A Multimodal Framework for Estimating Six-Dimensional Force During Dynamic Human-Robot Interaction Using Deep Learning

Gao Lin, Fei Wang, *Member, IEEE*, Xu Zhong, Zida An, and Shuai Han

***Abstract*—Interaction force estimation using surface electromyography (sEMG) is a popular technique for applications such as powered exoskeletons, robotics, and rehabilitation. However, since the sEMG is non-stationary, it is difficult to estimate high-dimensional interaction forces accurately during physical human-robot interaction (pHRI). This work proposes an end-to-end six-dimensional force estimation framework that can accurately predict forces during dynamic pHRI from cartesian and joint space. Firstly, the sEMG, pose, and velocity of the upper limb during pHRI are synchronously acquired. Subsequently, sEMG and kinematic information are deeply fused by tensor fusion and cross-attention fusion. Finally, a spatio-temporal neural network (STNN) is utilized to extract the features of the fused information and estimate the interaction force. Six subjects interact with the robot at four stiffness: 100N/m, 150N/m, 200N/m, and 250N/m. The proposed multimodal fusion scheme achieves excellent performance in different types of STNNs and is validated in different kinematic spaces. Among them, the highest $\mathbf{R^2}$ of 0.969 is achieved using ConvLG in cartesian space. Compared to solely employing sEMG, the $\mathbf{R^2}$ of force estimation based on multiple modalities increases by 21.4%-42.0% ($\mathbf{p < 0.05}$). It shows the effectiveness of the presented approach and contributes a new way to estimate high-dimensional force during dynamic pHRI.**

***Index Terms*—Physical human-robot interaction (pHRI), Force estimation, Surface electromyography (sEMG), Multimodal fusion, Spatio-temporal neural network (STNN).**

I. Introduction

As social needs become increasingly diverse, people have high hopes for robots to interact more closely, safely, and efficiently with humans in diversified scenarios such as industrial manufacturing, life services, and rehabilitation medicine. Physical human–robot interaction (pHRI) is a form of collaboration between humans and robots that involves a task-oriented and continuous dynamic process with temporal characteristics in a shared physical operating space, which is more targeted than human–robot interaction (HRI) [1]. In pHRI systems, it is crucial to capture interaction force accurately. On the one hand, it can improve security and collaboration efficiency [2]. The other aspect is imitation learning, as physical contact is one of the most effective ways for robots to learn human skills [3]. Force sensors can measure interaction force. However, they have several limitations, including high cost, extra weight, and complex design and implementation in limited space and harsh environments [4]. Therefore, accurate estimation of interaction force is necessary for various applications such as powered exoskeletons, robotics, and rehabilitation systems.

Since demonstrating the relationship between muscle electrical activity and generated force in humans [5], surface electromyography (sEMG) has been widely used as a non-invasive estimator of the generated force and joint torque. Decoding sEMG can provide natural and intelligent pHRI, which helps to realize human-centered interaction [6]. sEMG are physiological signals from the superposition of motor unit action potentials generated by muscle fibers. sEMG has the characteristic of being 30-150ms ahead of the movement, which has been widely employed to solve the human-robot coupling problem caused by the time lag of signal processing and system response [7]. Information regarding human dynamics, such as joint torque [8], [9] and stiffness [10], can be obtained by decoding sEMG. Consequently, sEMG is a potential characterization of human dynamics with great potential for estimating interaction force.

Traditional sEMG-force estimation is realized by building a musculoskeletal model (parametric model). For example, the Hill model, constructed based on the properties of muscles and tendons, is regarded as a practical method for estimating muscle forces [11]. Hayashibe *et al.* [12] propose a multiscale physiological model for force estimation with more minor modeling errors than Hill. Romero *et al.* [13] analyze the effectiveness of different Hill models in force estimation and suggest selecting appropriate Hill models to solve specific tasks. However, the force estimation methods based on the musculoskeletal models are highly parameter-dependent and sensitive. Furthermore, they are highly individualized and cannot be transferred, which limits their application.

Given the limitations of parametric models, there has been a growing interest in developing non-parametric methods

This work was supported in part by the Foundation of National Natural Science Foundation of China under Grant 62373086, 62373087, Liaoning Revitalization Talents Program under Grant XLYC2203013. *(Corresponding author: Fei Wang and Shuai Han).*

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 Northeastern University under Application No. NEU-EC-2023B031S. (Followed the World Medical Association's Declaration of Helsinki for medical research involving humans.)

Gao Lin, Fei Wang, and Zida An are with the Faculty of Robot Science and Engineering, Northeastern University, Shenyang 110000, China (e-mail: 2310767@stu.neu.edu.cn; e-mail: wangfei@mail.neu.edu.cn; e-mail: 2110702@stu.neu.edu.cn).

Xu Zhong is with the Medical Engineering Department, Affiliated Hospital of Yangzhou University, Yangzhou 225000, China (e-mail: 092690@yzu.edu.cn).

Shuai Han is with the Department of Neurosurgery, Shengjing Hospital of China Medical University, Shenyang 110000, China (e-mail: hanshuai19870217@outlook.com).

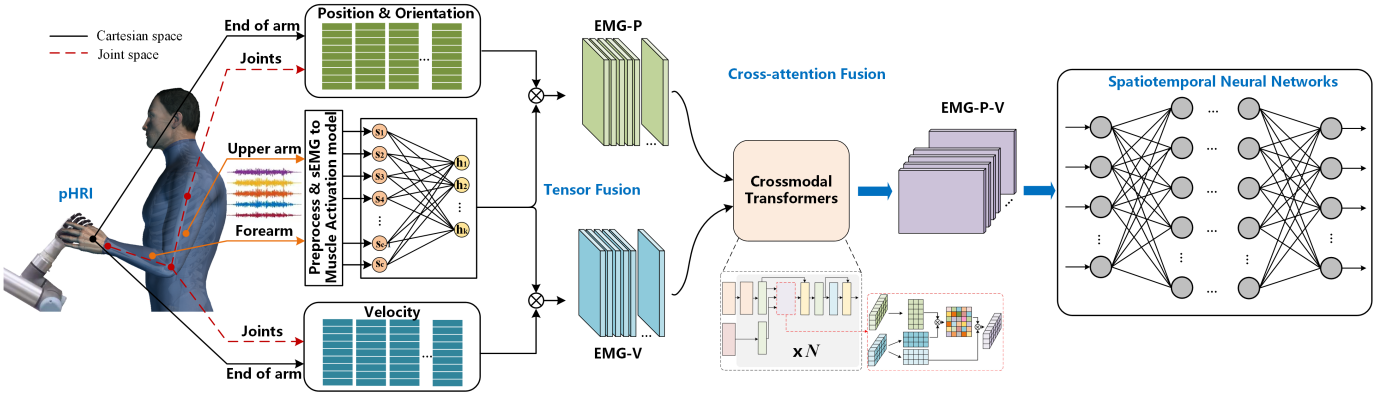

Fig. 1. The proposed method's scheme. The marker "$\otimes$" denotes the matrix multiplication. $s_i(i = 1, ..., c)$ represents the sEMG information in the muscle activity space and $h_i(i = 1, ..., k)$ is the activation information in the muscle synergy space. EMG-P, EMG-V, and EMG-P-V denote the inter-modal interactions of sEMG-pose, sEMG-velocity, and sEMG-pose-velocity, respectively.

based on machine learning and deep learning for sEMG-force estimation. Khoshde *et al.* [14] propose an artificial neural network (ANN) for estimating knee torque and realizing interaction with a rehabilitation robot. Jing *et al.* [15] propose a three-domain fuzzy wavelet neural network (TDFWNN), which outperforms the traditional radial basis function neural network (RBFNN) in estimating the hand force (2-Dof). Wei *et al.* [16] propose an end-to-end model integrating ResNet and bi-directional long short-term memory (Bi-LSTM) for estimating hand force (1-Dof). Xue *et al.* [17] employ a convolutional neural network (CNN) to predict finger grip force (1-Dof) during static interaction for manipulator control. Hang *et al.* [18] propose a deep CNN to predict the interaction force (1-Dof) at the end-of-arm, which outperforms the robustness of LSTM. Qin *et al.* [19] utilize a Bayesian filter and LSTM to predict thrust and torque (2-Dof) at the end-of-arm for the task of tightening a screw. However, a single sEMG modality has inherent limitations, including low robustness, stability, sensitivity, and resolution for complex motions. These make it challenging to conduct comprehensive and reliable force estimation for pHRI systems with high degrees of freedom (Dof). Therefore, the sEMG-based multimodal fusion scheme for interaction force estimation is worth exploring.

The fusion of sEMG with other modalities is employed to enhance the information expressed and facilitate data complementarity, yielding a more comprehensive, stable, and accurate estimation. The sEMG has been widely fused with inertial measurement units (IMU) [20], vision [21], ultrasound (US) [22], [23], acoustic signals [24], electroencephalogram (EEG) [25], and near-infrared (NIR) [26] to enhance the estimation performance. Regarding interaction force estimation, Qiang *et al.* [22] utilize sEMG fused with the US to predict ankle joint torque, and the performance is better than using sEMG alone. Zou *et al.* [23] employ sEMG and A-mode US as inputs to a self-attentive CNN (OLR-SACNN) to predict hand force during static interaction, and the performance is better than solely employing sEMG or US. Xiong *et al.* [27] propose a fusion scheme of sEMG and arm poses, combined with a generalized regression neural network (GRNN), to predict the interaction force of the end-of-arm in the direction of horizontal or vertical motion. Ning *et al.* [28] predict elbow torque by fusing sEMG and elbow angles for upper limb rehabilitation robot control. Gelareh *et al.* [29] propose a deep multimodal CNN extracts features from EMG and elbow angle and aggregates them to obtain elbow torque, and the performance improves significantly when the kinematic information is included. In conclusion, current multimodal force estimation is still focused on tasks with static interaction or low-Dof. The upper limb is a complex and highly flexible system of the human body. The accurate and stable estimation of the upper limb interaction force under complex pHRI with multiple-Dof remains a challenge.

