropped from 19.82 mm to 11.61 mm (a 41.4% reduction). This suggests that longer sequences provide richer temporal context, allowing the model to better learn motion patterns and reduce drift in trajectory estimation.

However, further increasing the frame count to 70 did not yield continued improvements; instead, performance slightly degraded. This decline may be due to redundant frames or increased noise within longer sequences, which can obscure meaningful motion cues. Moreover, extended sequences are more susceptible to non-rigid motion artifacts, including tissue deformation and probe slippage, which will potentially impact pose estimation consistency. Therefore, we can say that 50 frames provided the best balance between computational efficiency and trajectory estimation accuracy.

### 5.1.2 Model Comparison

The second part of the study is to compare the proposed model with three baseline models.

As can be seen from the results, our model achieves the best performance across all five metrics. Specifically, DE is reduced to 8.69 mm, which is a 75% improvement over the basic CNN model[26] and a 57% reduction compared to ResNet18. Compared to DCL-Net[28], the proposed method achieves a 49% decrease.

n terms of the FD, which reflects the accumulated deviation at the end of the sequence, the ResNet18+ViT model achieves 11.61 mm, in stark contrast to 29.59 mm for DCL-Net, 37.97 mm for ResNet18, and 55.24 mm for CNN. Compared to CNN, the final drift is reduced by nearly 79%, and compared to ResNet18 and DCL-Net, the improvements are 69% and 61%, respectively. These significant reductions demonstrate the enhanced ability of our model to maintain long-term trajectory consistency and mitigate cumulative errors over time.

The ResNet18+ViT model also perform best on the other three metrics, achieving an MD of 13.76 mm, SD of 3803.40 mm, and HD of 13.73 mm, outperforming all baseline methods by large margins. In contrast, the CNN model shows the poorest performance with MD = 58.77 mm, SD = 16257.12 mm, and HD = 41.68 mm. This corresponds to an overall reduction of 76.6% in MD, 76.6% in SD, and 67.1% in HD, respectively. Such improvements clearly demonstrate that our model is significantly more resilient to both localized outliers and long-term accumulated errors.

Compared to DCL-Net, which performs second-best on these metrics ,

the ResNet18+ViT model still shows consistent superiority, achieving 53.9% lower MD, 50.1% lower SD, and 54% lower HD. These results indicate that even in challenging frames or over long sequences, our model maintains accurate positional estimates with minimal drift and deformation.

In summary, the quantitative results demonstrate that the ResNet18+ViT model excels in capturing both spatial texture and long-range temporal dependencies, which is crucial for accurate and stable pose estimation in 3D freehand ultrasound reconstruction tasks. The proposed architecture significantly outperforms established baselines across all key performance metrics, validating the effectiveness of combining convolutional encoders with Transformer-based attention mechanisms.

In addition to quantitative performance, qualitative analysis of the predicted trajectories further supports the superiority of the proposed model.

When comparing the reconstruction trajectories produced by different models, the simple CNN exhibits significant deviations from the actual scanning path. The predicted trajectory shows considerable fluctuations and drift, indicating instability in pose estimation across frames. This instability results in a large MD and SD, suggesting that the shallow network architecture lacks the capacity to effectively capture the complex spatiotemporal dynamics required for accurate path reconstruction.

In contrast, the trajectory reconstructed by ResNet18 more closely resembles the overall shape of the ground truth and demonstrates relatively consistent length, implying improved spatial feature extraction and temporal coherence. However, the trajectory exhibits a noticeable downward tilt, indicating persistent directional bias. This directional deviation accumulates over time and results in significant drift, which undermines the overall accuracy of the reconstruction.

For DCL-Net, while the direction of the predicted trajectory aligns reasonably well with the ground truth, the overall path length is noticeably shorter. This compression effect indicates a failure to maintain global scale consistency, likely due to the model's limited ability to preserve long-range spatial correlations over extended frame sequences.

The proposed ResNet18+ViT architecture delivers notably improved accuracy and stability in 3D ultrasound reconstruction. The predicted scanning trajectories demonstrate strong consistency with the groundtruth in terms of geometric shape, movement direction, and spatial scale. Moreover, the model exhibits low levels of drift and fluctuation throughout the entire sequence, highlighting its robustness in maintaining trajectory fidelity over time.

