Towards Neurorobotic Interface for Finger Joint Angle Estimation: A Multi-Stage CNN-LSTM Network with Transfer Learning

Yun Chen, Xinyu Zhang, Hui Li, Hongsheng He, Wan Shou, and Qiang Zhang

***Abstract*— To maximize the autonomy of individuals with upper limb amputations in daily activities, leveraging forearm muscle information to infer movement intent is a promising research direction. While current prosthetic hand technologies can utilize forearm muscle data to achieve basic movements such as grasping, accurately estimating finger joint angles remains a significant challenge. Therefore, we propose a Multi-Stage Cascade Convolutional Neural Network with Long Short-Term Memory Network, where an upsampling module is introduced before the downsampling module to enhance model generalization. Additionally, we designed a transfer learning (TL) framework based on parameter freezing, where the pre-trained downsampling module is fixed, and only the upsampling module is updated with a small amount of out-of-distribution data to achieve TL. Furthermore, we compared the performance of unimodal and multimodal models, collecting surface electromyography (sEMG) signals, brightness mode ultrasound images (B-mode US images), and motion capture data simultaneously. The results show that on the validation set, the US image had the lowest error, while on the prediction set, the four-channel sEMG achieved the lowest error. The performance of the multimodal model in both datasets was intermediate between the unimodal models. On the prediction set, the average normalized root mean square error values for the four-channel sEMG, US images, and sensor fusion models across three subjects were 0.170, 0.203, and 0.186, respectively. By utilizing advanced sensor fusion techniques and TL, our approach can reduce the need for extensive data collection and training for new users, making prosthetic control more accessible and adaptable to individual needs.**

I. Introduction

Driven by neurorobotics, prosthetic hands enhance autonomy for individuals with physical disabilities [1]. For those who retain considerable functionality in their forearms, modern technology reconstructs movement intention and controls the prosthetic hand via muscle responses to brain signals.

Surface electromyography (sEMG) signals are extensively utilized for this purpose, as they can record the electrical activities initiated by muscle contractions [2], [3]. To obtain data from various muscle locations, multiple channels of signals are collected. A common method is to use a high-density EMG array [4] or a combination of multiple sEMG [5] sensors placed at different locations. By applying machine learning algorithms to analyze these multi-channel signals, it is possible to decode the intended movements of the human body based on the sEMG data [6]–[8]. Nevertheless, sEMG signals typically struggle with noise and artifacts. At the same time, crosstalk between adjacent muscles makes it difficult to isolate the activity of specific muscles, especially in the study of forearm muscles [9], [10]. In addition, the sliding of electrodes during movement will also affect the consistency of the signal.

In recent years, a new method for acquiring muscle activities using ultrasound (US) has emerged [11]–[13]. These sensors emit high-frequency sound waves that produce echoes when they encounter tissues of varying densities, thereby capturing signals. Brightness mode (B-mode) US imaging provides detailed two-dimensional information, showing cross-sectional views with varying brightness to indicate different echo intensities. The advantage of this technology lies in its ability to observe the structural changes of muscles and other soft tissues in real-time and non-invasively [14]. This makes it particularly valuable for studying the relationship between limb motion and muscle activity.

Up to now, many studies have achieved gesture and finger movement prediction based on sEMG signals or US signals [15]–[23]. With the development of DL methods [24]–[26], these predictions have become more accurate and reliable. Additionally, the benefits of using sensor fusion between different signals for motion/motion intent prediction on upper [27]–[29] and lower limbs [30]–[33] have been demonstrated. However, the sensor fusion approaches for upper limb applications mainly focus on discrete hand gesture recognition [34], thus missing the continuous hand motion prediction with sensor fusion. In the current study, we conducted the pilot study of the application of deep learning (DL) models to sensor fusion datasets and explored the use of transfer learning (TL) to address variations in muscle patterns across individuals when applying the model to different users, the detailed pipeline is shown in Fig. 1. The contributions of this study are as follows: (1) We propose a Multi-Stage Cascade Convolutional Neural Network with a Long Short-Term Memory (CNN-LSTM) network that leverages the strong temporal dependence of muscle information during finger movements; (2) To support cross-user model application, we introduce a TL framework using parameter freezing; (3) We developed a method to synchronize and process sEMG, US images, and motion capture data. Additionally, we investigated the performance of different sensor data and fusion datasets in predicting metacarpophalangeal (MCP) joint angles.