This work proposes an end-to-end estimation method of six-dimensional interaction force under complex pHRI. The general framework of the proposed method is shown in Fig. 1. Fusion information containing sEMG and kinematics is obtained by tensor fusion and cross-attention fusion, and interaction forces are regressed with a spatio-temporal neural network (STNN). This method does not require prior information for parametric models and manually selected features for non-parametric models. The main novelty of this research can be twofold: In terms of force estimation, this is the inaugural study to utilize sEMG and kinematic data to predict six-dimensional interaction forces during dynamic pHRI. In terms of information fusion, the fusion scheme of sEMG, pose, and velocity that we designed has excellent prediction performance in intention recognition. The contributions of this work are as follows:

1) We analyze the synergistic activation effects of forearm and upper arm muscle groups under dynamic pHRI and provide a quantitative measurement method for the synergistic activation of upper limb muscles.

2) We propose a hierarchical multimodal fusion scheme to fuse sEMG and kinematic information deeply, which ameliorates the problem of low accuracy and poor stability of force estimation using sEMG.

3) We construct a variety of STNNs suitable for estimating interaction force, which provides models for reference in the field of force estimation.

4) We synchronously collect a new and comprehensive

dataset containing sEMG, poses and velocities of the end-of-arm, joint angles and angular velocities of the upper limb, and interaction force under various interaction stiffness.

5) The effects of neural networks, kinematic spaces, and interacting stiffness on the performance are discussed to provide a multi-perspective and comprehensive analysis of interaction force estimation.

The rest of this paper is organized as follows: Section II provides detailed information on the relevant methods. Section III presents the experimental protocol and results. Section IV discusses this work. The last part is about the conclusion and future work.

II. METHODS

This section will introduce muscle synergy extraction, multimodal fusion technique, force estimation models, baseline methods, and evaluation standards.

A. Muscle synergy extraction

The synergistic activation information of the forearm and upper arm muscle groups is extracted for multimodal fusion. Algorithms commonly used for muscle synergy extraction include non-negative matrix factorization (NMF), principal component analysis (PCA), independent component analysis (ICA), and factor analysis (FA). PCA, ICA, and FA translate many indicators into a few. NMF decomposes data into target forms, and the decomposed elements have non-negative natural sparse characteristics. The decomposition form and results of NMF are more interpretable for muscle activation [30]. First, the raw sEMG signals are rectified, and max-min normalized. Subsequently, we employ NMF to convert the preprocessed sEMG signals from muscle activity space to synergy space:

$$S_{c\times n} = W_{c\times k}H_{k\times n} + E_{c\times n}, \tag{1}$$

where W is the synergy matrix representing the basis vectors of the synergy space. H is the activation coefficient matrix from the muscle synergy space. E is the residual matrix. c, n, and k are the number of sEMG channels, data samples, and the dimension of muscle synergies, respectively. The optimal W and H are obtained by minimizing the Euclidean distance between S and the reconstructed data $W \times H$:

$$\min\|\frac{1}{2}|S - WH\|^2, W \geq 0, H \geq 0. \tag{2}$$

The higher VAF indicates that more synergies contain more useful information of the sEMG. The synergy extraction is repeated, and the number of synergies increases from one to sixteen. The number of synergies is then selected as the minimum number that achieved an averaged VAF > 0.95 [31]. The VAF value is calculated by

$$VAF = 1 - \frac{\|S - S_r\|}{\|S - mean(S)\|^2}, \tag{3}$$

where S_r is the reconstructed muscle activation matrix. Muscle synergy space H is used for multimodal fusion.

B. Tensor fusion

Tensor fusion maps information from different modalities into a high-dimensional tensor through an outer product operation. This high-dimensional tensor can represent higher-order interactions between different modalities, thus capturing more complex interactions and enhancing the expression of the model. Compared with concatenation or weighted summation, tensor fusion can capture the dynamics between modalities and provide richer information expression [32]. An outer product forms this dynamic modeling method between modalities, so it has no learnable parameters. The likelihood of overfitting is low, even though the output tensor is high-dimensional [33]. Extensive research has shown that fusing pose information can improve the stability and accuracy of sEMG prediction [27], [28]. Reference [34] has demonstrated that motion velocity significantly impacts the efficacy of sEMG estimation. In order to retain as much valuable information as possible, we perform information fusion from the data layer. The inter-modal interactions of sEMG-pose and sEMG-velocity are established by

$$X_{\alpha}^{L\times d} = X_e^L \otimes X_p^d, \tag{4}$$

$$X_{\beta}^{L\times k} = X_e^L \otimes X_v^k. \tag{5}$$

X_e, X_p, and X_v respectively represent sEMG, pose, and velocity information in the same unit time. L, d, and k denote the data dimensions of the sEMG, pose, and velocity. X_α and X_β are the interaction matrices of the sEMG-pose and sEMG-velocity, respectively. By means of the above outer product operation, the two modalities are generated into a high-dimensional tensor ($X_{\alpha}^{L\times d}$ or $X_{\beta}^{L\times k}$), whose elements $X^{ij}(i = 1, ..., L; j = 1, ..., d/k)$ represent all possible combinations between the corresponding components in the two modalities.

C. Cross-attention fusion

This section describes the proposed cross-attentional fusion module. The crossmodal transformer learns attention between two modalities (sEMG-pose and sEMG-velocity). It repeatedly reinforces the other modality with low-level information from one modality, capturing inter-modal correlations. The cross-modal transformer we designed does not have an encoder-decoder structure but consists of multiple stacks of cross-attention blocks. The cross-attention blocks are designed to focus on low-level information while eliminating self-attention. Furthermore, a multi-head attention mechanism is utilized to enhance pattern richness.

1) Cross-attention: The cross-attention operation shown in Fig. 2 can be defined by

$$CA(X_\alpha, X_\beta) = softmax(\frac{X_\alpha W_{Q_\alpha} W_{K_\beta}^T X_\beta^T}{\sqrt{d_k}})X_\beta W_{V_\beta}. \tag{6}$$

Here, $X_\alpha \in R^{L_\alpha \times d_k}$ and $X_\beta \in R^{L_\beta \times d_k}$ represent two different modalities. X_α as query, X_β as key and value. W_{Q_α}, W_{K_β}, and W_{V_β} are the weights corresponding to query, key, and value. T denotes transpose operation. Furthermore, the

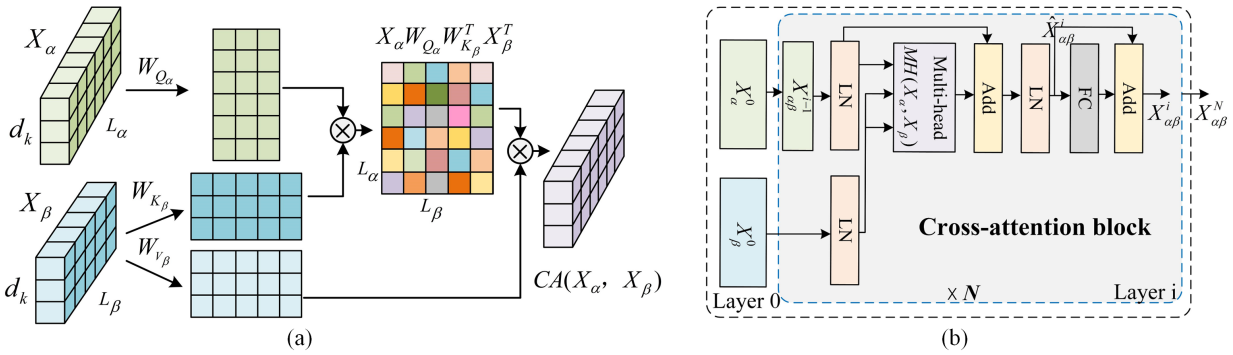

Fig. 2. Architectural elements of cross-attention fusion. (a) Cross-attention mechanism. (b) Crossmodal Transformer.

cross-attention can be implemented in a multi-head manner with a linear map W_{MH}:

$$MH(X_\alpha, X_\beta) = Concat(H_1, H_2, \cdots, H_h)W_{MH}. \quad (7)$$

Each attentional head H_i is calculated with the learnable weight set $\theta_i^{(Q,K,V)}$:

$$H_i = CA(X_\alpha, X_\beta | \theta_i^{(Q,K,V)}, L_\alpha \times L_\beta). \quad (8)$$

The dimension of the attention coefficient is $L_\alpha \times L_\beta$.

2) Crossmodal transformer: Crossmodal transformer consists of N layers of cross-attention blocks. Formally, a crossmodal transformer computes feed-forwardly for $i = 1, \cdots, N$ layers:

$$X_{\alpha\beta}^0 = X_\alpha^0, \quad (9)$$

$$\hat{X}_{\alpha\beta}^i = MH_{\alpha\beta}^i(LN(X_{\alpha\beta}^{i-1}), LN(X_\beta^0)) + LN(X_{\alpha\beta}^{i-1}), \quad (10)$$

$$X_{\alpha\beta}^i = FC_{\alpha\beta}^i(LN(\hat{X}_{\alpha\beta}^i)) + LN(\hat{X}_{\alpha\beta}^i). \quad (11)$$

X_α^0 and X_β^0 represent the initial information of the two modalities, respectively. $X_{\alpha\beta}^i$ is the output of the i-th cross-attention block. When $i = 0$, $X_{\alpha\beta}^0$ is equal to X_α^0. $MH_{\alpha\beta}^i$ represents the multi-head attention of the i-th cross-attention block. This work sets $i = 5$, $MH = 3$, $N = 5$. $FC_{\alpha\beta}^i$ is a fully connected layer, LN means layer normalization. $\hat{X}_{\alpha\beta}^i$ represents the output of the LN layer of the i-th cross-attention block. At each level of the cross-attention block, the low-level signals X_β^0 from modality X_β are transformed into a distinct set of key/value pairs to interact with the $X_{\alpha\beta}^{i-1}$, which repeatedly reinforces inter-modal interactions.

