A Novel Transformer-Based Approach for Simultaneous Recognition of Hand Movements and Force Levels in Amputees Using Flexible Ultrasound Transducers

Xinhao Peng, Yan Liu, Fangning Tan, Weicen Chen, Zhiyuan Liu, Teng Ma, *Member, IEEE*, Xiangxin Li, *Member, IEEE*, and Guanglin Li, *Senior Member, IEEE*

***Abstract*— Accurate hand motion intention recognition is essential for the intuitive control of intelligent prosthetic hands and other human-machine interaction systems. Sonomyography, which can detect the changes in muscle morphology and structure precisely, is a promising signal source for fine hand movement recognition. However, sonomyography measured by traditional rigid ultrasound probes may suffer from poor acoustic coupling because the rigid probe surfaces cannot accommodate the curvilinear shape of the human body, particularly in the case of small and irregular residual limbs in amputees. In this study, we used a self-designed lightweight, flexible, and wearable ultrasound transducer to acquire muscle ultrasound images, and proposed a sonomyography transformer (SMGT) model for simultaneous recognition of hand movements and force levels. The performance of SMGT was systematically compared to two commonly used image processing methods, HOG and Gray Gradient, as well as a deep CNN model, in simultaneously recognizing ten classes of hand/finger movements and three force levels. Additionally, ten subjects including seven non-disabled subjects and three trans-radial amputees who are the end users of prosthetic hands were recruited to evaluate the effectiveness of SMGT. Results showed that our proposed method achieved average classification accuracies of 98.4% ± 0.6% and 96.2% ± 3.0% in non-disabled subjects and amputee subjects, respectively, which are much higher than those of other methods. This study provided a valuable approach for ultrasound-based hand motion recognition that may promote the applications of intelligent prosthetic hands.**

***Index Terms*— Sonomyography, wearable muscle ultrasound, hand motion intention recognition, transformer, trans-radial amputees.**

I. Introduction

IN THE last few decades, hand movement recognition has become one of the most important technologies in the fields of human-machine interaction and rehabilitation engineering. Decoding hand motion intentions from human physiological signals, such as surface electromyography (sEMG) and electroencephalography (EEG), can provide intuitive control for prosthetic hands and other assistive rehabilitation devices [1]. Since the sEMG signal is simple to collect and has the properties of non-invasiveness and high temporal resolution, it has gained widespread use for hand movement recognition. However, sEMG still has some inherent limitations that greatly hindered its clinical application. One of the major drawbacks is its low spatial resolution, as sEMG collected at the surface of the skin are a combination of electrical signals generated by different muscle fibers. This makes it difficult to detect the activities of small and deep muscles from sEMG, and thus leads to a poor performance in recognizing fine hand movements [2]. Therefore, it is necessary to find a more reliable and precise signal source for fine hand movement recognition.

Sonomyography is an effective approach to detect the changes in muscle morphology and structure precisely, and therefore can be considered as a promising complement or alternative to sEMG for fine hand movement recognition. Recently, some researchers have worked on the studies of

Manuscript received 13 June 2023; revised 7 October 2023; accepted 10 November 2023. Date of publication 15 November 2023; date of current version 23 November 2023. This work was supported in part by the Shenzhen Basic Research Program under Grant JCYJ20220818101407016 and Grant JCYJ20210324101601005 and in part by the Jinan 5150 Program for Talents Introduction. *(Corresponding authors: Teng Ma; Xiangxin Li.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee of the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences under Approval No. SIAT-IRB-221115-H0626.

Xinhao Peng and Fangning Tan are with the CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China, and also with the Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences, Shenzhen, Guangdong 518055, China.

Yan Liu, Zhiyuan Liu, and Xiangxin Li are with the CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: lixx@siat.ac.cn).

Weicen Chen and Teng Ma are with the Institute of Biomedical and Health Engineering, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: teng.ma@siat.ac.cn).

Guanglin Li is with the CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China, and also with Shandong Zhongke Advanced Technology Company Ltd., Jinan 25000, China.

Digital Object Identifier 10.1109/TNSRE.2023.3333008

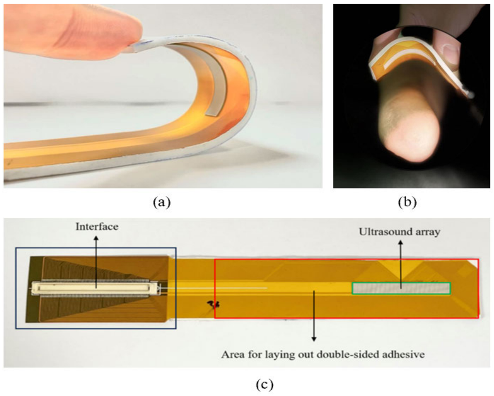
© 2023 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

sonomyography based hand motion recognition since the significant developments were achieved in flexible ultrasonic sensing technology [3]. As of now, three ultrasound modes, namely A-mode, B-mode, and M-mode, have been explored for hand movement recognition [4].

A-mode ultrasound, a kind of one-dimensional sonomyography that can reflect structural information in a single direction, was first introduced as a novel human machine interface method by Guo et al. in 2008 [5]. Subsequently, Yang et al. [6], [7] developed a wearable A-mode ultrasound acquisition device and reach a real-time recognition accuracy of 95.4% ± 8.7% for 12 movements. Zhou et al. [8], [9] combined sEMG features and A-mode ultrasound features, achieving 4% and 20% improvement in recognition accuracy compared to using the separate sEMG features and A-mode ultrasound features, respectively. M-mode ultrasound is a technique that captures a series of A-mode scans over time to visualize the dynamic motion of the one-dimensional structure. Li et al. [10] using M-mode ultrasound and a linear fitting method, achieved a recognition accuracy of 98.70% ± 0.99% for 13 wrist and finger movements.

Different from A-mode and M-mode ultrasounds which only provide one-dimensional structural information, B-mode ultrasound is a two-dimensional imaging technique that offers a more intuitive and interpretable visualization of anatomical structures. The changes in muscle morphology and structure caused by muscle contraction during hand movement can be clearly observed in B-mode ultrasound images. Zheng et al. [11] were the first to utilize B-mode ultrasound for prosthetic control and achieved an average correct rate of 94.05% for five finger movements [12]. Huang et al. [13], [14], [15] compared the effectiveness of sEMG and B-mode ultrasound for gesture recognition and discovered that B-mode ultrasound achieved better performance and long-term effectiveness. Furthermore, Fernandes et al. [16], [17], [18] utilized B-mode ultrasound and proposed a gray gradient feature to predict finger movements and various flexion angles. McIntosh et al. [19] investigated the impact of data acquisition location on classification accuracy and found that the wrist region was most effective for hand motion recognition. Akhlaghi et al. [20] investigated the robustness of B-mode ultrasound-based gesture recognition against different arm positions and found that the effect of arm position on classification was not significant.

Although B-mode ultrasound technology has made some achievements in human motion recognition, it is still not widely used in the applications of human machine interaction. Two major reasons are the bulky ultrasound transducers and the insufficiently developed recognition algorithms. For the transducer, the conventional medical ultrasound transducers are hard, bulky and rely on mechanical fixation, which are prone to displacement during data acquisition. Furthermore, these transducers may not be suitable for small and irregular residual limbs amputees. These probe defects usually cause location changes during use, which seriously affect the recognition accuracy and the application scenarios of ultrasound-based human machine interaction. For the recognition algorithms, most of previous studies only focused on the movement recognition, and there is a lack of studies focusing on the recognition of force levels [21]. However, accurate force recognition is essential in various clinical applications, particularly in controlling prosthetic hands to achieve precise manipulations such as grasping objects of different weights and levels of strength. Importantly, the methods based on B-mode ultrasound proposed in previous studies were primarily tested on non-disabled subjects, and there is few research that specifically involves trans-radial amputees, who are the intended end users of intelligent prosthetic hands. To solve the issues mentioned above, we used a lightweight, flexible, and wearable B-mode ultrasound transducer to acquire data from both non-disabled subjects and amputees, and proposed a Transformer-based approach, called Sonomyography Transformer (SMGT), for simultaneous recognition of hand movements and force levels.