These findings suggest that the proposed architecture has not only improved prediction accuracy but also enhanced stability and generalization ability, making it a promising solution for future development in 3D muscle reconstruction and motion tracking using freehand ultrasound.

## 5.2 Limitation

Although the proposed model demonstrates strong overall performance in terms of 3D reconstruction accuracy and stability, noticeable variations are observed when evaluating different individual scans. As can be seen from table 4.3, even when three test sequences share similar probe trajectories, and acquisition settings, the reconstruction results vary significantly across scans. For example, some scans exhibit considerably higher FD and MD values compared to others.

Several factors may contribute to this inconsistency. First, despite comparable scanning paths, subtle variations introduced by different operators—such as slight deviations in probe angle or variations in scanning speed—can affect the appearance of tissue structures in the ultrasound images, thereby challenging the model's feature extraction and matching capabilities. Second, individual physiological differences, such as muscle thickness or subcutaneous fat layers, may cause variations in image texture, making it difficult for the model to generalize effectively to unseen subjects. Furthermore, the limited size of the dataset might lead to insufficient representation of certain anatomical patterns or motion behaviors.

As a result, while the model performs well on average, improving its robustness to inter-subject variability remains a critical direction for future research. This may involve collecting a more diverse training dataset, adopting more balanced sampling strategies, or introducing personalized adaptation mechanisms to enhance the model's reliability and practical applicability in clinical settings.

Another limitation is that the model struggles to accurately capture small fluctuations and subtle jitters that occur during the scanning process. These minor variations—caused by slight hand manipulation, local probe oscillations, or tissue deformation—are often overlooked by the model, resulting in a smoothed trajectory that may not fully reflect the fine-grained dynamics of the actual probe motion.

This limitation suggests that although the network effectively models global motion patterns and long-term spatial dependencies, its sensitivity to short-term, high-frequency variations remains insufficient. This could

impact the reconstruction quality in scenarios that demand precise anatomical localization or detailed biomechanical analysis.

# Chapter 6

# Conclusions and Future Work

## 6.1 Conclusions

This thesis presented a hybrid deep learning framework for 3DfUS reconstruction of the TA muscle, aiming to eliminate the reliance on external tracking systems. A ResNet18+ViT architecture was proposed, which combines the local feature extraction capability of ResNet18 with the global temporal modeling power of ViT. Extensive experiments were conducted to determine the optimal number of input frames and evaluate the model's performance against three existing baselines.

The results clearly demonstrate that the proposed model outperforms all baselines across multiple reconstruction accuracy metrics. Visual trajectory comparisons further confirmed that our method provides smoother, more accurate, and drift-less reconstructions, thereby validating its effectiveness for robust pose estimation in freehand ultrasound applications.

Overall, this work highlights the importance of incorporating long-term temporal dependencies in motion estimation networks and establishes a strong foundation for markerless, lightweight, and accurate 3D ultrasound systems for muscle assessment.

## 6.2 Future Work

Although the proposed ResNet18+ViT demonstrates strong performance in freehand 3D ultrasound reconstruction of the TA muscle, several avenues remain open for future exploration and improvement.

The current study is limited by the size and diversity of the dataset, which includes only a small number of subjects and a consistent scanning protocol.

Future work will focus on expanding the dataset to include a broader range of participants with different anatomical characteristics, scanning angles, and probe orientations. Introducing additional scanning modalities, such as longitudinal sweeps or angled sweeps, can help the model generalize to more complex clinical scenarios.

On the other hand, the current implementation is focused on the TA muscle, the proposed model can potentially be adapted to other regions. Evaluating the model's generalizability across different anatomical structures will further validate its applicability in broader clinical contexts.

# References

[1] C. A. Adriaans, M. Wijkhuizen, L. M. van Karnenbeek, F. Geldof, and B. Dashtbozorg, "Trackerless 3d freehand ultrasound reconstruction: A review," *Applied sciences*, vol. 14, no. 17, pp. 7991–, 2024. [Pages 1 and 10.]