D. Force estimation model

The fusion information is the spatio-temporal sequences. Six STNNs are constructed based on the current mainstream and advanced network structures, capable of predicting the interaction forces end-to-end. The STNNs include CNNGRU, ConvLG (ConvLSTM + GRU), LS-TCN (large-scale temporal convolutional network), CNN-att (CNN-attention), TF (Transformer), and Swin-TF (Swin Transformer). The proposed STNNs are trained by a GPU server with an RTX 4060, and all parameters are the best that we have tried. Furthermore, the related code will be provided for reference via https://github.com/l4655mh?tab=repositories.

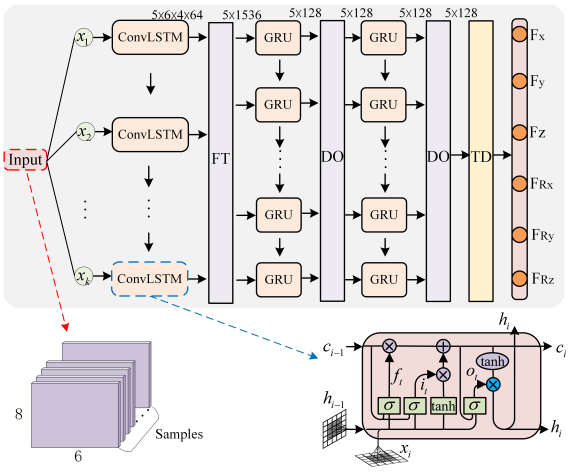

Fig. 3. The structure illustration of ConvLG. Abbreviations FT, DO, and TD represent the flatten layer, dropout layer, and TimeDistributed layer in respective.

1) CNNGRU: CNN and gated recurrent unit (GRU) are two common networks. CNN is widely used in computer vision to extract image features. GRU is widely used for time series prediction and is more lightweight than LSTM. In this work, CNN extracts hidden features in the space and combines them layer by layer to generate abstract high-level features. However, CNN does not consider the temporal correlation of time series data. Consequently, we employ the GRU to capture long-term dependencies in time series data. The input data dimension of the CNNGRU is 5×8×6. The convolutional layer comprises 64 kernels with a size of 1×2, and the activation function is ReLU. The maximum pooling size is 1×2, and the subsequent layers and parameters are identical to ConvLG's.

2) ConvLG: Xingjian [35] proposes multi-layer ConvLSTM as a solution to the problem of precipitation nowcasting. ConvLSTM determines the future state of a certain cell in multi-dimensional features by its locally adjacent input and past states, enabling it to capture spatiotemporal correlations effectively. The inputs x_i, cell outputs c_i, hidden states h_i, and gates(f_t, i_t, and o_t) of the ConvLSTM are 3D tensors. The structure of the presented ConvLG is illustrated in Fig. 3. Firstly, the spatio-temporal features of the fusion information

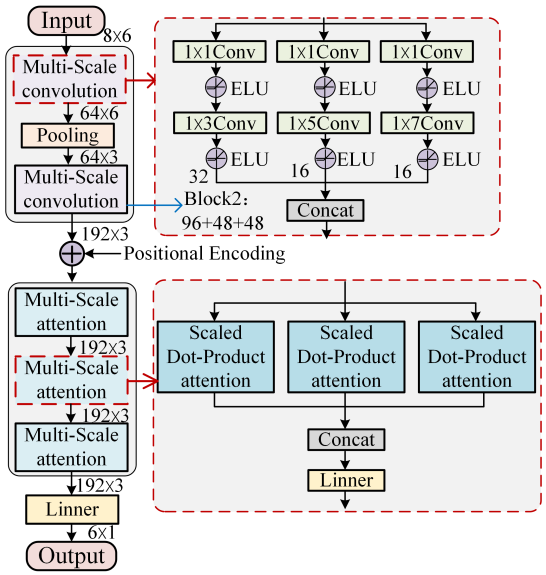

Fig. 4. The structure illustration of CNN-att.

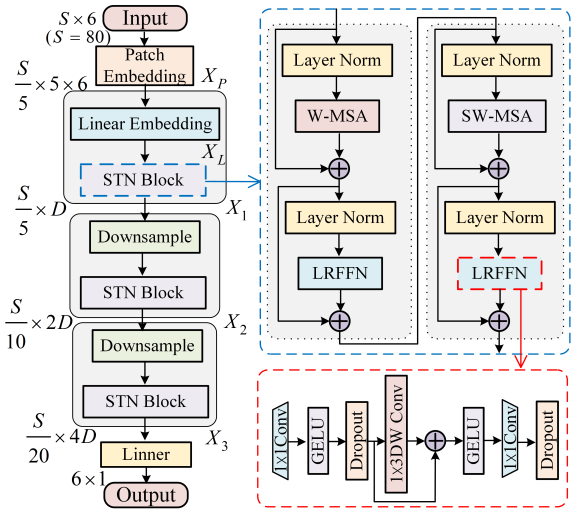

Fig. 5. The structure illustration of Swin-TF. W-MSA and SW-MSA are window-based and shifted window-based self-attention modules, respectively.

are extracted by a layer of ConvLSTM. Then, the features are dimensionally transformed to meet the input requirements of the GRU. Finally, long-term dependencies in the data are captured by multi-layer GRU.

The input data dimension of the ConvLG is 5×8×6×1. The time step is set to 5, with the dimension of the time step data being 8×6×1. The number of convolution kernels in the ConvLSTM layer is set to 64, with the kernel size being 3×3. ReLU is employed as the activation function. The flatten layer is employed to transform the output of the ConvLSTM layer (6×4×64) into a format compatible with the input requirements of the GRU layer. The GRU layer and the dropout layer are then connected in series. Each GRU cell contains 128 hidden units, and ReLU is the activation function. The dropout rate is 0.2 to avoid overfitting. Finally, the interaction forces in six directions (X, Y, Z, RX, RY, and RZ) are output from the TimeDisributed layer. The batch size is 50, with a learning rate of 0.0002. The optimization method is Adam, and the maximum number of iterations is 100.

3) LS-TCN: Chen [36] proposes an LS-TCN to address the limitations of TCN in fully extracting temporal features from sEMG. The LS-TCN employs causal dilation convolutions and residual connections, enabling the modeling of long-term sequences. The proposed LS-TCN is constructed by stacking five layers of TCN for interaction force estimation. The internal structure of the TCN block can be referred to [36]. The convolution kernel size is 2, the dilation rate is [1, 2, 4], and the number of convolution channels is [32, 64, 64, 32, 6]. Finally, the six-dimensional forces are output through two fully connected (FC) layers.

4) CNN-att: CNN-Attention employs a multi-scale convolutional structure with diverse filter sizes, enabling the representation of features by extracting fine and coarse information [37]. Based on research [37], we adopt the convolutional structure of Inception [38] and the design methodology of Transformer [39] to construct the CNN-att, which incorporates a multi-scale convolutional structure, absolute position coding, and a multi-head attention structure. The structure of the presented CNN-att is illustrated in Fig. 4. The output dimensions for the first and second multi-scale convolutional blocks are set to [32, 16, 16] and [96, 48, 48], respectively. The main parameters of the multi-head attention block are as follows: the embedding size is 192, and the dimensions of K, Q, and V are 64, 64, and 128, respectively. We employ three multi-head attention blocks with three-head attention.

5) TF & Swin-TF: The transformer can establish global dependencies of sequence data by a self-attention mechanism and using absolute position encoding to introduce sequence information into the model. Subsequently, Liu et al. [40] propose the Swin Transformer, which effectively reduces the computational complexity by sliding window self-attention. This approach leads to tasks such as image segmentation and target detection. Transformer and Swin Transformer are used for human continuous motion recognition in this work. For the Swin-TF, a locally enhanced feedforward neural network (LEFNN) is constructed to enhance the model's capacity to extract local features, as illustrated in Fig. 5.

For the TF, the primary parameters are as follows: the number of encoders of the decoder layer is 3, the number of heads in multi-head attention is 4, and $d_k = d_q = d_v = 6$. Attn-dropout and fc-dropout are set to 0.5. Six-dimensional forces are output by a fully connected layer.

For the Swin-TF, the input dimensions are [sample/10, 8×10, 6]. In the patch embedding layer, a convolutional layer with a kernel of 5 and a step size of 5 is employed to transform the input matrix into feature vectors. These are done to achieve concatenation of neighboring features in the channel dimensions. Subsequently, the output is reshaped to obtain X_P. The linear embedding layer encodes X_P, and the features are projected into a higher-dimensional linear space to obtain the encoded feature vectors X_L. D is the dimension of the feature space, and D is set to 100. Then, window self-attention is computed by the STN block to enhance the representation

of features. The window size is 10, and the shift size is 5. A LEFFN is employed to integrate local and global information, enhancing the model's capacity to extract local features. (The LEFFN comprises a 1×1 convolution that maps feature vectors to a high-dimensional space, a 1×3 depthwise convolution that extracts local features of the channels, and a residual connection that fuses global and local features. The GELU activation function is utilized to enhance the model's nonlinear representation ability. Finally, the feature vectors are mapped back to the low-dimensional space by 1×1 convolution.). The number of multi-heads in each stage is 2, 4, and 6, respectively. The output of "stage1" is X_1. An output channel of $2\times D$ is employed in the downsample layer for twofold downsampling. The output of "stage2" is X_2. By analogy, the output of "stage3" is X_3. Finally, six-dimensional forces are output by a fully connected layer.