(a)

(b)

(c)

Fig. 1. (a) The flexible ultrasound transducer can be bent into specific shapes and angles to accommodate the curvilinear shape of the human body. (b) Despite the small size and irregular surface of amputees' residual limb, the flexible ultrasound transducer allows for a seamless and conforming fit. (c) The architecture of our flexible ultrasound transducer.

The main contributions of this paper are as follows:(1) We utilized a self-developed lightweight, flexible, and wearable B-mode ultrasound transducer for data acquisition; (2) We developed the SMGT model for simultaneous recognition of hand movements and force levels, achieving an average accuracy of 98.4% ± 0.6% in seven non-disabled subjects; (3) The proposed method was validated in three trans-radial amputee subjects, with an average accuracy of 96.2% ± 3.0% achieved; and (4) A systematic comparison was conducted between our SMGT and three commonly used methods, demonstrating the superior of our method.

II. Methodology

A. Wearable Ultrasound Transducer

To accommodate the small size and irregular surface of amputees' residual limb, as well as to avoid the impact of probe displacement on recognition performance, we utilized a lightweight, flexible, and wearable ultrasound transducer developed by our institute to acquire data. As shown in Fig. 1(a), a 128-element flexible linear array ultrasound transducer was bent into curved shape. The transducer has an average center frequency of 6.35 MHz, an average −6-dB bandwidth of 69.2%, and can achieve a minimum concave bend diameter of 20 mm and a minimum convex bend diameter of 25 mm. Detailed beamforming process and elements positioning method can be found in [22].

The size of the ultrasound array is only about 40 mm × 5 mm, and it is placed on a flexible backing layer to support

TABLE I
SUBJECTS INFORMATION

Sub	Age	Circumference of Arm (cm)	Amputated Limb	Time of Amputation
Abs 1	24	19.5	/	/
Abs 2	27	20.5	/	/
Abs 3	24	19.0	/	/
Abs 4	24	18.0	/	/
Abs 5	25	18.5	/	/
Abs 6	24	19.5	/	/
Abs 7	35	18.0	/	/
Amp 1	38	17.0	Left	2006
Amp 2	27	16.0	Left	2018
Amp 3	30	14.0	Right	2006

the circuitry and interfaces, thus making the total size of the flexible transducer to be 177 mm × 26 mm. The weight of the whole transducer is 12.2 grams, so wearing it will not cause any discomfort to the user, and the maximum average power consumption for conventional imaging pulses is 10W. Through a wired connection to the ultrasound system, the flexible transducer allows for long term functional imaging without additional mechanical clamping, and it can be bent to any shape to fit the body perfectly, as shown in Fig. 1(b). It should be noted that it is still necessary to apply ultrasound gel to achieve high-quality imaging, and we used double-sided adhesive and medical elastic bandage to fix the probe on the arm and prevent probe dislocations during motions. The lightweight, flexible and wearable properties of the transducer prevent interference with movement and effectively eliminate transducer displacement, which significantly affects recognition accuracy.

B. Experiment Setup and Protocol

1) Subjects and IRB Approval: Ten subjects including seven non-disabled subjects (Abs 1-7) and three trans-radial amputees (Amp 1-3) with ages between 23-38 years old were recruited for this study. Their specific individual information is shown in Table I. Prior to the experiment, subjects were informed about the experiment protocol and signed an informed consent form. The experimental protocol was approved by the ethics committee of the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences and in accordance with the declaration of Helsinki, and the approval number is SIAT-IRB-221115-H0626.

2) Experiment Setup: In the experiment, we used a research ultrasound system (Vantage 256, Verasonics Inc., Kirkland, Washington, USA) to acquire ultrasound images, which was used in conjunction with the flexible ultrasound transducer mentioned in part A. The ultrasound images were displayed and saved on a computer. The imaging depth was set to 30 mm, and the received frame rate is resulted in 10 Hz within the constraints of ultrasound frequency, imaging depth, image quality, and transmission speed. The probe is placed one-half of the way up the forearm (for the amputees, the probe location is one half of the stump to elbow) and covered the flexor digitorum superficialis and the flexor digitorum profundus of the forearm, as shown in Fig. 2.

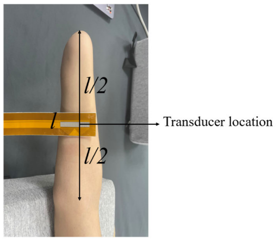


Fig. 2. The transducer location on the forearm.

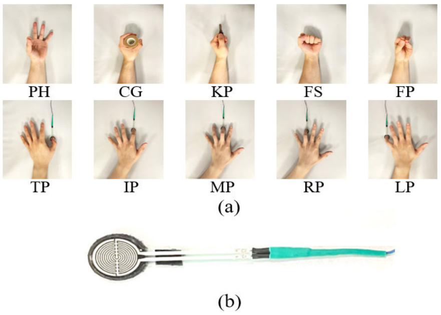
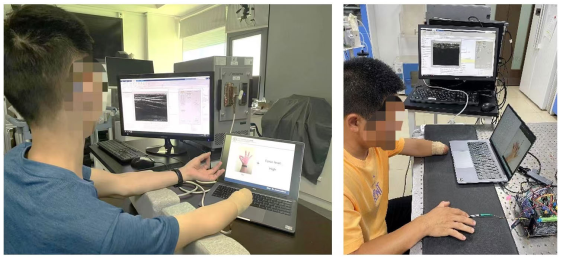


Fig. 3. (a). 10 movements including five functional grasp movements: pinch (PH), cylindrical grasp (CG), key pinch (KP), fist (FS), five finger pinch (FP) and five finger press movements: thumb press (TP), index finger press (IP), middle finger press (MP), ring finger presa (RP), and little finger press (LP). (b). The force sensor FSR402 used in our self-developed finger force platform.

3) Experiment Protocol: In the experiment, each subject was asked to performance 10 classes of movements at three force levels respectively, according to the guidance on the computer screen. The 10 classes of movements are shown in Fig. 3(a), including five functional grasp movements that are commonly used in daily life: pinch (PH), cylindrical grasp (CG), key pinch (KP), fist (FS), five finger pinch (FP), and five fine finger movements that can helps with fine manipulation: thumb press (TP), index finger press (IP), middle finger press (MP), ring finger press (RP), and little finger press (LP). In the experiment every movement at each force level will be repeated eight times. In each repetition, the subject was instructed to perform and hold the movements at the corresponding force level for seven seconds. To avoid potential muscle fatigue, a 60 second rest was taken between two consecutive movements.

According to subjects' feedback, all subjects can perceive and execute the functional grasp movements at three different force levels (low, medium, and high). The medium force level is the moderate effort that is naturally produced by the subjects; the high force level is higher than the moderate level and almost equal to the maximum voluntary contraction (MVC); the low force level is lower than the moderate level exerted by the subjects. To ensure the stability and reliability of the data collected from the residual limbs of amputees, a mirrored bilateral training strategy [23] was employed during the data acquisition process. In this training strategy, amputees were instructed to exert the same movement and force simultaneously using both their intact hand and their phantom hand, as shown in Fig. 4(a).