*This work was supported by the ROH funding at the University of Alabama with the FOAP of 13009-214271-200. Corresponding author: Qiang Zhang (qiang.zhang@ua.edu).

Yun Chen and Qiang Zhang are with the Department of Mechanical Engineering, the University of Alabama, Tuscaloosa, 35487, USA.

Xinyu Zhang, Hui Li and Hongsheng He are with the Department of Computer Science, the University of Alabama, Tuscaloosa, 35487, USA.

Wan Shou is with the Department of Mechanical Engineering, University of Arkansas, Fayetteville, 72701, USA

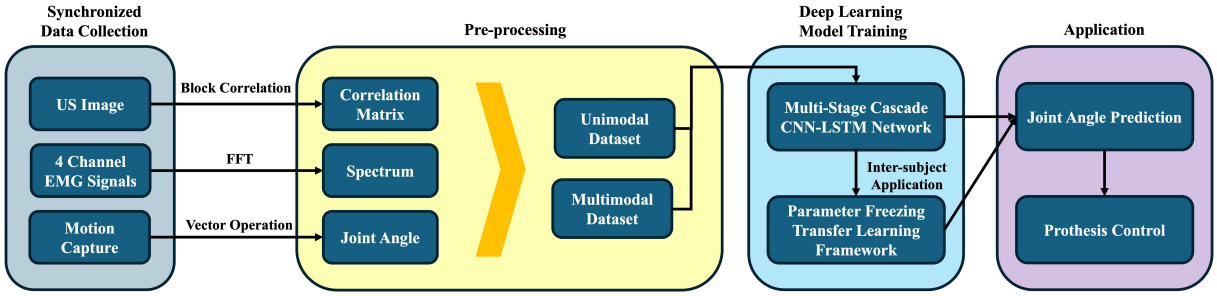

Fig. 1: Detailed workflow of data collection and processing for deep learning (DL) model and prosthetic hand control.

II. Related Work

Numerous studies have focused on classifying and predicting hand motion using sEMG and US signals from forearm muscles [35], [36]. Some research targets gesture classification. For example, Qi et al. used principal component analysis and generalized regression neural networks to develop a recognition system, collecting sEMG signals for nine static gestures and achieving a 95.1% recognition rate [37]. Similarly, Lu et al. extracted features from A-mode US signals and employed SVM for real-time classification, reaching an accuracy of $83.8\% \pm 6.9\%$ [38].

Other studies estimate finger movements from muscle signals [39]–[41]. Lee et al. introduced an encoder-decoder network with attention mechanisms to estimate 14 finger joint angles, achieving an average error of $10.86\% \pm 9.82\%$ for MCP joints [42]. Another study used a CNN-RNN structure to calculate finger movement angles from US images, with errors under 0.1 rad [43]. Additionally, TL has been employed to reduce data collection and training burdens [44]. One study [6] aggregated signals from multiple users to train a DL model, followed by TL for individual data. This approach maintained high accuracy while minimizing individual data needs. Testing on various datasets, this study used raw sEMG, spectrograms, and continuous wavelet transform (CWT) as inputs. The CWT-based CNN achieved a 98.31% offline accuracy on gesture classification.

Furthermore, sensor fusion methods have been explored in rehabilitation research [12]. A study applied a multimodal multilevel converged attention network (MMCANet), which combined sEMG and A-mode US signals for gesture recognition. Using CNN-LSTM for feature extraction and a transformer encoder for fusion, MMCANet outperformed linear discriminant analysis by 5.15%. Single-modal gains were 14.31% for sEMG and 3.80% for AUS [34].