E. Baseline models

We compare the performance of our proposed models with that of force estimation models proposed by other studies. The models we compare are as follows:

1) CNN: The majority of current studies employ CNN to estimate force [17], [18], [29]. We construct the force estimation model proposed in the research [29]. Initially, the Conv2d module is employed to extract spatial information. Subsequently, the features are weighted through a shallow neural network (FC), and a regression layer is employed to output the force information. The model parameters are identical to those described in [29]. The convolution kernel size is 3×3, the pooling size is 3×3, and the number of neurons in the FC layer is 128. The training batch size is 100, the maximum number of iterations is 100, and the dropout rate is 0.5.

2) FFNN: The feedforward fully connected neural network (FFNN) has been demonstrated to perform with significant efficacy in tasks of force estimation [22], [41]. We reproduce the model from the research [41]. The FFNN comprises four fully connected blocks and one regression block. Each FC block contains one linear, ReLU, and dropout layer. The number of neurons in each block is 128, 64, 32, and 16, respectively. The batch size for training is 50, the maximum number of iterations is 100, and the dropout rate is 0.3.

3) GCN-LSTM: Graph convolution network (GCN), representing a more recent generation of neural networks, has recently been employed in the domain of force estimation [8]. We replicate the model structure from the research [8]. Firstly, GCN is combined with a two-layer LSTM to capture spatiotemporal features. Subsequently, force information is output through two fully connected layers. The fusion information is spatially distributed in an 8×6 array in a regular Euclidean graph. Each vertex uses its adjacent vertices as features, and the feature dimension is 4. The adjacency matrix size is 48×48. The fused information is divided into segments of length six along the time dimension and input into the graph convolution layer as a graph tensor of 6 timesteps×48 vertices×4 feature sizes. The LSTM, fully connected layer, and model training parameter settings are consistent with the research [8].

F. Performance evaluation standards

The root mean square error ($RMSE$) and R-square (R^2) are frequently utilized as performance evaluation standards for EMG-based regression tasks. These standards will be employed to assess the regression performance of the interaction force. The formulas are as follows:

$$RMSE = \sqrt{\frac{1}{Q}\sum_{i=1}^{Q}(F_i^{real} - F_i^{pred})^2}, \tag{12}$$

where F^{real} and F^{pred} are the measured and estimated interaction force, respectively. Q is the length of the sequence. The $RMSE$ gives the index of the square root of the mean of all the square errors.

$$R^2 = 1 - \frac{\sum_{i=1}^{Q}(F_i^{real} - F_i^{pred})^2}{\sum_{i=1}^{Q}(F_i^{real} - \overline{F_i^{pred}})^2}. \tag{13}$$

The value of R^2 can be used to reflect the quality of the model. This value ranges from 0 to 1. The closer R^2 is to 1, the stronger the interpretation ability of the model to the estimated result. We made a one-way Analysis of Variance (ANOVA) to provide statistical analysis under a significant level of 0.05. If $p < 0.05$ is met, one approach will be significantly better than the other.

III. Experiment and Result

A. Experiment protocol

Six male subjects (age: 25.2 ± 1.2years, weight: 67.3 ± 7.6kg, height: 174 ± 4.9cm, mean ± SD) participated in the experiment to validate the proposed method. Firstly, EMG and IMU sensors will be attached to their body. Subsequently, all IMU sensors will be calibrated while the subject is in the static pose shown in Fig. 7(a). Finally, the subject will interact with the robot at stiffness of 100N/m, 150N/m, 200N/m, and 250N/m. The sEMG signals of the forearm and upper arm, the poses and velocities of the end-of-arm, and the angles and angular velocities of the joints are acquired simultaneously. The frequency of data acquisition is 200Hz. Five data trials are acquired under each stiffness. One data set is used for training, and the other four are used for testing. Then, we perform a cross-validation. The length of each trial is 1 minute and a break of 3 minutes between two trials. There is a 10-minute rest before changing stiffness to avoid muscle fatigue. The study was approved by the Northeastern University Human Participants Ethics Committee (NEU-EC-2023B031S), and consents were obtained from the subjects.

B. Data acquisition

To facilitate the system's portability and wearability, the sEMG acquisition device employed is an eight-channel Myo armband from Thalmic Labs. Two Myo armbands are worn following the official standard, as illustrated in Fig. 6(a). The Myo armband is well-known in the field of sEMG-based recognition. It is a low-cost consumer-grade device with good portability, easy wearability, and non-stickiness. There is no need to worry about the electrodes falling off during motion. The Myo armband transmits data via Bluetooth at a frequency

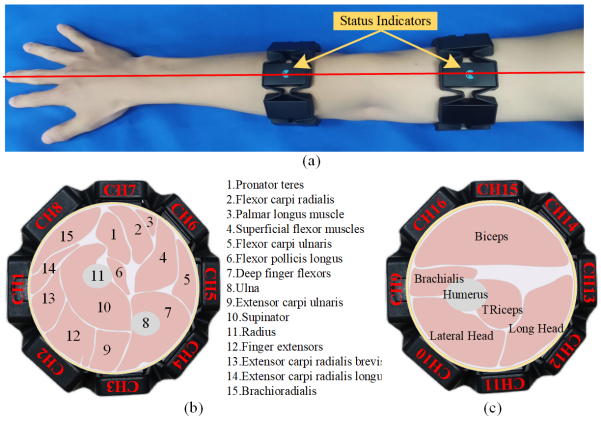

Fig. 6. (a) illustrate the Myo sensor's location. (b) and (c) illustrate the position of the muscle groups corresponding to each electrode, respectively.

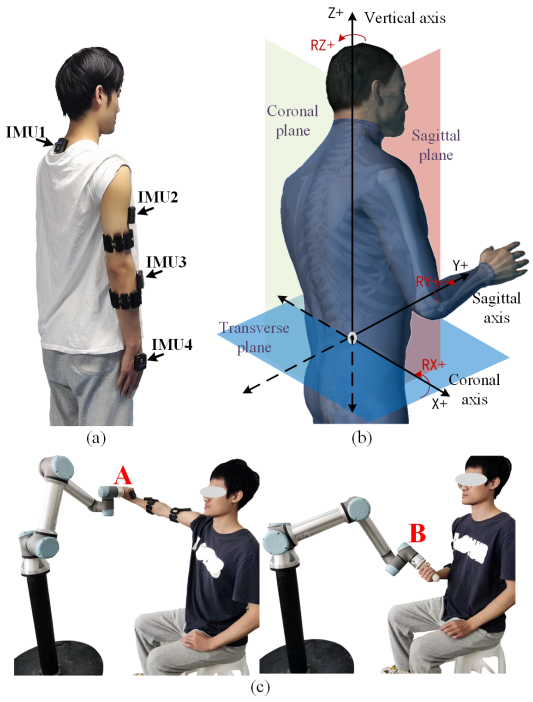

Fig. 7. (a) IMU sensor's location. (b) Data acquisition standard. (c) Interaction paradigm

of 200Hz. The forearm and upper arm muscles corresponding to the 16 acquisition channels are shown in Fig. 6(b) and Fig. 6(c), respectively.

Upper limb joint angles (shoulder flexion/extension, shoulder adduction/abduction, shoulder internal/external rotation, elbow flexion/extension, forearm pronation/supination, wrist flexion/extension, and wrist adduction/abduction) and corresponding angular velocities are acquired using Noraxon myoMotion sensors with an acquisition frequency of 200Hz. As illustrated in Fig. 7(a), four myoMotion sensors are affixed to the participant's anterior segment of the upper spine, upper arm, forearm, and hand, respectively. The joint angles and corresponding angular accelerations will be automatically calculated by Noraxon MR3 software. The angular velocity of

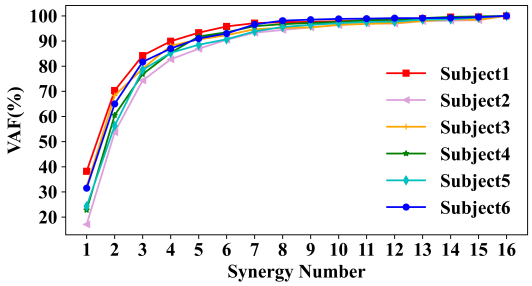

Fig. 8. Averaged variance accounted for (VAF) curve for synergy numbers.

each joint is obtained by integrating the angular acceleration against time and removing the drift effect of the accelerometer by a polynomial fit [42].

The UR5e cobot is employed to capture interaction forces, poses, and velocities of the end-of-arm. The UR5e has 6-Dof, a 5kg payload, and an 850mm working radius. It is furnished with a six-dimensional force sensor at the end to fulfill the experimental requirements. When collecting sEMG, the UR5e's acquisition interfaces are called to synchronously collect the tool coordinate system's interaction forces, poses, and velocities. The acquisition frequency is 200Hz.

C. Interaction paradigm

In order to analyze the synergy activation of muscle groups during pHRI from a biomechanical perspective and formulate a unified data acquisition standard, we overlap the orientations of the human section coordinate system established based on anatomy and the base coordinate system of the UR5e cobot. As illustrated in Fig. 7(b), the UR5e's X, Y, and Z directions correspond to the coronal, sagittal, and vertical axes of the basic section of the human body, respectively. The analysis of interaction force in a single dimension enables the composition of muscle activation during compound motion to be understood and the synergistic activation of muscles in the forearm and upper arm to be analyzed biomechanically. Furthermore, we design an interaction motion that simultaneously covers 6-Dof in cartesian space for estimating six-dimensional interaction forces. As illustrated in Fig. 7(c), an interaction motion can be expressed from state A to state B and back to state A. The motion process encompasses both translations in the X, Y, and Z directions and rotations in the RX, RY, and RZ directions.

D. Result

1) Influence of muscle synergy numbers: Fig. 8 shows the averaged VAF curves across different interacting stiffness for each subject, with the synergy number ranging from one to sixteen. According to the criteria that VAF > 0.95 and a small increment with an additional muscle synergy, eight muscle synergies are appropriate to describe the muscle activation during pHRI. Therefore, we select the synergy number as eight for the subsequent experiments.