But some subjects reported that they have difficulty in perceiving and executing accurate and stable force levels for the fine finger movements. Therefore, we used a self-developed force feedback system to help the subjects to execute different force levels. As shown in Fig. 4(b), amputee subjects perform



(a) Fuctional movements (b) Fine finger movements

Fig. 4. Experimental scene.

each fine finger movement with both intact hand and amputated hand. The pressure data recorded by the force sensor (FSR402, Interlink Electronics Inc, Irvine, California, USA) of each fingertip of intact hand was transmitted through a microcontroller board (Arduino UNO, Arduino LLC, Boston, Massachusetts, USA) and displayed on the computer screen in front of the subject. By using the mirrored bilateral training strategy coupled with real-time force feedback, subjects said that they were able to produce the stable force levels. The three force levels were set to 15%∼25%, 45%∼55%, and 75%∼85% of the maximum voluntary contraction (MVC), in which the MVC was pre-measured prior to data collection.

C. Data Pre-Processing

Before using the data for simultaneous recognition, the raw ultrasound image frames recorded from 2s to 6s of each movement were extracted by a video split-frame operation, since the muscle contractions and force levels are more stable in the duration. After that, we get the raw images with size of 688 × 544 pixels. To reduce the computation time, the raw images are resized into 384 × 384 pixels.

D. Sonomyography Transformer (SMGT)

Transformer is a deep sequence model based on self-attention mechanism, and achieved significant performance in natural language and image processing recently [24], [25]. In this study, we designed a SMGT model to simultaneously recognize the hand movements and force levels. Additionally, considering that the transformer-based models typically require a large amount of data for model training, which is time-consuming and boring, two data augmentation methods were integrated with our proposed SMGT model to improve the performance of our model. The architecture of the model is shown in Fig. 5(a) and depicted as follows.

1) Data Augmentation: Data augmentation is a useful technique to improve the generalization ability and robustness of the model by expanding the size and diversity of the train dataset. In this study, two data augmentation methods of Cutout [26] and Mixup [27] were used to generate new samples for the training dataset, as depicted in Fig. 5(b). In Cutout method, random portions of each ultrasound image were removed and replaced with 0 pixel values. As shown in formula (1), in which *(i, j)* represent locations of the cutout pixels. Specifically, we set the probability of performing the operation to 0.5, the size of the clipped portion to be from 2% to 40% of the total area of the image, the aspect ratio of the clipped portion to be from 0.4 to 2.5, and the location of the clipping to be randomized. While Mixup is a proportional mix of two random samples and their label, as shown in formula (2) and (3). In which the $x1$, $x2$, $y1$, $y2$ represent the image and label respectively, and the decisive parameter λ, which determines the mixing rate, takes a value between 0 and 1 and follows a beta distribution *beta (0.5,0.5)*.

$$X_{cutout}(i,j) = \begin{cases} 0 & a \le i < b,\ c \le j < d \\ X(i,j) & otherwise \end{cases} \tag{1}$$

$$\tilde{x} = \lambda x1 + (1-\lambda)x2 \tag{2}$$

$$\tilde{y} = \lambda y1 + (1-\lambda)y2 \tag{3}$$

2) Class and Position Embedding: Before applying class and position embedding, the input 3-channel image with a size of 384 × 384 × 3, is divided into 576 patches, each with a size of 16 × 16 × 3, where 576 is the resulting number of $(384/16)^2$. Then each patch is flattened into a 768-dimension vector, where 768 is the resulting number of 16 × 16 × 3. The vectors from all 576 patches are then spliced together to form a 576 × 768 matrix. This matrix is then fed into a fully connected layer, denoted as E, to obtain the patch tokens as shown in formula (4). Token is a concept from machine translation, but the token in this model is a 768-dimension vector that obtained from each patch. Then a learnable token $x$class was embedded to patch tokens. At the end of the transformer encoder this $x$class will be extracted as the output y for classification, as shown in formula (7). After class embedding, position embedding is added to the patches to retain positional information. The class and position embedding can be described as below:

$$Z_0 = [x_{class}; x_p^1 E; x_p^2 E; \cdots; x_p^N E] + E_{pos}$$
$$\text{where } E \in \mathbb{R}^{768\times768},\ E_{pos} \in \mathbb{R}^{577\times768} \tag{4}$$

3) Transformer Encoder: It consists of alternating layers of multiheaded self-attention (MSA) and multilayer perception (MLP) blocks. MSA enables the model to jointly focus on information from different representation subspaces at different locations [25], while MLP applies nonlinear transformations and mappings to the input to enhance the model's expressive power and nonlinearity. Additionally, to enhance the training efficiency and accuracy of the model, a normalization layer is applied before each block and residual connections are used after each block [28]. The depth of the transformer encoder, denoted as L, was set to be 12 in our model. The overall structure is illustrated in Fig. 5(d) and can be described as formula (5) and (6):

$$Z'_l = MSA(LN(Z_{l-1})) + Z_{l-1}, \quad l = 1\cdots L \tag{5}$$

$$Z_l = MLP(LN(Z'_l)) + Z'_l, \quad l = 1\cdots L \tag{6}$$

$$y = LN(Z_L^0) \tag{7}$$

4) Multiheaded Self-Attention (MSA): MSA consists of several "Scaled Dot-Product Attention" layers, as shown in Fig. 5(c). For each element in an input Z, it was multiplied with three different weight matrices to obtain the query vector (Q), the key vector (K), and the value vector (V). The Scaled Dot-Product Attention compute the dot product of the query with all keys, divide each by $\sqrt{dk}$, and apply a SoftMax function to obtain the weights on the values. So, the output of Multiheaded Self-attention can be described

TABLE II
DIFFERENT FORCE LEVELS COLLECTED OF EACH MOVEMENTS

	Functional Grasp Movements					Fine Finger Movements				
	PH	CG	KP	FS	FP	TP	IP	MP	RP	LP
Force Levels	Low	Low	Low	Low	Low	20 ± 5%	20 ± 5%	20 ± 5%	20 ± 5%	20 ± 5%
	Medium	Medium	Medium	Medium	Medium	50 ± 5%	50 ± 5%	50 ± 5%	50 ± 5%	50 ± 5%
	High	High	High	High	High	80 ± 5%	80 ± 5%	80 ± 5%	80 ± 5%	80 ± 5%

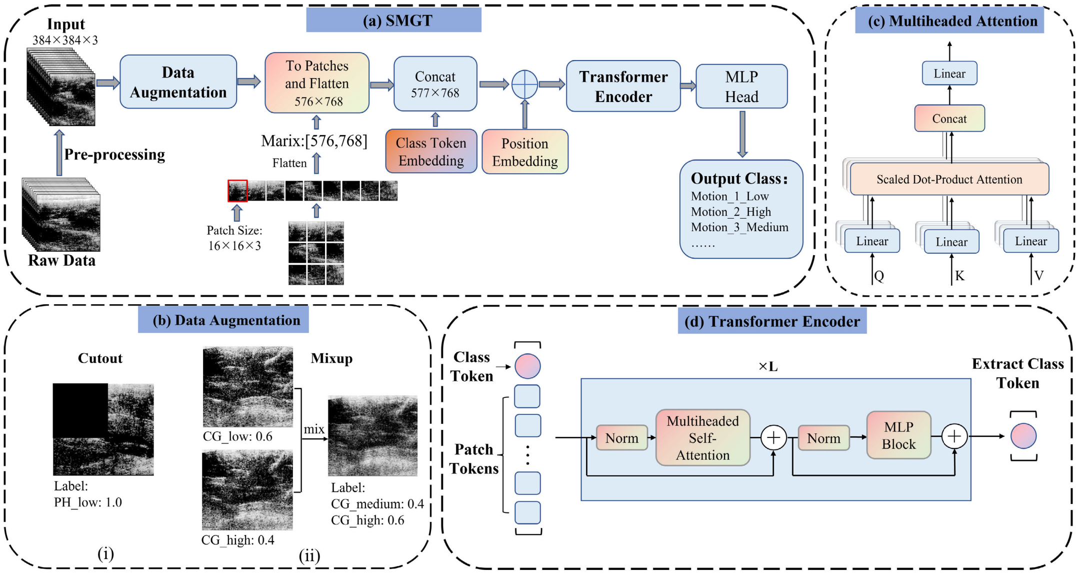


Fig. 5. The architecture of our Sonomyography Transformer (SMGT) model: (a) The overall structure of SMGT, (b) Transformer encoder, (c) Multiheaded attention, (d) two data augmentation methods used in this study: (i) Cutout, (ii) Mixup.

as formula (8), (9), and (10):

$$[Q, K, V] = Z[Wq, Wk, Wv] \tag{8}$$

$$H = Attention(Z) = SoftMax(\frac{QK^T}{\sqrt{dk}})V \tag{9}$$

$$MSA(Z) = Concat[H1; H2; \cdots ; Hh]Wo \tag{10}$$

In this study, we used 12 parallel attention layers ($h = 12$), which is commonly referred to as 12-head attention.