[2] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. doi: 10.1109/5.726791 [Page 1.]

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021. [Online]. Available: https://arxiv.org/abs/2010.11929 [Pages ix, 1, and 17.]

[4] A. Fenster and D. B. Downey, "Three-dimensional ultrasound imaging," *Annual Review of Biomedical Engineering*, vol. 2, no. Volume 2, 2000, pp. 457–475, 2000. doi: https://doi.org/10.1146/annurev.bioeng.2.1.457. [Online]. Available: https://www.annualreviews.org/content/journals/10.1146/annurev.bioeng.2.1.457 [Pages 2 and 7.]

[5] M. H. Mozaffari and W.-S. Lee, "Freehand 3-d ultrasound imaging: A systematic review," *Ultrasound in Medicine Biology*, vol. 43, no. 10, pp. 2099–2124, 2017. doi: https://doi.org/10.1016/j.ultrasmedbio.2017.06.009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0301562917302776 [Pages 2 and 7.]

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1512.03385 [Pages 3, 14, and 30.]

[7] K. Moore and A. Dalley, *Clinically Oriented Anatomy*. Wolters kluwer india Pvt Ltd, 2018. ISBN 9789387963689. [Online]. Available: https://books.google.se/books?id=9CvvDwAAQBAJ [Pages 5 and 6.]

[8] N. Zielinska, R. S. Tubbs, F. Paulsen, B. Szewczyk, M. Podgórski, A. Borowski, and . Olewnik, "Anatomical variations of the tibialis anterior tendon insertion: An updated and comprehensive review," *Journal of Clinical Medicine*, vol. 10, no. 16, 2021. doi: 10.3390/jcm10163684. [Online]. Available: https://www.mdpi.com/2077-0383/10/16/3684 [Page 6.]

[9] S. Martin-Rodriguez, J. J. Gonzalez-Henriquez, V. Galvan-Alvarez, S. Cruz-Ramírez, J. A. Calbet, and J. Sanchis-Moysi, "Architectural anatomy of the human tibialis anterior presents morphological asymmetries between superficial and deep unipennate regions," *Journal of Anatomy*, vol. 243, no. 4, pp. 664–673, 2023. doi: https://doi.org/10.1111/joa.13864. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/joa.13864 [Pages ix and 6.]

[10] I. Di Giulio, C. N. Maganaris, V. Baltzopoulos, and I. D. Loram, "The proprioceptive and agonist roles of gastrocnemius, soleus and tibialis anterior muscles in maintaining human upright posture," *The Journal of Physiology*, vol. 587, no. 10, pp. 2399–2416, 2009. doi: https://doi.org/10.1113/jphysiol.2009.168690. [Online]. Available: https://physoc.onlinelibrary.wiley.com/doi/abs/10.1113/jphysiol.2009.168690 [Page 6.]

[11] J. N. Maharaj, A. G. Cresswell, and G. A. Lichtwark, "Tibialis anterior tendinous tissue plays a key role in energy absorption during human walking," *Journal of Experimental Biology*, vol. 222, no. 11, p. jeb191247, 06 2019. doi: 10.1242/jeb.191247. [Online]. Available: https://doi.org/10.1242/jeb.191247 [Page 6.]

[12] M. Perry, S. Carville, I. Smith, O. Rutherford, and D. Newham, "Strength, power output and symmetry of leg muscles: Effect of age and history of falling," *European journal of applied physiology*, vol. 100, pp. 553–61, 07 2007. doi: 10.1007/s00421-006-0247-0 [Page 6.]

[13] J. W. Ramsay, M. A. Wessel, T. S. Buchanan, and J. S. Higginson, "Poststroke muscle architectural parameters of the tibialis anterior

and the potential implications for rehabilitation of foot drop," *Stroke Research and Treatment*, vol. 2014, no. 1, p. 948475, 2014. doi: https://doi.org/10.1155/2014/948475. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1155/2014/948475 [Page 6.]