TABLE I
RMSE (FORCE ± STD) OF STNNS IN CARTESIAN SPACE

Direction	CNNGRU	ConvLG	LS-TCN	CNN-att	TF	Swin-TF	CNN	FFNN	GCN-LSTM
X	2.840±0.440	**2.729±0.444**	3.149±0.444	3.117±0.432	3.414±0.512	6.402±0.745	4.775±0.592	4.166±0.538	3.120±0.431
Y	2.556±0.362	**2.491±0.396**	2.741±0.428	2.857±0.364	2.904±0.454	5.143±0.643	3.856±0.468	3.416±0.478	2.859±0.382
Z	2.612±0.419	**2.490±0.404**	2.906±0.471	2.926±0.406	3.250±0.592	5.851±0.731	4.233±0.547	3.645±0.454	2.923±0.424
RX	0.129±0.021	**0.126±0.020**	0.133±0.021	0.142±0.020	0.136±0.022	0.192±0.026	0.154±0.026	0.160±0.027	0.143±0.019
RY	0.199±0.032	**0.194±0.029**	0.227±0.038	0.228±0.039	0.252±0.052	0.423±0.062	0.309±0.047	0.311±0.042	0.230±0.038
RZ	0.162±0.032	**0.158±0.031**	0.177±0.036	0.183±0.037	0.193±0.041	0.277±0.056	0.216±0.046	0.221±0.047	0.181±0.036
$\overline{\text{xyz}}$	2.612±0.419	**2.570±0.414**	2.932±0.448	2.921±0.401	3.189±0.519	5.799±0.706	4.288±0.535	3.742±0.490	2.967±0.412
$\overline{\text{Rxyz}}$	0.163±0.028	**0.159±0.027**	0.179±0.032	0.177±0.032	0.194±0.038	0.297±0.048	0.226±0.039	0.231±0.039	0.185±0.031

TABLE II
RMSE (FORCE ± STD) OF STNNS IN JOINT SPACE

Direction	CNNGRU	ConvLG	LS-TCN	CNN-att	TF	Swin-TF	CNN	FFNN	GCN-LSTM
X	2.779±0.458	**2.706±0.418**	3.128±0.640	3.348±0.571	3.545±0.627	6.234±0.974	4.701±0.849	4.275±0.942	3.352±0.643
Y	2.469±0.383	**2.419±0.364**	2.691±0.526	2.965±0.499	2.974±0.576	5.069±0.909	3.913±0.770	3.913±0.568	2.687±0.551
Z	2.564±0.399	**2.516±0.387**	2.938±0.544	3.248±0.548	3.456±0.667	6.122±1.044	4.320±0.875	3.789±0.679	3.248±0.702
RX	**0.132±0.025**	**0.132±0.024**	0.138±0.026	0.151±0.024	0.145±0.026	0.205±0.029	0.159±0.029	0.168±0.035	0.149±0.024
RY	0.225±0.051	**0.219±0.046**	0.262±0.061	0.282±0.059	0.294±0.074	0.498±0.107	0.347±0.078	0.338±0.062	0.285±0.067
RZ	0.169±0.037	**0.164±0.031**	0.185±0.045	0.200±0.045	0.205±0.045	0.303±0.069	0.229±0.053	0.228±0.048	0.208±0.044
$\overline{\text{xyz}}$	2.604±0.413	**2.547±0.390**	2.919±0.570	2.923±0.539	3.325±0.623	5.808±0.975	4.311±0.831	3.847±0.730	3.096±0.632
$\overline{\text{Rxyz}}$	0.175±0.037	**0.172±0.034**	0.195±0.044	0.194±0.043	0.215±0.049	0.335±0.068	0.245±0.054	0.245±0.049	0.214±0.045

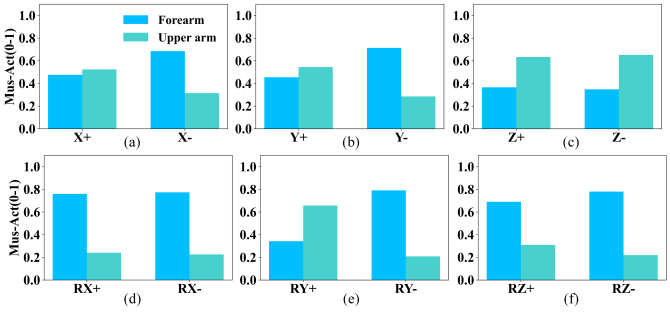

Fig. 9. The averaged synergistic activation of the upper arm and forearm muscle groups under different interacting dimensions.

2) Muscle groups synergistic activation: Fig. 9 shows the averaged synergistic activation of the upper arm and forearm muscle groups under different interacting dimensions for all subjects. Fig. 9(a) shows the translational interactions of the upper limb along the horizontal plane to the right (X+) and the left (X-). The rightward translation with abduction and external rotation of the shoulder and flexion and extension of the elbow. Upper arm muscle activation is slightly higher than the forearm. The leftward translation is performed with adduction and internal rotation of the shoulder and flexion and extension of the elbow. The forearm and upper arm muscle activation ratio is about 7:3. Fig. 9(b) represents the forward (Y+) and backward (Y-) translational interactions. The forward interaction corresponds to shoulder flexion and elbow extension. At this point, there is an external inward thrust on the upper limb, and the upper arm is slightly more activated than the forearm. The backward interaction corresponds to shoulder extension and elbow flexion. At this time, the external manifests as an outward pull on the upper limb, with a muscle activation ratio of about 7:3 for the forearm and upper arm. Fig. 9(c) shows the upward and (Z+) downward (Z-) translational interactions, corresponding to vertical elbow flexion and extension, respectively. At this time, the muscle activation is mainly dominated by the upper arm, which is about 60%. In Fig. 9(d) and Fig. 9(f), RX+, RX-, RZ+, and RZ- correspond to radial flexion, ulnar flexion, palmar flexion, and dorsi flexion of the wrist, respectively. The ratio of the forearm and upper arm muscle activation for radial flexion, ulnar flexion, and dorsiflexion is 8:2, while that of palmar flexion is 7:3. Fig. 9(e) represents the interaction of pronation and supination of the forearm. Our findings indicate that the upper arm muscles are more active than the forearm muscles during supination, with an activation ratio of approximately 6.5:3.5. Conversely, pronation is 2:8. Because the primary moving muscle is the supinator when the forearm is supinated, the auxiliary muscle is the biceps and the triceps also has moderate activity. When the forearm is pronated, the primary moving muscle is the pronator quadratus, the auxiliary muscles are the pronator teres and flexor carpi radialis, and the triceps brachii have slight activity. These quantitative results are consistent with the results of biomechanical analysis.

3) Regression performance: The $RMSE$ performance of three convolution and RNN-based schemes involve CNNGRU, ConvLG, and GCN-LSTM; two temporal and attentional convolution-based schemes involve LS-TCN and CNN-att; two attention-based methods involving TF and Swin-TF, and two FC-based methods involve CNN and FFNN are in Table 1 and Table 2. There is a significant disparity in the RMSE corresponding to X, Y, and Z and RX, RY, and RZ. This is because the measurement standards for force and torque are distinct. In general, the $RMSE$ is within 7.3N and 0.61rad/N for all methods for all forces and torques, respectively. The R^2 of these methods is in Fig .10. R^2 can be a better standard in that the measurement standards of force and torque are different, and the force and torque ranges of different people are diverse.

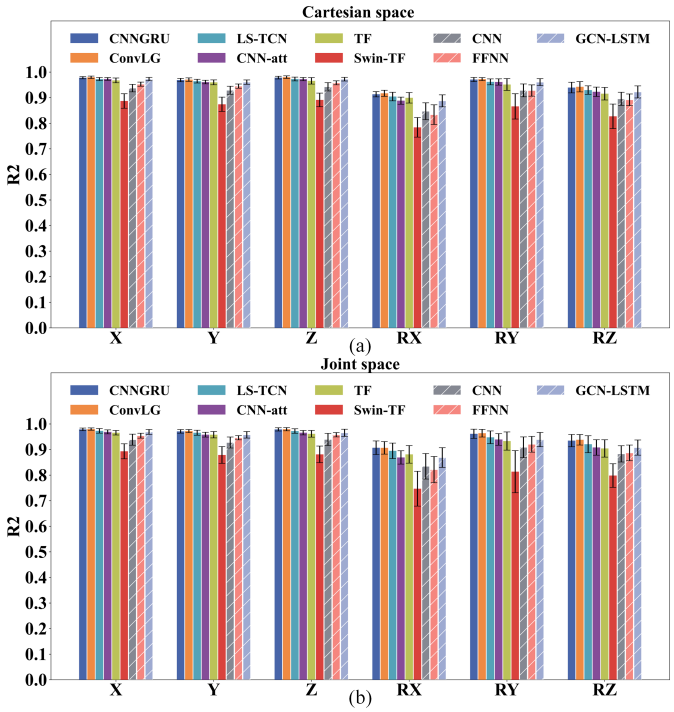

Fig. 10. The R^2 for X, Y, Z, RX, RY, and RZ of proposed and baseline methods. (a) Cartesian space. (b) Joint space.