5) Training Setup and Data Split: Our model was implemented on Pytorch framework and the training process was accelerated by an NVIDIA TITAN V GPU. For the training process, we set the training epoch to 100, the learning rate to 1e-4, and the weight decay to 1e-4, respectively. The dataset for each subject was divided into two parts: the first four repetitions and the last four repetitions. The first half of the dataset were used as the training set and the second half were used as the test set. To ensure the reliability of the experimental results, we also performed an additional iteration where the training and test sets were exchanged, and then the results from both iterations were averaged.

E. Other Recognition Algorithms for Comparison

The performance of our proposed SMGT was compared with three other methods in motion recognition based on B-mode ultrasound. There are two commonly used image recognition methods that extract features of gray gradient and histogram of oriented gradients from the B-mode ultrasound images, respectively, and then classify the features by using the support vector machine (SVM). Additionally, because convolutional neural network (CNN) models were reported to have good performance in image recognition, a deep CNN model of Resnet152 was adopted as a comparison.

1) Gray Gradient Based on Region of Interest: It has been proved that this feature can reflect local changes of the image and is effective in gesture recognition and finger flexion angle prediction [18]. It was obtained by fitting the grayscale values of the region of interest (ROI) in an image with a 3D plane, and the extraction procedure is shown in formulas (11) and (12). Firstly, a number of circular ROIs was divided from the image, which can be described as:

$$ROI_i = \{(x, y) : (x - x_i)^2 + (y - y_i)^2 \le r^2\} \tag{11}$$

In each ROI, a plane was found to represent the region by regression, and the three parameters α, β, and γ were extracted as features of every ROI.

$$G(ROI_i) \approx \alpha_i(x_i - x) + \beta_i(y_i - y) + \gamma_i \tag{12}$$

2) Histogram of Oriented Gradients (HOG): This feature [29] can capture the edge information of images and rich texture information of local position. Ortenzi et al. [30] firstly used HOG in ultrasound-based motion recognition and proved its effectiveness. In the calculation process of HOG, an image was divided into small cells and the gradient magnitude and direction of the pixels in each cell was computed. These gradients are then quantized into orientation bins to form a histogram of the cell. Next, the cells are grouped into larger blocks, and the feature vectors of each block are concatenated to form a global feature vector, the whole process can be shown as Fig. 6.

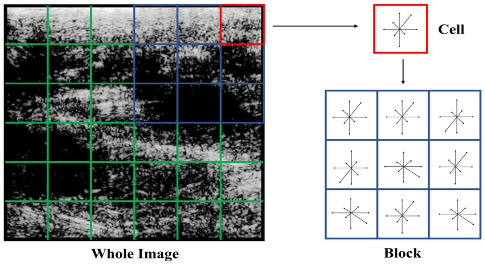


Fig. 6. The calculation process of HOG features.

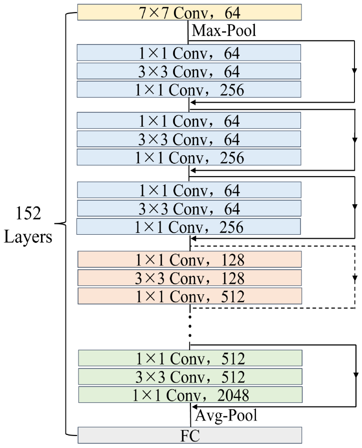


Fig. 7. The architecture of Resnet152.

3) Resnet152: Resnet152 [28] is an image classification model developed by Microsoft Research. It has strong ability to automatically extract and represent complex features, and achieved state-of-the-art performance on various benchmark datasets, such as ImageNet, CIFAR-10, and CIFAR-100. As shown in Fig. 7, the architecture of ResNet152 consists of 152 layers, including convolutional layers, pooling layers, and residual blocks. It uses a 7 × 7 kernel in the initial convolutional layer to capture global information, and subsequent layers utilize 3 × 3 kernels to extract local patterns, while the inclusion of 1 × 1 kernels are incorporated to reduce dimensionality. Moreover, the model's residual learning framework enables it to learn residual mappings instead of direct mappings, allowing for training of very deep neural networks without encountering the problem of vanishing gradients.

F. Performance Evaluation and Statistical Analysis

In this study, four metrics of classification accuracy (CA), precision, recall, and F1 score were used to evaluate the performance of our proposed method. The formulas for these metrics are shown in (13), (14), (15), and (16), where TP is the true positive of motion classifications, FN is the false negative, FP is the false positive and TN is the true negative [31].

Moreover, to assess the significance of the experimental results, a one-way ANOVA was conducted on the classification accuracy in this study with a significance level of $p = 0.05$.

$$CA = \frac{correctly\ classified\ samples}{the\ whole\ test\ samples} \times 100\% \quad (13)$$

$$Precision = \frac{TP}{TP + FP} \quad (14)$$

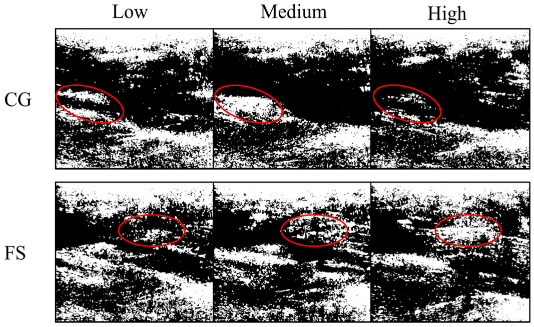


Fig. 8. Binarized ultrasound images of forearm muscles under two different movements and three force levels.

$$Recall = \frac{TP}{TP + FN} \quad (15)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (16)$$

In addition, we also investigated the effectiveness of data augmentation and effects of two important parameters of depth L and number of heads h on the model performance.

III. Results

A. The Ultrasound Images of the Forearm Muscle

Using the flexible ultrasound transducer, we collected ultrasound images of ten movements and three force levels. From the binarized images shown in Fig. 8, it was observed that there are distinctions between different movements and force levels. It indeed revealed the morphological changes during different movements and force levels. However, we also observed that the variation between different force levels is localized and subtle, making it more challenging to distinguish compared to the differences between different movements.

B. Performance of SMGT on the Non-Disabled Subjects

Fig. 9(a) displays the comparative results of four different methods on seven non-disabled subjects. The results indicate that, for the task of this study, our proposed SMGT method achieved a highest average accuracy of 98.4% ± 0.6%. The SMGT method exhibited superior performance in terms of classification accuracy for all seven non-disabled subjects, outperforming the other three methods. Moreover, statistical analysis also proved that the proposed SMGT approach significantly outperformed the other three methods with p-value < 0.05.

In addition to the classification accuracy, we also used precision, recall, and F1 score to measure our experimental results, which are presented in Fig. 10. It can be observed that our SMGT approach performed better than other three methods across all the three metrics, and also exhibited lower variance across different subjects. That indicates its superior robustness on various subjects. The average confusion matrices of seven non-disabled subjects are shown in Fig. 11. It can be seen that the recognition accuracy for all movements and force levels are above 99%, except for the low force level of FP, which achieved an accuracy of 93.6%. These results demonstrated the reliability of our SMGT in recognizing various movements and forces in non-disabled subjects.