[14] B. S. Varghese A, "Ultrasound of tibialis anterior muscle and tendon: anatomy, technique of examination, normal and pathologic appearance," *J Ultrasound*, pp. 113–123, Dec. 2013. doi: 10.1007/s40477-013-0060-7. [Online]. Available: https://rdcu.be/ekSnx [Page 6.]

[15] P. Tyler and A. Saifuddin, "The imaging of myositis ossificans," *Seminars in Musculoskeletal Radiology*, vol. 14, no. 2, pp. 201–216, 2010. doi: 10.1055/s-0030-1253161 [Page 7.]

[16] L. Mercier, T. Langø, F. Lindseth, and D. L. Collins, "A review of calibration techniques for freehand 3-d ultrasound systems," *Ultrasound in Medicine Biology*, vol. 31, no. 4, pp. 449–471, 2005. doi: https://doi.org/10.1016/j.ultrasmedbio.2004.11.015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0301562905001195 [Page 7.]

[17] Z. Wang, "Quantification of skeletal muscle morphology and mechanical properties using medical imaging," PhD dissertation, KTH Royal Institute of Technology, 2025. [Online]. Available: https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-361815 [Pages ix and 9.]

[18] J. Welch, J. Johnson, M. Bax, R. Badr, and R. Shahidi, "A real-time freehand 3d ultrasound system for image-guided surgery," in *2000 IEEE Ultrasonics Symposium. Proceedings. An International Symposium (Cat. No.00CH37121)*, vol. 2, 2000. doi: 10.1109/ULTSYM.2000.921630 pp. 1601–1604 vol.2. [Page 8.]

[19] Q. Huang and Z. Zeng, "A review on real-time 3d ultrasound imaging technology," *BioMed Research International*, vol. 2017, pp. 1–20, 03 2017. doi: 10.1155/2017/6027029 [Page 8.]

[20] J.-F. Chen, J. B. Fowlkes, P. L. Carson, and J. M. Rubin, "Determination of scan-plane motion using speckle decorrelation: Theoretical considerations and initial test," *International journal of imaging systems and technology*, vol. 8, no. 1, pp. 38–44, 1997. [Page 9.]

[21] M. Ghanbari, "The cross-search algorithm for motion estimation (image coding)," *IEEE transactions on communications*, vol. 38, no. 7, pp. 950–953, 1990. [Page 9.]

[22] G. Farneback, J. Bigun, and T. Gustavsson, "Two-frame motion estimation based on polynomial expansion," in *IMAGE ANALYSIS, PROCEEDINGS*, ser. Lecture Notes in Computer Science. Germany: Springer Berlin / Heidelberg, 2003, vol. 2749, pp. 363–370. ISBN 9783540406013 [Pages 9 and 11.]

[23] H. Rivaz, R. Zellars, G. Hager, G. Fichtinger, and E. Boctor, "9c-1 beam steering approach for speckle characterization and out-of-plane motion estimation in real tissue," in *2007 IEEE Ultrasonics Symposium Proceedings*, 2007. doi: 10.1109/ULTSYM.2007.200 pp. 781–784. [Page 10.]

[24] N. Afsham, M. Najafi, P. Abolmaesumi, and R. Rohling, "A generalized correlation-based model for out-of-plane motion estimation in freehand ultrasound," *IEEE transactions on medical imaging*, vol. 33, no. 1, pp. 186–199, 2014. [Page 10.]

[25] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature (London)*, vol. 521, no. 7553, pp. 436–444, 2015. [Page 10.]

[26] R. Prevost, M. Salehi, S. Jagoda, N. Kumar, J. Sprung, A. Ladikos, R. Bauer, O. Zettinig, and W. Wein, "3d freehand ultrasound without external tracking using deep learning," *Medical Image Analysis*, vol. 48, pp. 187–202, 2018. doi: https://doi.org/10.1016/j.media.2018.06.003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841518303712 [Pages ix, 10, 11, 22, 25, 30, and 34.]

[27] R. Housden, A. H. Gee, R. W. Prager, and G. M. Treece, "Rotational motion in sensorless freehand three-dimensional ultrasound," *Ultrasonics*, vol. 48, no. 5, pp. 412–422, 2008. doi: https://doi.org/10.1016/j.ultras.2008.01.008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0041624X08000139 [Page 11.]