III. Methodology

A. Experimental Implementation

In this study, four sEMG sensors (Trigno System, Delsys, USA) and a B-mode US diagnostic system with a US transducer (ArtUs EXT-2H, Telemed, Lithuania) were simultaneously placed on the subject's forearm, with markers positioned on the back of the hand for motion capture. The sEMG signals and motion capture were directly synchronized using the motion capture system software Nexus 12.0 (Vicon, Oxford, UK). Synchronization between the US image signals and Nexus was achieved through a connection to Simulink of MATLAB R2020b with Quanser being connected. A trigger in Simulink simultaneously initiated signal acquisition in Nexus and screen recording software for capturing US image signals. For convenient data processing, the sampling rate of the sEMG signals was set at 960 Hz, while the US image and motion capture sampling rates were 60 Hz. All study procedures were approved by the institutional review board at the University of Alabama (Protocol ID: 24-01-7231).

The arrangement of the sensors is illustrated in Fig. 2a and 2b, where sEMG sensors were placed on specific locations of the forearm to capture electrical signals from different muscles corresponding to finger movements, as detailed in Table I. The US transducer was positioned on the inner side of the forearm at a 30-degree angle to the radius. This configuration allows for a comprehensive observation of muscle deformation in the cross-section of the forearm, as well as muscle stretching and contraction along the radial direction. For this study, markers are placed on the proximal phalanges and metacarpal bones of the fingers, excluding the thumb, as shown in the diagram. The movement of the thumb is not considered in this analysis because its primary movements are controlled by the carpometacarpal and MCP joints of the thumb [45], which are affected by intrinsic muscles outside the range of our imaging technology.

In each experimental set, subjects were required to stand and position their forearms at a 45-degree angle on the tabletop. To capture the continuous angle changes of each targeted independent finger, the starting position was designated as fully extended (with the MCP joint at 0 degrees), and the maximum flexion served as the end position. The process where the finger moves from the starting position to maximum flexion and then returns to the starting position was defined as one cycle. Subjects were instructed to perform finger movements at a rate of one cycle every two seconds, following a metronome set at 60 beats per minute. Each trial involved 13 cycles of movement for each targeted finger. When pre-processing and forming a data set, the first three seconds and the last three seconds of each finger's motion data were eliminated, and the middle 20 seconds were taken as valid data. Data were collected from three subjects, each undergoing three experimental sets. Two of these sets were used as the training and validation datasets, while the third

set was excluded from the training process and utilized as a prediction set to evaluate the model's generalizability.

(a) (b) (c)

Fig. 2: (a) Sensor placement of inner and outer arm. (b) Experiment setup of synchronized data acquisition. (c) Side view of hand. Gray dots are markers for motion capture.

TABLE I: Target muscles and corresponding movements for each channel sEMG

Channel	Target Muscle	Target Movement
1	Flexor digitorum superficialis	Middle finger flexion
2	Left extensor digitorum	Index, Middle, Ring, Little finger extension
3	Flexor Digiti Minimi Brevis	Little finger flexion
4	Flexor digitorum profundus	Ring, Little finger flexion

B. Data Pre-Processing And DL Dataset Formation

1) US Signals: B-mode US images are large, consuming significant computational resources and hindering real-time monitoring efficiency. Therefore, for the collected US images, we filtered out points with grayscale values below 10 using threshold processing. We then applied a block correlation method. Specifically, the image was evenly divided into multiple blocks, and the correlation between consecutive frames was calculated for each block, forming a correlation matrix, which is shown in Fig. 3d. This method captures motion changes and dynamic features in muscle images during finger movements. To enable the LSTM layer to extract temporal features, every four consecutive correlation matrices were concatenated along the time dimension to form the dataset. In this study, we cropped the US images to a size of 560*560, then set each block to 16*16 pixels, and used the block correlation method to reduce the image data to a 35*35 correlation matrix. Fig. 3e shows a 5 * 5 block from the correlation matrix.