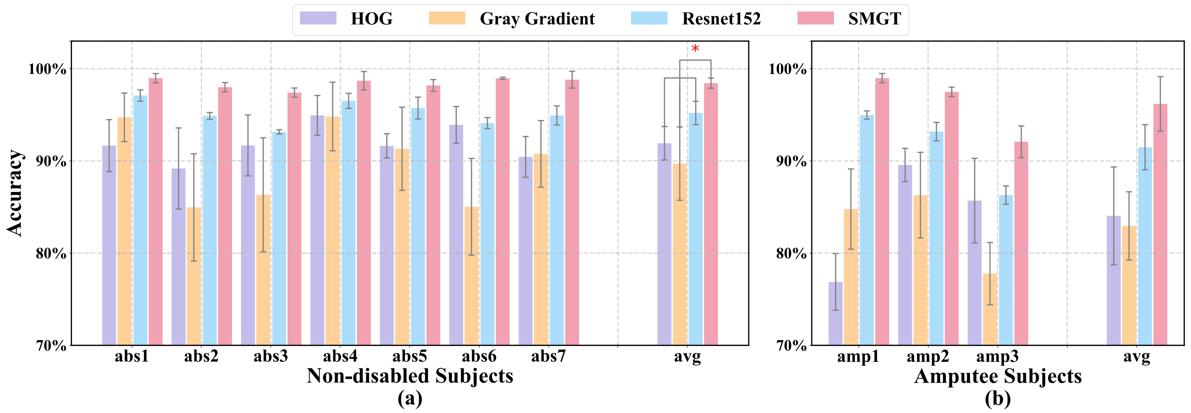


Fig. 9. Classification accuracy of the four methods for both non-disabled and amputee subjects. (a) The result of seven non-disabled subjects. (b) The result of three amputee subjects. (The error bars are the standard errors of two-fold cross-validation).

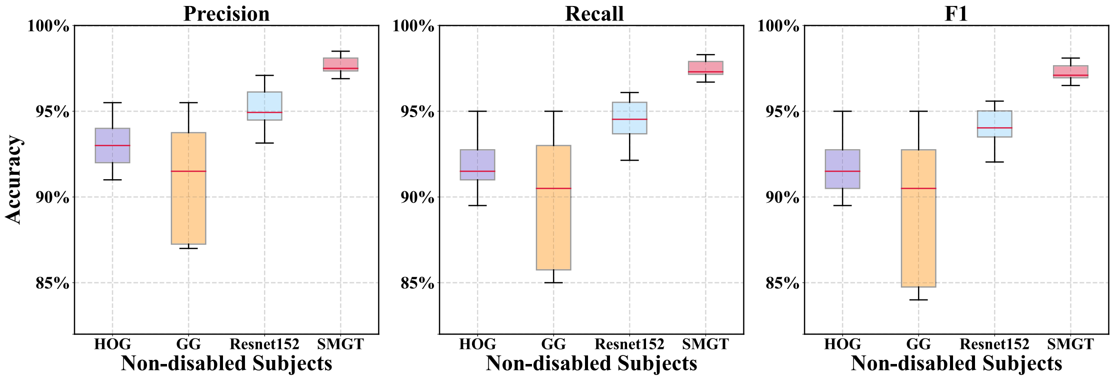


Fig. 10. The comparison of Precision, Recall, and F1 score of the four methods for seven non-disabled subjects. (GG refers to Gray Gradient).

TABLE III
The Precision, Recall and F1 Score of the Four Methods for Three Amputee Subjects

Subjects	HOG			Gray Gradient			Resnet152			**SMGT**		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Amp 1	0.769	0.781	0.770	0.848	0.855	0.846	0.954	0.958	0.947	**0.986**	**0.988**	**0.985**
Amp 2	0.896	0.905	0.894	0.863	0.872	0.862	0.947	0.955	0.942	**0.975**	**0.976**	**0.974**
Amp 3	0.857	0.861	0.856	0.778	0.794	0.772	0.864	0.869	0.862	**0.938**	**0.954**	**0.935**

C. Performance of SMGT on the Amputee Subjects

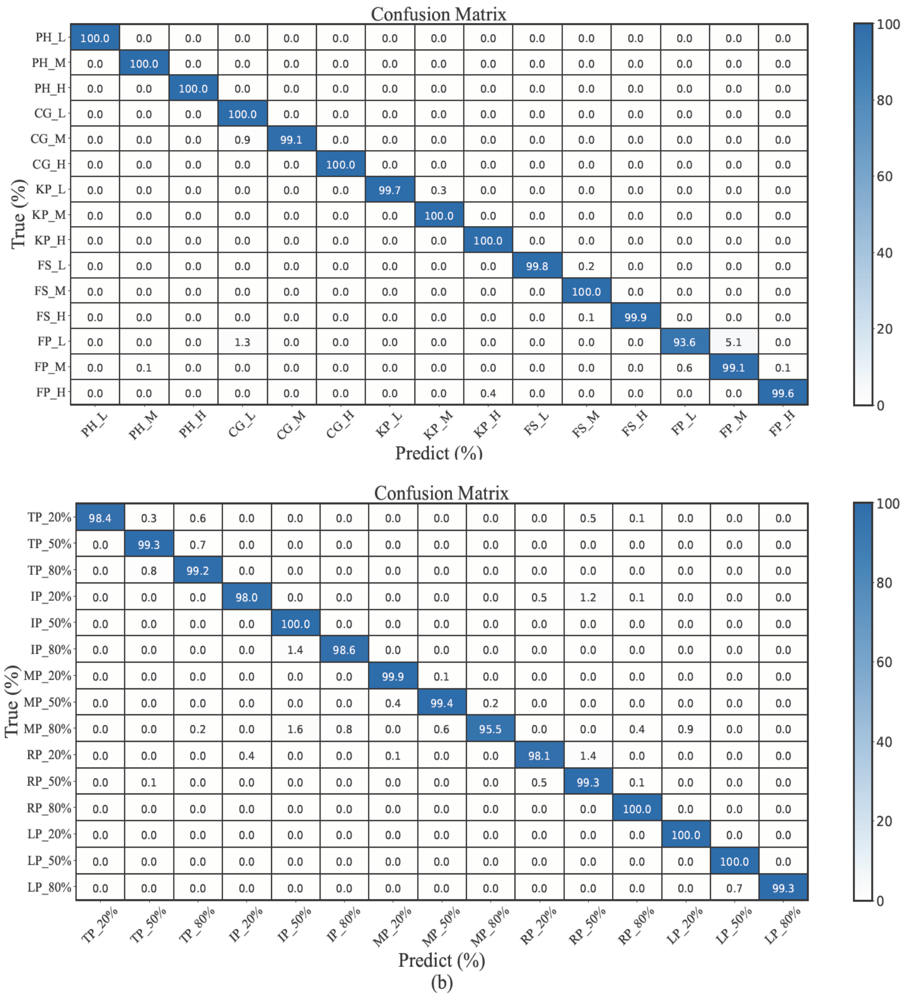
The classification accuracy for the three amputee subjects is presented in Fig. 9(b), which indicates that our proposed SMGT method achieved the highest average classification accuracy of 96.1% ± 2.9% among the four different methods. Although the classification accuracy of each method decreased compared to that in the non-disabled subjects, the relative ranking between methods did not change: SMGT > Resnet152 > HOG > Gray Gradient. The precision, recall, and F1 score of the three amputees are shown in Table III, and the performance of our method is still the best under these three metrics. These findings demonstrate the superiority of our method when applied to amputee subjects. As shown in Fig. 12, although the recognition accuracy dropped to about 82% for some movements and force levels, most misclassified samples are within adjacent force levels, with relatively few misclassifications for different movements classes.