Among the models, ConvLG achieves the smallest $RMSE$ and the highest R^2 in each dimension. The prediction performance of ConvLG is superior to that of other schemes. In the three convolution and RNN-based schemes(CNNGRU, ConvLG, and GCN-LSTM), ConvLG has superior $RMSE$ and R^2 than CNNGRU in cartesian and joint space. It indicates that ConvLSTM is more adept at extracting spatial information than CNN, which is attributed to the fact that ConvLSTM considers the temporal correlation of information when extracting spatial features. Additionally, the $RMSE$ and R^2 of CNNGRU and ConvLG are superior to those of GCN-LSTM. Since GCN employs convolutional operations based on adjacency and feature matrices, this approach cannot ensure that every vertex in the graph is related to its surrounding vertices, which can lead to data distortion. According to $RMSE$ and R^2, the force estimation performance of GCN-LSTM is comparable to that of LS-TCN and CNN-att. In comparison to other STNNs, attention-based TF and Swin-TF exhibited larger $RMSE$ and lower R^2. TF is generally suited to larger-scale data sets, and the force estimation performance based on global attention is superior to window attention in our study. Except for Swin-TF, the two fully connected regression-based baseline methods, CNN and FFNN, show the largest $RMSE$ and standard deviation and the smallest R^2 due to their inability to capture long-term temporal dependencies. The larger standard deviation indicates that the CNN and FFNN models are less stable. Furthermore, the force regression performance of FFNN is slightly better than that of CNN, which shows that the feature extraction of deep FC is more effective than that of local and shallow networks.

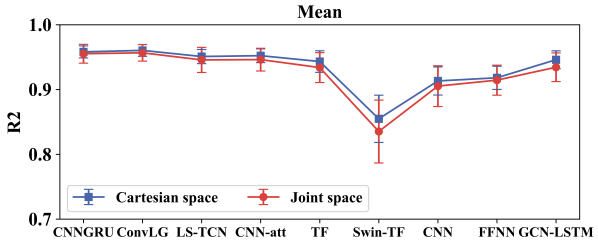

Fig. 11. Overall performance of proposed and baseline methods.

4) Forces of X, Y, Z, RX, RY, and RZ: As illustrated in Table 1 and Table 2, the force of X exhibits the most significant $RMSE$ values, while the force of RX exhibits the smallest. However, it cannot be verified that the force of RX has optimal estimation performance because their force or torque boundary is different. Therefore, the R^2 can be a better standard to evaluate the prediction performance. The R^2 of the different dimensions for all methods on average are illustrated in Fig. 10. The R^2 of the seven methods for the three directions of X, Y, and Z are similar and have small variances. In contrast, the R^2 difference in the RX, RY, and RZ directions is more obvious and has a larger variance. It indicates that the prediction of the translational directions is more accurate and stable. The interactive motion causes this because the force changes in the translational directions are smoother and have a greater range of variation than in the rotational directions. Moreover, the R^2 of RY is considerably higher than that of RX and RZ, which opposes $RMSE$. The reason can be the complicated interaction between the hand and the robot or the effect of the contact area.

5) Overall performance: The overall R^2 for each method is illustrated in Fig. 11. Based on the maximum R^2 criteria, ConvLG achieves the best force estimation performance under both cartesian and joint space. In cartesian space, the average R^2 of three temporal convolution-based CNNGRU, ConvLG, and LS-TCN are 0.958 ± 0.009, 0.960 ± 0.009 and 0.951 ± 0.011, within the range of 0.940-0.969. The three attention-based CNN-att, TF, and Swin-TF are 0.952 ± 0.011, 0.943 ± 0.017, and 0.855 ± 0.036, within 0.819-0.963. The three baseline methods, CNN, FFNN, and GCN-LSTM, are 0.913 ± 0.022, 0.918 ± 0.018, and 0.946 ± 0.014, within 0.891-0.960. The other methods are significantly better than Swin-TF ($p < 0.05$). Methods except Swin-TF are significantly better than the two baseline methods, CNN and FFNN ($p < 0.05$). CNNGRU and ConvLG are better than LS-TCN, CNN-att, TF, and GCN-LSTM ($p < 0.05$), but they have no statistically significant difference ($p > 0.05$). In addition, there is no statistically significant difference among LS-TCN, CNN-att, TF, and GCN-LSTM ($p > 0.05$), which indicates that GCN-LSTM as an STNN has almost the same force estimation performance as the above three methods. The statistical analysis results in joint space are consistent with cartesian space. We randomly select the results of subject 2 at 150N/m to plot his interaction force curve. The tracking trajectories of the ConvLG in cartesian and joint space are shown in Fig. 12(a) and Fig. 12(b). The results show that the proposed

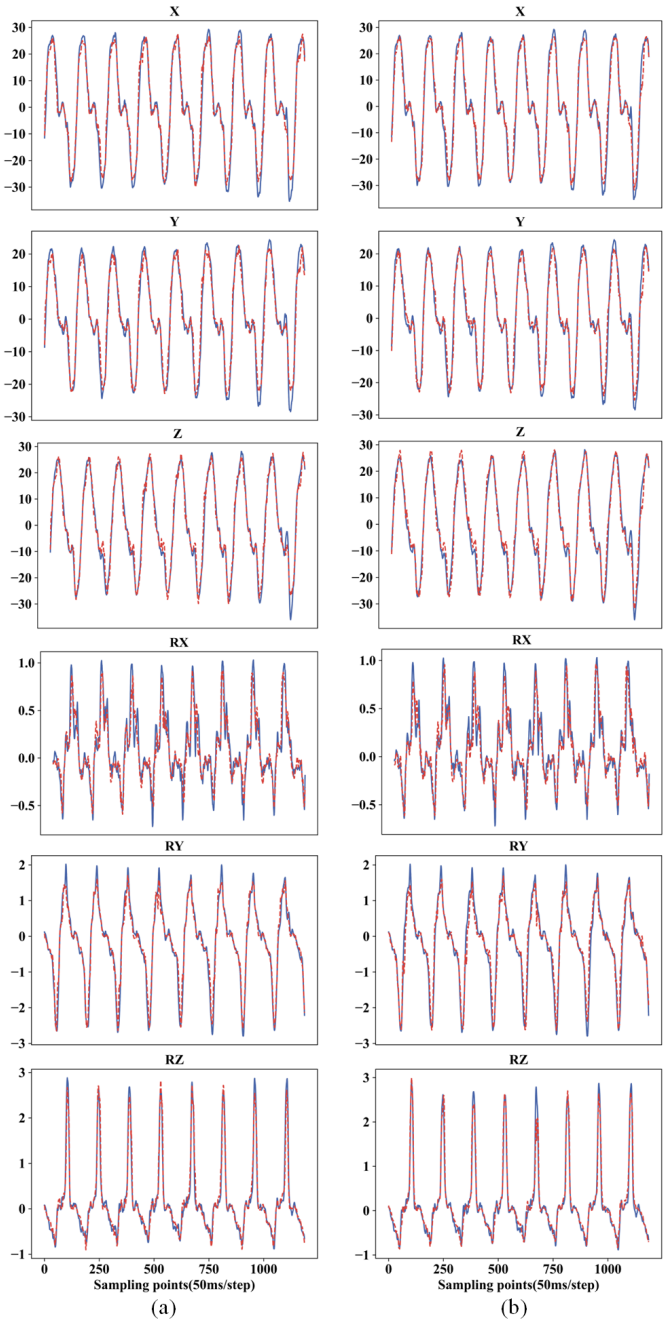

Fig. 12. Interaction force curve of subject 2 at 150N/m. (a) and (b) are the prediction effects in cartesian and joint space, respectively. The solid blue line is the target force, while the dotted red line is the estimated force. The horizontal axes are time with a step of 50ms/point, and the longitudinal axes are force(N) or torque(rad/N).

STNNs have excellent prediction performance. The interactive motion we designed considers the physical interaction in six different dimensions. The proposed methods are evaluated quantitatively with standards of $RMSE$ and R^2 to analyze the performance, and excellent performance has been achieved.

6) Effects of kinematic spaces: Table 1, Table 2, and Fig. 10 show that the performance of force estimation in cartesian and joint space are each high or low in a single dimension. For example, the cartesian space will have better estimation

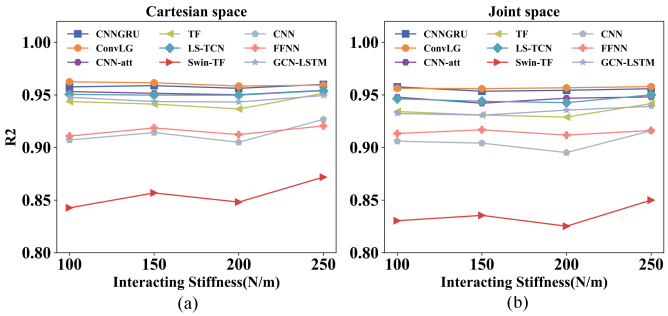

Fig. 13. Overall R^2 under different interacting stiffness. (a) Cartesian space. (b) Joint space.

performance in the translational directions with the ConvLG but worse performance than the joint space in the rotational directions. The R^2 of CNNGRU, ConvLG, LS-TCN, CNN-att, TF, Swin-TF, CNN, FFNN, and GCN-LSTM in cartesian space are 0.958 ± 0.009, 0.960 ± 0.009, 0.951 ± 0.011, 0.952 ± 0.011, 0.943 ± 0.017, 0.855 ± 0.036, 0.913 ± 0.022, 0.918 ± 0.018, and 0.946 ± 0.014, respectively. The R^2 corresponding to the joint space are 0.955 ± 0.015, 0.957 ± 0.013, 0.946 ± 0.019, 0.946 ± 0.018, 0.934 ± 0.023, 0.835 ± 0.049, 0.905 ± 0.032, 0.915 ± 0.023, and 0.934 ± 0.022, respectively. The average R^2 scores are higher in cartesian space than in joint space, indicating that the estimates are more accurate in cartesian space. Moreover, the standard deviation in cartesian space is smaller, indicating that the estimation performance is stabler. The kinematic information of the upper limb in joint space is 7-Dof, whereas in cartesian space it is 6-Dof. 6-Dof is sufficient to represent the spatial motion of the end-of-arm, so there is a redundancy of information in joint space. This redundant information can lead to distortion by mutual crosstalk and more difficult extraction of valuable information, resulting in a degradation of the estimation performance.