2) sEMG Signals: sEMG signals typically contain complex noise that requires elimination through a series of preprocessing steps. First, a fourth-order Butterworth filter is applied for band-pass filtering in the range of 20-450 Hz to remove low-frequency noise and high-frequency interference. The signal is then rectified and demeaned to eliminate DC offset. Then, a fourth-order Butterworth filter with a 6 Hz low-pass frequency is used to smooth the signal and extract the envelope of sEMG signals. Finally, each column of sEMG data is normalized separately, raw sEMG signals and processed signals are shown in Fig. 3a and Fig. 3b.

To align with the US image dataset on the time axis for sensor fusion network training, a sliding window method is applied to the sEMG signals, with a window size of 96 and a stride of 960/60=16. Each windowed signal is transformed into a spectrum using a Fast Fourier Transform. Every four consecutive spectra of sliding windows are concatenated along the time dimension to form the dataset. Transforming signals into spectra reveals frequency components, aiding in the identification and analysis of features across different frequencies. The synchronized formation of US image, sEMG signals, and MCP joint angle is shown in Fig. 3f.

3) Hand Motion Capture: The Vicon system enables the real-time tracking of markers, which allows for precise measurement of the spatial locations of markers attached to the fingers. This capability is crucial for accurately calculating the angles of the MCP joints.

Taking the index finger as an example, a marker is positioned near the MCP joint on the back of the hand. A side view of the hand is shown in Fig. 2c, in this setup, $\vec{MP}1$ refers to the proximal phalanx of the index finger, and $\vec{MQ}$ denotes the metacarpal bone of the index finger. The angle of the MCP joint is calculated by determining the angle between the vectors representing these anatomical structures. The cross-product is employed to calculate the angle accurately, ensuring that our estimations and subsequent normalization of finger angles are precise. Finally, the computed MCP joint angle data, used as ground truth, was smoothed using a 6 Hz mean filter, as shown in Fig. 3c. Additionally, the MCP angles for each finger were normalized individually.

C. Network Design

Although CNNs have shown commendable recognition performance in US image recognition and sEMG signal processing [46], [47], they often struggle in complex scenarios due to insufficient local information, artifacts, and feature loss during forward propagation. To address this, we propose a CNN that uses upsampling to expand and decode information. Additionally, we employ US-sEMG multimodal feature fusion to compensate for potential feature loss in single modalities. For inter-subject scenarios, parameter freezing is applied to the pre-trained model on in-distribution (ID) datasets. By freezing the optimal parameters of the downsampling encoding layers, these parameters serve as prior information and non-empirical hyperparameters during training on out-of-distribution (OOD) datasets.

1) Multi-Stage Cascade CNN-LSTM Network: The general network for feature extraction from sEMG and US signals is divided into three stages, as shown in Fig. 4. The first stage involves a feature expansion operation composed of three upsampling layers. In the upsampling layer, transposed convolution is used to increase the spatial resolution of the feature maps. Transposed convolution works by reversing the process of a regular convolution, effectively enlarging the input size. The use of transposed convolution serves two purposes. Firstly, it partially restores the precision of the original US signal through transpose convolution, allowing the feature maps to be resized to a more reasonable scale and