D. Data Augmentation Effectiveness

The necessity and effectiveness of data augmentation in our SMGT model were verified through a comparative experiment conducted with and without data augmentation. The results are summarized in Table IV. It can be observed that the addition of data augmentation improved the accuracy for all ten subjects, with an average improvement of 1.81%. Notably, Amp 3 who had the lowest classification accuracy of 89.9% among the ten subjects achieved the highest improvement of 3.9% when data augmentation applied.

E. Parameter Sensitivity of the Proposed SMGT

The parameter depth L and number of heads h are the most important parameters in transformer encoder. Thus, we assessed the effects of these two parameters on the model performance. The results in Fig. 13(a) shows that, as the depth L gradually increases from 1, the model performance



	PH_L	PH_M	PH_H	CG_L	CG_M	CG_H	KP_L	KP_M	KP_H	FS_L	FS_M	FS_H	FP_L	FP_M	FP_H
PH_L	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PH_M	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PH_H	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
CG_L	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
CG_M	0.0	0.0	0.0	0.9	99.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
CG_H	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
KP_L	0.0	0.0	0.0	0.0	0.0	0.0	99.7	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0
KP_M	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
KP_H	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0
FS_L	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	99.8	0.2	0.0	0.0	0.0	0.0
FS_M	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0
FS_H	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	99.9	0.0	0.0	0.0
FP_L	0.0	0.0	0.0	1.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	93.6	5.1	0.0
FP_M	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	99.1	0.1
FP_H	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0.0	99.6

	TP_20%	TP_50%	TP_80%	IP_20%	IP_50%	IP_80%	MP_20%	MP_50%	MP_80%	RP_20%	RP_50%	RP_80%	LP_20%	LP_50%	LP_80%
TP_20%	98.4	0.3	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.1	0.0	0.0	0.0
TP_50%	0.0	99.3	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TP_80%	0.0	0.8	99.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
IP_20%	0.0	0.0	0.0	98.0	0.0	0.0	0.0	0.0	0.0	0.5	1.2	0.1	0.0	0.0	0.0
IP_50%	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
IP_80%	0.0	0.0	0.0	0.0	1.4	98.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MP_20%	0.0	0.0	0.0	0.0	0.0	0.0	99.9	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MP_50%	0.0	0.0	0.0	0.0	0.0	0.0	0.4	99.4	0.2	0.0	0.0	0.0	0.0	0.0	0.0
MP_80%	0.0	0.0	0.2	0.0	1.6	0.8	0.0	0.6	95.5	0.0	0.0	0.4	0.9	0.0	0.0
RP_20%	0.0	0.0	0.0	0.4	0.0	0.0	0.1	0.0	0.0	98.1	1.4	0.0	0.0	0.0	0.0
RP_50%	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	99.3	0.1	0.0	0.0	0.0
RP_80%	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0
LP_20%	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0
LP_50%	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0
LP_80%	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	99.3

Fig. 11. Average confusion matrices of non-disabled subjects. (a) Average confusion matrix for functional grasp movements and corresponding force levels. (H, L, and M represent high, low, and medium force levels, respectively.) (b) Average confusion matrix for fine finger movements and corresponding force levels. (20%, 50%, and 80% represent different force levels relative to MVC).

TABLE IV

EFFECTIVENESS OF DATA AUGMENTATION IN SMGT

Sub	Without(%)	With(%)
Abs 1	99.17	99.48
Abs 2	95.95	98.50
Abs 3	97.21	97.91
Abs 4	98.00	99.70
Abs 5	97.52	98.82
Abs 6	97.47	99.08
Abs 7	96.98	99.73
Amp 1	96.27	98.27
Amp 2	96.26	97.46
Amp 3	89.90	93.80
Average	96.47 ± 2.37	98.28 ± 1.66

improves. However, at a depth of 5, the model reached a saturation point with a recognition accuracy of 97.88%, and further increasing the depth L did not lead to a significant improvement. Moreover, increasing depth L lead to an increase in computational cost due to the growing number of model parameters. Additionally, in Fig. 13(b), it can be observed that increasing the number of heads h did not significantly improve the performance of the model. However, the optimal performance is achieved when h equals 12, with a recognition accuracy of 97.93%.

TABLE V

THE ALGORITHM COMPUTATION TIME OF DIFFERENT METHODS

Num	Methods	Time Cost per Image (ms)
1	HOG	13.3
2	Gray Gradient	39.1
3	Resnet152	54.7
4	SMGT	28.5

F. Algorithm Computation Time

We compared the algorithm computation time of the SMGT method with those of the other three methods, as shown in Table V. The algorithm computation time here refers to the duration taken to process a raw image and output the corresponding class label for that image using the pretrained model. As Resnet152 and our SMGT are deep learning methods that require high computing power, they were conducted on a GPU (TITAN V, NVIDIA Inc, Santa Clara, California, USA), the other two machine learning methods of HOG and Gray Gradient were conducted on a CPU (Ryzen 7 6800H, AMD Inc, Santa Clara, California, USA), the results are presented

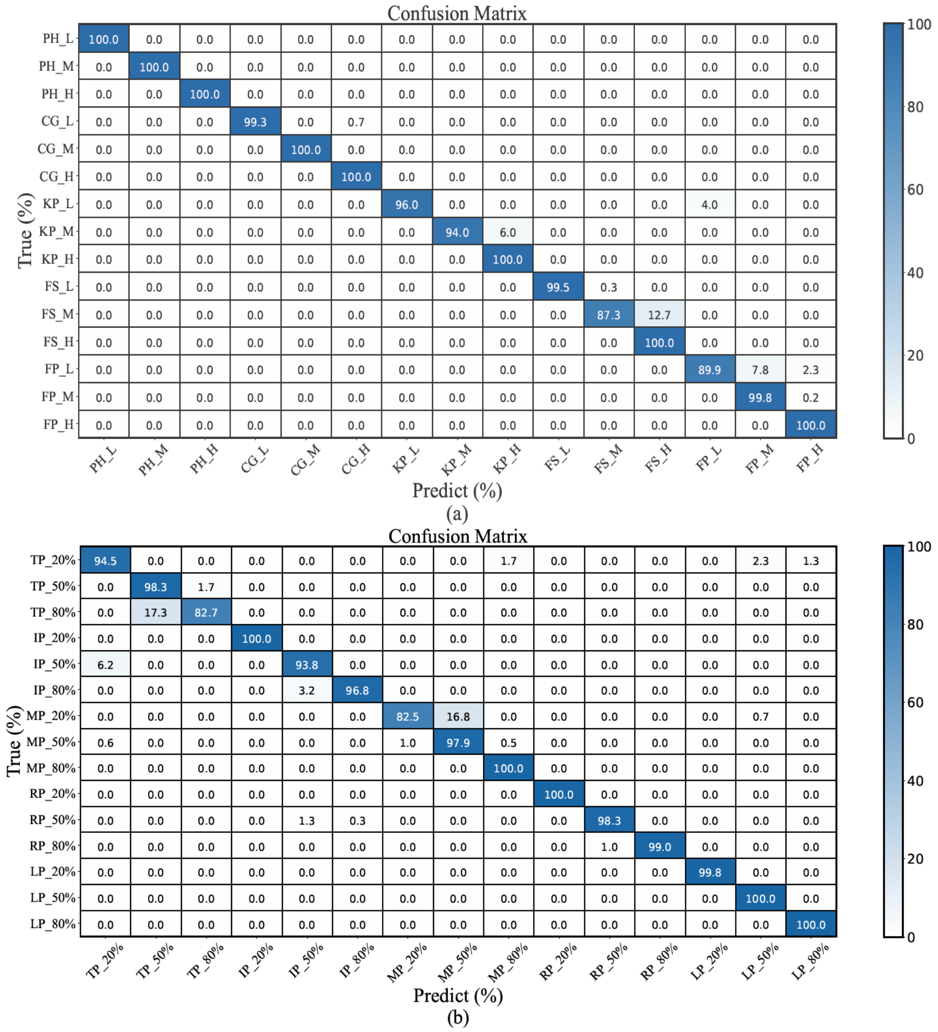


Fig. 12. Average confusion matrices of amputee subjects. (a) Average confusion matrix for functional grasp movements and corresponding force levels. (b) Average confusion matrix for fine finger movements and corresponding force levels.