[28] H. Guo, S. Xu, B. Wood, and P. Yan, "Sensorless freehand 3d ultrasound reconstruction via deep contextual learning," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga,

S. K. Zhou, D. Racoceanu, and L. Joskowicz, Eds. Cham: Springer International Publishing, 2020. ISBN 978-3-030-59716-0 pp. 463–472. [Pages 12, 22, 24, 25, 30, and 34.]

[29] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," 2016. [Pages 12 and 15.]

[30] K. Miura, K. Ito, T. Aoki, J. Ohmiya, and S. Kondo, "Localizing 2d ultrasound probe from ultrasound image sequences using deep learning for volume reconstruction," in *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis*, Y. Hu, R. Licandro, J. A. Noble, J. Hutter, S. Aylward, A. Melbourne, E. Abaci Turk, and J. Torrents Barrena, Eds. Cham: Springer International Publishing, 2020. ISBN 978-3-030-60334-2 pp. 97–105. [Page 12.]

[31] P. Fischer, A. Dosovitskiy, E. Ilg, P. Häusser, C. Hazırbaş, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," 2015. [Online]. Available: https://arxiv.org/abs/1504.06852 [Page 12.]

[32] Y. Xie, H. Liao, D. Zhang, L. Zhou, and F. Chen, "Image-based 3d ultrasound reconstruction with optical flow via pyramid warping network," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, 2021. doi: 10.1109/EMBC46164.2021.9630853 pp. 3539–3542. [Page 13.]

[33] M. Luo, X. Yang, H. Wang, L. Du, and D. Ni, "Deep motion network for freehand 3d ultrasound reconstruction," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li, Eds. Cham: Springer Nature Switzerland, 2022. ISBN 978-3-031-16440-8 pp. 290–299. [Page 13.]

[34] M. Luo, X. Yang, Z. Yan, J. Li, Y. Zhang, J. Chen, X. Hu, J. Qian, J. Cheng, and D. Ni, "Multi-imu with online self-consistency for freehand 3d ultrasound reconstruction," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, H. Greenspan, A. Madabhushi, P. Mousavi, S. Salcudean, J. Duncan, T. Syeda-Mahmood, and R. Taylor, Eds. Cham: Springer Nature Switzerland, 2023. ISBN 978-3-031-43907-0 pp. 342–351. [Page 13.]

[35] Q. Li, Z. Shen, Q. Li, D. C. Barratt, T. Dowrick, M. J. Clarkson, T. Vercauteren, and Y. Hu, "Long-term dependency for 3d reconstruction of freehand ultrasound without external tracker," *IEEE Transactions on Biomedical Engineering*, vol. 71, no. 3, pp. 1033–1042, 2024. doi: 10.1109/TBME.2023.3325551 [Page 14.]

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. [Online]. Available: https://arxiv.org/abs/1706.03762 [Page 16.]

[37] P. Merriaux, Y. Dupuis, R. Boutteau, P. Vasseur, and X. Savatier, "A study of vicon system positioning performance," *Sensors*, vol. 17, no. 7, 2017. doi: 10.3390/s17071591. [Online]. Available: https://www.mdpi.com/1424-8220/17/7/1591 [Page 20.]

[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017. [Online]. Available: https://arxiv.org/abs/1412.6980 [Page 23.]

[39] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," 2019. [Online]. Available: https://arxiv.org/abs/1912.01703 [Page 23.]

[40] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016. ISBN 978-3-319-46475-6 pp. 694–711. [Page 24.]

[41] M. Luo, X. Yang, X. Huang, Y. Huang, Y. Zou, X. Hu, N. Ravikumar, A. F. Frangi, and D. Ni, "Self context and shape prior for sensorless freehand 3d ultrasound reconstruction," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, Eds. Cham: Springer International Publishing, 2021. ISBN 978-3-030-87231-1 pp. 201–210. [Page 25.]

TRITA-EECS-EX-2025:587
Stockholm, Sweden 2025

www.kth.se