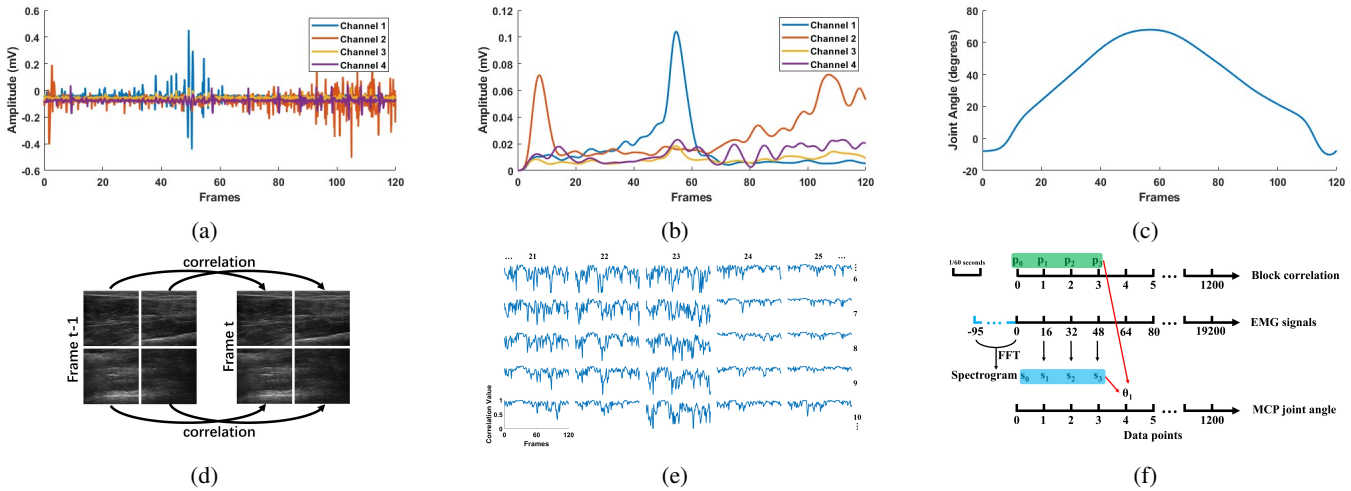

Fig. 3: (a) sEMG signals before filters. (b) sEMG signals after filters. (c) Joint angle for middle finger of one movement cycle. (d) Correlation calculation. (e) 120 frames Correlation value of a 5*5 blocks from 35*35 correlation matrix. (f) The synchronization of different sensors with joint angles.

compensating for the loss of precision and features during segmentation. Secondly, it maps the sEMG signals from the electrode collection area to the surrounding skin region, increasing the features while reducing the impact of electrode movement during experiments, thereby enhancing the correlation between sEMG signals and motion recognition and smoothing the subsequent feature compression process. The second stage comprises five down-sampling layers, this stage compresses and encodes the feature maps transmitted from the previous stage. The final stage consists of a flattened layer, LSTM layer, and Multilayer Perceptron (MLP) layer.

2) Sensor fusion model of US and sEMG: US signals capture the morphological characteristics of muscles, while sEMG signals focus on the electrical aspects of muscle contraction. For sensor fusion by feature, we concatenate features from both signals within the original model to create a fused feature space that, along with temporal variations, is fed into the final MLP layer for regression. Notably, we balance the total number of flattened features to prevent uneven composition. The detailed parameters of the fusion network are shown in Table II, the numbers in the table represent: Channels – Kernel Size – Stride.

TABLE II: Condensed Layer Details for Ultrasound and sEMG Channels

Layer Name	Ultrasound	sEMG
Input	4×1×35×35	4×1×96×4
ConvTranspose2d-1	8 — 2×2 — 2×2	8 — 2×2 — 2
ConvTranspose2d-2	3 — 2×2 — 2×2	3 — 2×2 — 2
ConvTranspose2d-3	1 — 2×2 — 2×2	1 — 2×2 — 2
Conv2d-1	3 — 2×2 — 1×1	3 — 2×2 — 1×1
Conv2d-2	8 — 2×2 — 1×1	8 — 2×2 — 1×1
Conv2d-3	16 — 2×2 — 1×1	16 — 2×2 — 1×1
Conv2d-4	32 — 2×2 — 1×1	32 — 2×2 — 1×1
Conv2d-5	64 — 2×2 — 1×1	64 — 2×2 — 1×1
Conv2d-6	64 — 1×1 — 1×1	-
Flatten	-	-
LSTM / FCN	hidden size = 100	neurons = 100

3) Parameter freezing transfer learning framework: To further enhance the model's generalization ability and robustness, mitigating the effects caused by individual differences and electrode repositioning during testing, we propose a framework based on frozen parameters, as shown in Fig. 5. The training parameters of the downsampling encoder from the pre-trained model on the ID dataset are frozen (specifically, the parameters from stage 2) and used as hyperparameters during OOD dataset training. In this process, the parameters of the downsampling layers remain fixed, while the upsampling layers reconstruct the expanded feature maps, which are then passed through the frozen downsampling layers for adaptation to the OOD data features. By freezing certain parameters, the model updates only a subset of its parameters, reducing training time and preserving its feature extraction capabilities. This approach helps prevent overfitting and minimizes instability when there are significant variations in sample features.