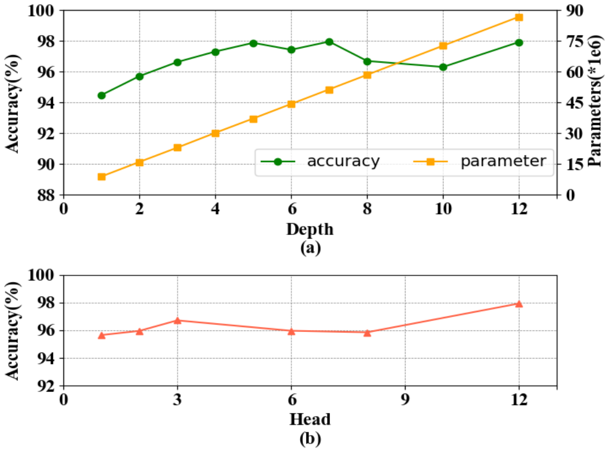


Fig. 13. Parameter sensitivity. (a) Effect of depth on performance and the number of parameters. (b) Effect of the number of heads on model performance.

in Table V. The results showed that with CUDA (Compute Unified Device Architecture) acceleration on the GPU, our method can achieve an average prediction time of only 28.5 ms per image.

IV. Discussion

To the best of our knowledge, although there have been some studies on motion recognition using B-mode ultrasound, few of them have specifically focused on the simultaneous recognition of movements and force levels. In this study, we proposed a novel SMGT model to recognize the movements and force levels simultaneously, and tested the feasibility and effectiveness of our model in both non-disabled and amputee subjects.

A. Feasibility and Superiority of the Proposed SMGT

In this study, we evaluated the performance of four different methods, and the proposed SMGT achieved the highest average classification accuracy of 98.4% ± 0.6% in non-disabled subjects and 96.2% ± 3.0% in amputees. In contrast, the methods of HOG and Gray Gradient only achieved the worst classification accuracy of 91.9% ± 1.8%, 89.7% ± 3.0% in non-disabled subjects, and 84.0% ± 3.7%, 82.9% ± 2.4% in amputees respectively. The challenges of precise recognition in this study arise from the subtle morphological variations in muscles between different force levels of the same movements, as depicted in Fig. 8. These subtle difference makes the simultaneous recognition task in this study resembling a fine-grained image classification task, where the objective is to distinguish fine-grained images that with similar appearance features but difficult to differentiate. Consequently, traditional image feature-based methods, like HOG and Gray Gradient, faced difficulties to accurately capture these subtle differences. These observations are consistent with the findings of Ortenzi et al. [30], who also found difficult in recognizing force using these two image features. Moreover, our Transformer-based SMGT method outperformed the CNN-based Resnet152, with an average recognition accuracy improvement of 3.72%. This is supported by the findings presented in [32], which also proved that transformer structure is more effective than CNN structure when solving fine-grained image classification tasks.

Additionally, we found that amputees achieved worse performance than non-disabled subjects in all the four methods. The possible reason may be the muscle atrophy in the residual limb of amputees, which limited muscle morphological and structural information provided in the ultrasound images which is crucial for hand motion recognition. It is supported by the individual information of subjects in Table I, which indicates amputees have smaller arm circumference than able-bodied subjects. Specifically, amputee 3 who had the smallest arm circumference, exhibited the lowest recognition accuracy among all the subjects, as illustrated in Fig. 9(b).

Furthermore, compared with the results in [33], which also focused on simultaneous recognition of movements and force levels using sEMG, our achieved accuracy of 98.4% ± 0.6% in non-disabled subjects and 96.2% ± 3.0% in amputees are obviously higher than their 86.5% in non-disabled subjects, and 76.3% in amputees. This discrepancy may be due to the fact that B-mode ultrasound can detect the activities of small and deep muscles, which are triggered by fine movements and force variations. In contrast, sEMG which was recorded on the skin surface may not provide the same level of sensitivity and detailed muscle activity information. It shows the distinct advantages of B-mode ultrasound over conventional sEMG in hand movement recognition.

As depicted in Fig. 11 and Fig. 12, the average accuracy for various movements and force levels remained at a high level, demonstrating great robustness to different movements and force levels. Although there were still a few misclassified samples, most of which were classified into adjacent force level classes. Notably, in the average confusion matrix of amputees, we found more samples were misclassified into adjacent force levels. This is probably because even if we have conducted a mirrored bilateral training, it is still difficult to guarantee that amputees can perform movements with their phantom hand at the exact force level for every time.

B. Effectiveness of Data Augmentation

In order to improve the generalization ability of our SMGT model, two data augmentation methods of Cutout and Mixup were applied. According to the results in Table IV, classification accuracies of all the ten subjects were improved by data augmentation ranging from 0.31% to 3.90%. In particular, for Amp 3 (the subject with the poorest performance), the classification accuracy improved from 89.9% to 93.8% with the application of data augmentation. These results demonstrate the necessity of data augmentation and its positive impact on the performance of B-mode ultrasound image based motion recognition. Specifically, Cutout masks a random region of an image by setting the pixel values in that region to 0. This forces the model to learn to recognize features from the unmasked regions of the image, encouraging it to pay attention to the local details. While Mixup combines two randomly chosen samples from the training data to create a new sample. This effectively creates a smooth interpolation between the two samples, forcing the model to learn from the combinatorial features from both samples rather than from one individual sample. Thus Mixup can improve the model's generalization to new and unseen data and reduce overfitting. The analysis above confirmed the appropriateness of the two chosen data augmentation methods.

C. Parameter Effects and Algorithm Computation Time

We explored the effects of two crucial parameters, depth L and number of heads h, on the performance of SMGT. From the results in Fig. 13(a), we found that when the depth L reached 5, the model reached a saturation point with a recognition accuracy of 97.88%. Additionally, increasing the depth beyond 5 did not yield a substantial improvement in performance. However, it did result in an increase in the number of model parameters, leading to longer computation time. The results depicted in Fig. 13(b) indicate that the optimal classification accuracy of 97.93% was achieved when the number of heads was set to 12. However, we did not observe obvious positive correlation between the number of heads and the model performance. This finding is consistent with the results reported in [34], which used a transformer encoder for EEG recognition and found MSA module is not sensitive to the number of heads. These results demonstrated the importance of selecting the appropriate values for the depth and number of heads.

The algorithm computation time of four different methods was evaluated and compared. The results in Table V indicated that our SMGT not only outperformed Resnet152 in terms of accuracy, but also in terms of algorithmic computation time, with a prediction time of only 28.5 ms per image. This indicates that our model can accurately recognize the correct movement and force level based on a muscle ultrasound image within 28.5 ms. Therefore, the proposed SMGT approach exhibits excellent potential for real-time applications.

D. Potential and Benefits of the SMGT for Amputees

By capturing the intricate morphological and structural details of muscles, the SMGT presents a good classification performance of functional grasp movements, fine finger movements and their different force levels. According to the previous study in [35], the response time of a control system should not to be more than 300 milliseconds, so our response time of 128.5 milliseconds is acceptable for the real-time application of the control system, which does not introduce a user-perceived delay. Thus, it is potential to use the SMGT method to achieve a more natural and accurate control of prosthetic hands for amputees, which could greatly improve the quality of their daily lives.