7) Influence of interacting stiffness: The performance of the different interacting stiffness is shown in Fig. 13. The results demonstrate that all methods can achieve higher R^2 at 250N/m. The force estimation performance is influenced by two factors: the naturalness of the interaction and the range of the interaction force. When the stiffness is low, pHRI is more natural and fluid. It contributes to accomplishing more sets of motions in a given time, which in turn contributes to improved prediction performance. As stiffness increases, the resistance to interaction becomes greater, and the range of interaction force expands. Increasing the range of interaction force can enhance prediction performance. When stiffness exceeds 200N/m, the range of interaction force exerts a more pronounced influence on estimated performance.

IV. Discussion

This paper formulates the question of human-robot interactions as a force estimation task. The interaction forces in the six directions, X, Y, Z, RX, RY, and RZ, are predicted simultaneously. Once the interaction forces are available, human-robot interactions can be obtained. Further, the torque of each upper limb joint can be calculated by inverse kinematics. Hence, interaction forces estimation can be a new

TABLE III
PREDICTION PERFORMANCE(R^2) UNDER DIFFERENT FEATURE ENGINEERING

Method	X	Y	Z	RX	RY	RZ	Mean
Raw sEMG	0.700±0.064	0.692±0.054	0.693±0.065	0.540±0.103	0.613±0.159	0.610±0.120	0.641±0.094
ACT	0.701±0.079	0.703±0.077	0.701±0.081	0.547±0.132	0.614±0.208	0.628±0.137	0.651±0.119
ZC	0.707±0.048	0.686±0.051	0.684±0.067	0.539±0.117	0.585±0.795	0.601±0.129	0.634±0.101
ACT+ZC	0.766±0.065	0.749±0.067	0.745±0.078	0.594±0.125	0.648±0.190	0.661±0.125	0.694±0.108
Our work	**0.978±0.004**	**0.969±0.007**	**0.978±0.005**	**0.913±0.010**	**0.971±0.008**	**0.940±0.021**	**0.958±0.009**

TABLE IV
COMPARISON WITH RECENT RELATED RESEARCH

Ref.	Task	Method	Estimation method	Performance(R^2)
[16]	Elbow torque(1-Dof)	Window averages for EMG	ResNet-BiLSTM	0.99
[19]	Thrust and torque(2-Dof)	Bayesian filtering for EMG	LSTM	0.924
[22]	Ankle torque(1-Dof)	EMG and US fusion with features stitching	SVR or FFNN	SVN:0.95, FFNN:0.93
[23]	Static grip force(6-Dof)	EMG and US fusion with self-attention	OLR-SACNN	0.989
[29]	Elbow torque(1-Dof)	EMG and joint angles fusion with features stitching	CNN	0.91
Our work	Arm interaction force(6-Dof)	EMG and Kinematics fusion with tensor and cross-attention	STNNs	CNNGRU:0.958, ConvLG:0.960 LS-TCN:0.951, CNN-att:0.952 TF:0.943, Swin-TF:0.855

method for analyzing human-robot interactions. The fusion information proposed is suitable for various STNNs for end-to-end estimation of interaction forces. It reveals that sEMG and kinematic information are highly relevant to human-robot interactions.

In the previous sEMG-force estimation, two methods that can improve the prediction performance from the feature engineering perspective are selecting better features and combining more features as feature vectors. These methods can obtain richer expressions within sEMG signals. However, manual feature selection requires a significant amount of time for testing, and the predictive effect of different features or combinations is difficult to explain from the feature perspective. Table 3 illustrates the prediction effect on interaction forces with raw sEMG, nonlinear muscle activation(ACT), zero crossings(ZC), combined ACT and ZC features, and our approach. The formers employ GRU as the regression model. Given that the fusion information is spatio-temporal sequences, CNNGRU, the closest to GRU, is selected for comparison. Extracting the ACT feature enhances prediction performance compared to the raw sEMG, while extracting the ZC feature has the opposite effect. Consequently, the extraction of less relevant features will result in a reduction in the model's prediction performance. The extraction of multiple features can enhance the estimated performance. For instance, the simultaneous extraction of ACT and ZC features improves the R^2 by 0.053 compared to no feature extraction. However, the superposition of features doubles the input dimensions of the model, which increases the computational cost. Our work does not require manual feature selection. A comparison of the prediction performance of raw sEMG with that of our proposed method revealed an improvement in the R^2 of 0.214-0.420. Additionally, our method exhibits a reduction in variance, thereby enhancing prediction stability. Our work explores a new approach to interaction force estimation under complex pHRI. This work improves the sEMG-force prediction performance primarily due to three aspects. Firstly, the kinematic information (pose and velocity) is highly correlated with human-robot interactions, thereby enriching the representation of the information and facilitating the complementarity of the data. Secondly, the proposed multimodal fusion scheme deeply fuses the information from the data layer and maximally retains the valuable information in the original data by spatio-temporal sequences. Finally, the proposed STNNs can effectively and comprehensively extract the human-robot interactions from the fused information and predict the force end-to-end.

Deep learning has great potential to extract task-specific features from human physiological and physical level information, such as sEMG, kinematics, and other information, to achieve state-of-the-art prediction performance. A summary of recent deep learning-based force estimation research is presented in Table 4. It shows that our proposed scheme achieves state-of-the-art performance, which validates the effectiveness of our approach. Furthermore, the works in TABLE IV show that physiological and physical-level information is highly correlated with interaction force and can be mined by neural networks. Studies [16] and [19] show that accurate recognition of low-dimensional motion has been achieved by deep learning, but complex motion recognition requires further exploration. For methods based on machine learning or deep learning, the critical issue is how to extract valuable information from the data. Although enhanced performance can be achieved by selecting better features or optimizing model parameters, sEMG-based devices still cannot be adopted in commercial systems. Results from studies [22], [23], [29], and our work show that multimodal techniques can simultaneously process and analyze multiple data types. Multimodal fusion enriches the expression of information expression and promotes data complementarity, helping methods based on machine learning or deep learning to achieve more comprehensive and intelligent decision-making and interaction. Employing multimodal information for pHRI reclaims a new path for a more applicable human-robot interface using sEMG signals. It can help bridge the gap between laboratory tests and real-life

applications.

This work shows that the proposed multimodal fusion scheme performs excellent six-dimensional force estimation under dynamic pHRI. However, limitations still exist. On the one hand, since it is a dynamic interactive task, the electrodes may be offset or have poor contact, leading to abnormal or missing data. These non-ideal factors will affect the accuracy and stability of force estimation. A better approach is to use data generation or interpolation algorithms optimized for non-ideal sEMG signals to improve signal quality under non-ideal conditions. On the other hand, although this work focuses on force estimation under multiple constant interaction stiffnesses, there may be situations of variable stiffness during interaction with the environment, such as the interaction between humans and soft robots. Combining interaction force estimation and stiffness estimation can be a future innovative path more applicable to human-robot interaction in multiple complex environments.

V. Conclusion

This paper presents a 6-dimensional interaction force estimation framework based on multimodal fusion and STNNs. The prediction performance of six STNNs and three baseline methods are compared, with the ConvLG-based method performing best. The proposed framework achieves state-of-the-art performance compared with other recent related research, as listed in Table 4. We quantify the synergistic activation of upper limb muscles from a biomechanical perspective and discuss the effects of different types of neural networks, kinematic spaces, and interaction stiffness on the estimation performance of interaction force. The results show that our method is effective and explores a new way for interaction force analysis under complex pHRI.

The proposed multimodal force estimation scheme can link human-robot collaborative interaction. It can be applied to high-dimensional and complex pHRI scenarios that require contact forces measurement, further promoting "human-oriented" intelligence and safe HRI. Combining sEMG to study the regulation mechanism of human muscle activation and force characteristics can promote the development of neurocognitive science and provide ideas for the control interface and hardware design of collaborative robots. It helps to bridge the gap between robots and humans in performing specific collaborative tasks.

Future work in this research can be applied to human-robot collaboration fields such as multi-dimensional power-assisted rehabilitation robots (Standard rehabilitation of high-dimensional movements using an estimated force-driven exoskeleton or end-effector rehabilitation robots), industrial assembly (Estimating high-dimensional interaction forces for adjusting robots to achieve collaborative handling by humans and robots under complex conditions), remote impedance (Teleoperation in hazardous environments, such as nuclear power plants, fires and rescues, and other human-robot-environment interaction tasks), and home services (Force estimation in physical contact enables home service robots to learn human skills more efficiently). In terms of human intention recognition and analysis, this work can also be tried for more dynamic recognition tasks such as gait tracking and sign language translation.

References

[1] L. Chen, L. Chen, X. Chen, H. Lu, Y. Zheng, J. Wu, Y. Wang, Z. Zhang, and R. Xiong, "Compliance while resisting: A shear-thickening fluid controller for physical human-robot interaction," *Int. J. Robot. Res.*, 2024.

[2] L. Chen, S. Yang, X. Zhang, T. Li, Y. Long, and H. Pan, "Development of a capacitive force/torque sensor for lower limb rehabilitation robots for spasm detection," *IEEE Trans. Instrum. Meas.*, 2023.

[3] B. Zheng, S. Verma, J. Zhou, I. W. Tsang, and F. Chen, "Imitation learning: Progress, taxonomies and challenges," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–16, 2022.

[4] Z. Chua, A. M. Jarc, and A. M. Okamura, "Toward force estimation in robot-assisted surgery using deep learning with vision and robot state," in *Proc. IEEE Int. Conf. Robot. Automat.* IEEE, 2021, pp. 12 335–12 341.

[5] C. Wang, M. Sivan, D. Wang, Z.-Q. Zhang, G.-Q. Li, T. Bao, and S. Q. Xie, "Quantitative elbow spasticity measurement based on muscle activation estimation using maximal voluntary contraction," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–11, 2022.