E. Limitation and Future Work

One limitation of this study may be the limited number of subjects. Therefore, in the future work, we will recruit more non-disabled and amputee subjects to validate the effectiveness of the proposed method SMGT. Additionally, the size and weight of the ultrasound transducer can be further reduced to improve the wearing comfort of users. Furthermore, a wired connection between the ultrasound system and the computer was required for transmitting and processing the ultrasound images. The wired connection may lead to discomforts to the users. So, in the future work, we plan to develop wireless wearable ultrasound transducers those can greatly improve the comfort of users and be used for long-time monitoring of human activates.

V. Conclusion

In this study, we used a self-designed lightweight, flexible, and wearable ultrasound transducer for data acquisition and proposed a novel Sonomyography Transformer (SMGT) model for simultaneously recognizing hand movements and force levels. By tested on seven non-disabled subjects and three amputees, our approach achieved an average classification accuracy of 98.4% ± 0.6% in non-disabled subjects and 96.2% ± 3.0% in amputee subjects. Furthermore, through a systematically comparison with three commonly used methods, the superiority of our SMGT is proved in terms of recognition accuracy and computation time. In general, this study may promote the applications of intelligent prosthetic hands and rehabilitation engineering.

References

[1] T. A. Kuiken, "Targeted muscle reinnervation for real-time myoelectric control of multifunction artificial arms," *JAMA*, vol. 301, no. 6, p. 619, Feb. 2009.

[2] M. A. Oskoei and H. Hu, "Myoelectric control systems—A survey," *Biomed. Signal Process. Control*, vol. 2, no. 4, pp. 275–294, Oct. 2007.

[3] C. Wang et al., "Bioadhesive ultrasound for long-term continuous imaging of diverse organs," *Science*, vol. 377, no. 6605, pp. 517–523, Jul. 2022.

[4] V. Nazari and Y.-P. Zheng, "Controlling upper limb prostheses using sonomyography (SMG): A review," *Sensors*, vol. 23, no. 4, p. 1885, Feb. 2023.

[5] J.-Y. Guo, "Dynamic monitoring of forearm muscles using one-dimensional sonomyography system," *J. Rehabil. Res. Develop.*, vol. 45, no. 1, pp. 187–196, Dec. 2008.

[6] J. Yan, X. Yang, X. Sun, Z. Chen, and H. Liu, "A lightweight ultrasound probe for wearable human–machine interfaces," *IEEE Sensors J.*, vol. 19, no. 14, pp. 5895–5903, Jul. 2019.

[7] X. Yang, X. Sun, D. Zhou, Y. Li, and H. Liu, "Towards wearable A-mode ultrasound sensing for real-time finger motion recognition," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 6, pp. 1199–1208, Jun. 2018.

[8] W. Xia, Y. Zhou, X. Yang, K. He, and H. Liu, "Toward portable hybrid surface electromyography/A-mode ultrasound sensing for human–machine interface," *IEEE Sensors J.*, vol. 19, no. 13, pp. 5219–5228, Jul. 2019.

[9] J. Zeng, Y. Sheng, Y. Yang, Z. Zhou, and H. Liu, "Cross modality knowledge distillation between A-mode ultrasound and surface electromyography," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–9, 2022.

[10] J. Li, K. Zhu, and L. Pan, "Wrist and finger motion recognition via M-mode ultrasound signal: A feasibility study," *Biomed. Signal Process. Control*, vol. 71, Jan. 2022, Art. no. 103112.

[11] Y. P. Zheng, M. M. F. Chan, J. Shi, X. Chen, and Q. H. Huang, "Sonomyography: Monitoring morphological changes of forearm muscles in actions with the feasibility for the control of powered prosthesis," *Med. Eng. Phys.*, vol. 28, no. 5, pp. 405–415, Jun. 2006.

[12] J. Shi, J.-Y. Guo, S.-X. Hu, and Y.-P. Zheng, "Recognition of finger flexion motion from ultrasound image: A feasibility study," *Ultrasound Med. Biol.*, vol. 38, no. 10, pp. 1695–1704, Oct. 2012.

[13] Y. Huang, X. Yang, Y. Li, D. Zhou, K. He, and H. Liu, "Ultrasound-based sensing models for finger motion classification," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 5, pp. 1395–1405, Sep. 2018.

[14] X. Yang, D. Zhou, Y. Zhou, Y. Huang, and H. Liu, "Towards zero re-training for long-term hand gesture recognition via ultrasound sensing," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 4, pp. 1639–1646, Jul. 2019.

[15] J. He, H. Luo, J. Jia, J. T. W. Yeow, and N. Jiang, "Wrist and finger gesture recognition with single-element ultrasound signals: A comparison with single-channel surface electromyogram," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 5, pp. 1277–1284, May 2019.

[16] A. J. Fernandes, Y. Ono, and E. Ukwatta, "Evaluation of finger flexion classification at reduced lateral spatial resolutions of ultrasound," *IEEE Access*, vol. 9, pp. 24105–24118, 2021.

[17] K. Bimbraw, C. J. Nycz, M. Schueler, Z. Zhang, and H. K. Zhang, "Simultaneous estimation of hand configurations and finger joint angles using forearm ultrasound," *IEEE Trans. Med. Robot. Bionics*, vol. 5, no. 1, pp. 120–132, Feb. 2023.

[18] C. Castellini, G. Passig, and E. Zarka, "Using ultrasound images of the forearm to predict finger positions," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 20, no. 6, pp. 788–797, Nov. 2012.

[19] J. McIntosh, A. Marzo, M. Fraser, and C. Phillips, "EchoFlex: Hand gesture recognition using ultrasound imaging," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Denver, CO, USA, May 2017, pp. 1923–1934.

[20] N. Akhlaghi et al., "Real-time classification of hand motions using ultrasound imaging of forearm muscles," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 8, pp. 1687–1698, Aug. 2016.

[21] A. T. Kamatham, M. Alzamani, A. Dockum, S. Sikdar, and B. Mukherjee, "Sparse sonomyography-based estimation of isometric force: A comparison of methods and features," *IEEE Trans. Med. Robot. Bionics*, vol. 4, no. 3, pp. 821–829, Aug. 2022.

[22] W. Chen et al., "Flexible ultrasound transducer with embedded optical shape sensing fiber for biomedical imaging applications," *IEEE Trans. Biomed. Eng.*, vol. 70, no. 10, pp. 2841–2851, Oct. 2023.

[23] E. N. Kamavuako, D. Farina, K. Yoshida, and W. Jensen, "Estimation of grasping force from features of intramuscular EMG signals with mirrored bilateral training," *Ann. Biomed. Eng.*, vol. 40, no. 3, pp. 648–656, Mar. 2012.

[24] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–12.

[25] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[26] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," 2017, *arXiv:1708.04552*.

[27] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," 2017, *arXiv:1710.09412*.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[29] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jul. 2005, pp. 886–893.

[30] V. Ortenzi, S. Tarantino, C. Castellini, and C. Cipriani, "Ultrasound imaging for hand prosthesis control: A comparative study of features and classification methods," in *Proc. IEEE Int. Conf. Rehabil. Robot. (ICORR)*, Aug. 2015, pp. 1–6.

[31] S. Boughorbel, F. Jarray, and M. El-Anbari, "Optimal classifier for imbalanced data using Matthews correlation coefficient metric," *PLoS ONE*, vol. 12, no. 6, Jun. 2017, Art. no. e0177678.

[32] J. He et al., "TransFG: A transformer architecture for fine-grained recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 852–860.

[33] B. Fang et al., "Simultaneous sEMG recognition of gestures and force levels for interaction with prosthetic hand," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 2426–2436, 2022.

[34] Y. Song, Q. Zheng, B. Liu, and X. Gao, "EEG conformer: Convolutional transformer for EEG decoding and visualization," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 710–719, 2022.

[35] K. Englehart and B. Hudgins, "A robust, real-time control scheme for multifunction myoelectric control," *IEEE Trans. Biomed. Eng.*, vol. 50, no. 7, pp. 848–854, Jul. 2003.