[6] Y. Zhao, J. Zhang, Z. Li, K. Qian, S. Q. Xie, Y. Lu, and Z.-Q. Zhang, "Computationally efficient personalized emg-driven musculoskeletal model of wrist joint," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–10, 2022.

[7] L. Liu, M. Illian, S. Leonhardt, and B. J. Misgeld, "Iterative learning control for cascaded impedance-controlled compliant exoskeleton with adaptive reaction to spasticity," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–11, 2023.

[8] D. Li, P. Kang, Y. Yu, and P. B. Shull, "Graph-driven simultaneous and proportional estimation of wrist angle and grasp force via high-density emg," *IEEE J. Biomed. Health Inform.*, pp. 1–10, 2024.

[9] T. Zhang, H. Chu, and Y. Zou, "An online human dynamic arm strength perception method based on surface electromyography signals for human-robot collaboration," *IEEE Trans. Instrum. Meas.*, 2023.

[10] C. Shen, Z. Pei, W. Chen, Z. Li, J. Wang, J. Zhang, and J. Chen, "Stmi: Stiffness estimation method based on semg-driven model for elbow joint," *IEEE Trans. Instrum. Meas.*, 2023.

[11] F. E. Zajac, "Muscle and tendon: properties, models, scaling, and application to biomechanics and motor control," *Crit. Rev. Biomed. Eng.*, vol. 17, no. 4, pp. 359–411, 1989.

[12] M. Hayashibe and D. Guiraud, "Voluntary emg-to-force estimation with a multi-scale physiological muscle model," *Biomed. Eng. Online*, vol. 12, pp. 1–18, 2013.

[13] F. Romero and F. Alonso, "A comparison among different hill-type contraction dynamics formulations for muscle force estimation," *Mech. Sci.*, vol. 7, no. 1, pp. 19–29, 2016.

[14] V. Khoshdel and A. Akbarzadeh, "An optimized artificial neural network for human-force estimation: consequences for rehabilitation robotics," *Ind. Robot.*, vol. 45, no. 3, pp. 416–423, 2018.

[15] J. Luo, C. Liu, and C. Yang, "Estimation of emg-based force using a neural-network-based approach," *IEEE Access*, vol. 7, pp. 64 856–64 865, 2019.

[16] W. Lu, L. Gao, H. Cao, and Z. Li, "semg-upper limb interaction force estimation framework based on residual network and bidirectional long short-term memory network," *Appl. Sci.-Basel*, vol. 12, no. 17, p. 8652, 2022.

[17] J. Xue and K. W. C. Lai, "Dynamic gripping force estimation and reconstruction in emg-based human-machine interaction," *Biomed. Signal Process. Control*, vol. 80, p. 104216, 2023.

[18] H. Su, W. Qi, Z. Li, Z. Chen, G. Ferrigno, and E. De Momi, "Deep neural network approach in emg-based force estimation for human–robot interaction," *IEEE Trans. Artif. Intell.*, vol. 2, no. 5, pp. 404–412, 2021.

[19] Q. Zhang, L. Fang, Q. Zhang, and C. Xiong, "Simultaneous estimation of joint angle and interaction force towards semg-driven human-robot interaction during constrained tasks," *Neurocomputing*, vol. 484, pp. 38–45, 2022.

[20] Y. Celik, S. Stuart, W. L. Woo, E. Sejdic, and A. Godfrey, "Multi-modal gait: A wearable, algorithm and data fusion approach for clinical and free-living assessment," *Inf. Fusion*, vol. 78, pp. 57–70, 2022.

[21] N. E. Krausz, D. Lamotte, I. Batzianoulis, L. J. Hargrove, S. Micera, and A. Billard, "Intent prediction based on biomechanical coordination of emg and vision-filtered gaze for end-point control of an arm prosthesis," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 6, pp. 1471–1480, 2020.

[22] Q. Zhang, W. H. Clark, J. R. Franz, and N. Sharma, "Personalized fusion of ultrasound and electromyography-derived neuromuscular features increases prediction accuracy of ankle moment during plantarflexion," *Biomed. Signal Process. Control*, vol. 71, p. 103100, 2022.

[23] Y. Zou, L. Cheng, and Z. Li, "A multimodal fusion model for estimating human hand force: Comparing surface electromyography and ultrasound signals," *IEEE Robot. Autom. Mag.*, vol. 29, no. 4, pp. 10–24, 2022.

[24] N. S. Jong, A. G. S. de Herrera, and P. Phukpattaranont, "Multimodal data fusion of electromyography and acoustic signals for thai syllable recognition," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 6, pp. 1997–2006, 2020.

[25] M. S. Al-Quraishi, I. Elamvazuthi, T. B. Tang, M. Al-Qurishi, S. Parasuraman, and A. Borboni, "Multimodal fusion approach based on eeg and emg signals for lower limb movement recognition," *IEEE Sens. J.*, vol. 21, no. 24, pp. 27 640–27 650, 2021.

[26] E. Nsugbe, C. Phillips, M. Fraser, and J. McIntosh, "Gesture recognition for transhumeral prosthesis control using emg and nir," *IET Cyber-Syst. Robot.*, vol. 2, no. 3, pp. 122–131, 2020.

[27] P. Xiong, C. Wu, H. Zhou, A. Song, L. Hu, and X. P. Liu, "Design of an accurate end-of-arm force display system based on wearable arm gesture sensors and emg sensors," *Inf. Fusion*, vol. 39, pp. 178–185, 2018.

[28] N. Li, Y. Yang, G. Li, T. Yang, Y. Wang, W. Chen, P. Yu, X. Xue, C. Zhang, W. Wang *et al.*, "Multi-sensor fusion-based mirror adaptive assist-as-needed control strategy of a soft exoskeleton for upper limb rehabilitation," *IEEE Trans. Autom. Sci. Eng.*, 2022.

[29] G. Hajian, E. Morin, and A. Etemad, "Multimodal estimation of endpoint force during quasi-dynamic and dynamic muscle contractions using deep learning," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–11, 2022.

[30] Y. Sheng, J. Zeng, J. Liu, and H. Liu, "Metric-based muscle synergy consistency for upper limb motor functions," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–11, 2021.

[31] W. Zhong, X. Fu, and M. Zhang, "A muscle synergy-driven anfis approach to predict continuous knee joint movement," *IEEE Trans. Fuzzy Syst.*, vol. 30, no. 6, pp. 1553–1563, 2022.

[32] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," https://doi.org/10.48550/arXiv.1806.00064, 2018.

[33] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017.

[34] Z. Tang, H. Yu, H. Yang, L. Zhang, and L. Zhang, "Effect of velocity and acceleration in joint angle estimation for an emg-based upper-limb exoskeleton control," *Comput. Biol. Med.*, vol. 141, p. 105156, 2022.

[35] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Proc. Int. Conf. Neural Inf. Process. Syst.*, vol. 28, 2015.

[36] C. Chen, W. Guo, C. Ma, Y. Yang, Z. Wang, and C. Lin, "semg-based continuous estimation of finger kinematics via large-scale temporal convolutional network," *Appl. Sci.-Basel*, vol. 11, no. 10, p. 4678, 2021.

[37] Y. Geng, Z. Yu, Y. Long, L. Qin, Z. Chen, Y. Li, X. Guo, and G. Li, "A cnn-attention network for continuous estimation of finger kinematics from surface electromyography," *IEEE Robot. Autom. Lett.*, vol. 7, no. 3, pp. 6297–6304, 2022.

[38] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 1–9.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Proc. Int. Conf. Neural Inf. Process. Syst.*, vol. 30, 2017.

[40] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Conf. Comput. Vis.*, 2021, pp. 10 012–10 022.

[41] S. Ma, J. Zhang, C. Shi, P. Di, I. D. Robertson, and Z.-Q. Zhang, "Physics-informed deep learning for muscle force prediction with unlabeled semg signals," *IEEE Trans. Neural Syst. Rehabil. Eng.*, 2024.

[42] Q. Yuan and I.-M. Chen, "Localization and velocity tracking of human via 3 imu sensors," *Sens. Actuator A-Phys.*, vol. 212, pp. 25–33, 2014.

Gao Lin received the B.S. degree in automation from the Guangxi University, Nanning, China, in 2020. The M.S. degree in Electronic Information form the State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, China, in 2023. Now, he is pursuing a Ph.D. in Robotics Science and Engineering form the Northeastern University. He is committed to the research of human-machine interaction technology of end effector upper limb rehabilitation robot.

Fei Wang received the B.S., M.S. and Ph.D. degrees from Harbin Institute of Technology, Harbin, China, in 2000. He also received the Ph.D. degree in Electrical Engineering from the University of Tokushima, Japan, in 2004. He has long been committed to researching and solving key scientific theoretical and technical problems in the field of intelligent robots such as mechatronics, multi-modal perception and cognition, and human-machine interaction.

Xu Zhong received the B.S. degree in Xuzhou Medical University, Xuzhou, China,in 2020. The MS.degree form the State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, and the Department of Biomedical Engineering, China Medical University, Shenyang, China, in 2023. Now, he is working in the Affiliated Hospital of Yangzhou University. He is committed to the research of biomedical signal processing and clinical medical engineering.

Zida An received the B.S. degree in automation from the Northeastern University, Shenyang, China, in 2018. The M.S. degree in control science and engineering form the Northeastern University, in 2021. Now, he is pursuing a Ph.D. in Robotics Science and Engineering form the Northeastern University.He is committed to the research of human-machine interaction technology of rehabilitation robot.

Shuai Han received M.D. from China Medical University and is currently studying for his Ph.D. in 2018. He was the first doctor to use a Chinese-made robot to perform neurosurgery in northeast China. Nearly 1,000 neurosurgical robotic operations have been completed. He pioneered the use of dental registration to improve robotic surgery accuracy and reduce trauma.