D. DL Model Performance Evaluation

In the evaluation of model performance, RMSE and correlation coefficient are employed to assess the model's predictive accuracy on the dataset. These metrics are instrumental in understanding the model's accuracy in predicting MCP joint angles over continuous periods. The ultimate aim of this study is to control prosthetic hand movements in real-time based on forearm muscle information. Given that DL models typically address regression problems, we applied a low-pass filter at 2 Hz to the model outputs to achieve smoother results. The RMSE and correlation are calculated based on these smoothed outputs.

In the entire dataset, the movement of each finger is conducted sequentially. Therefore, when calculating the RMSE and correlation for each finger, we extract the data segment corresponding to that finger's movement. In addition, to statistically verify the improvement of performance brought

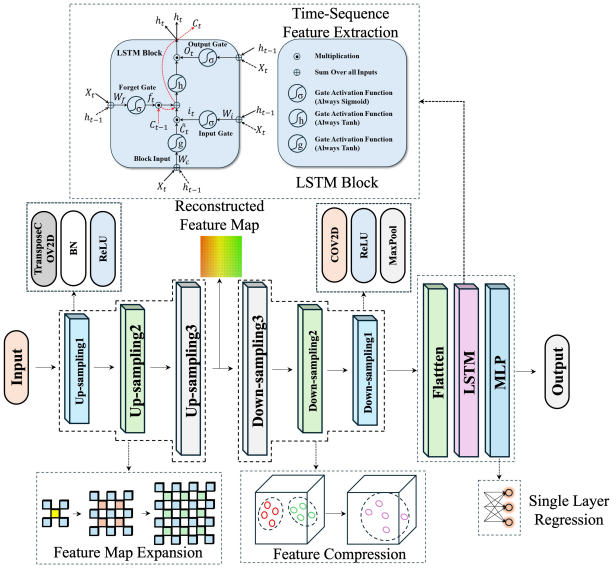

Fig. 4: The general DL model for US and sEMG data.

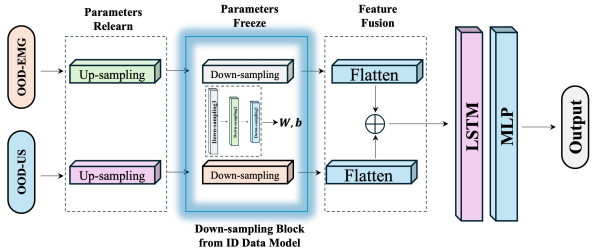

Fig. 5: The sensor fusion and TL framework.

by our proposed model, the Shapiro-Wilk normality test was applied to determine the data distribution of each metric across subjects. Then, on the two sets of data of three subjects which are required to be compared, and then conducted a repeated measures ANOVA test.

IV. Results And Discussion

A. Ablation Experiment

In our study on sensor fusion datasets, we trained both our proposed model and a variant without the upsampling module, retaining only the downsampling component. We compared their inference performance to the prediction set. Fig. 6 presents the results for all three subjects. All indicate that our proposed model has a lower error, with an average improvement of 0.014 ($p<0.05$), and the correlation coefficient increases by an average of 0.049 ($p<0.05$).

B. Inter-Subject Transfer Learning

The sensor fusion training set for each subject was used to train a base model individually, and then we utilized 20% of the training set from another subject for TL, conducting model training within a parameter-freezing framework. Fig. 7 shows an example of the performance of the TL models on the validation set. Each subject's training set was used to train a model as a pre-training step. We then employed

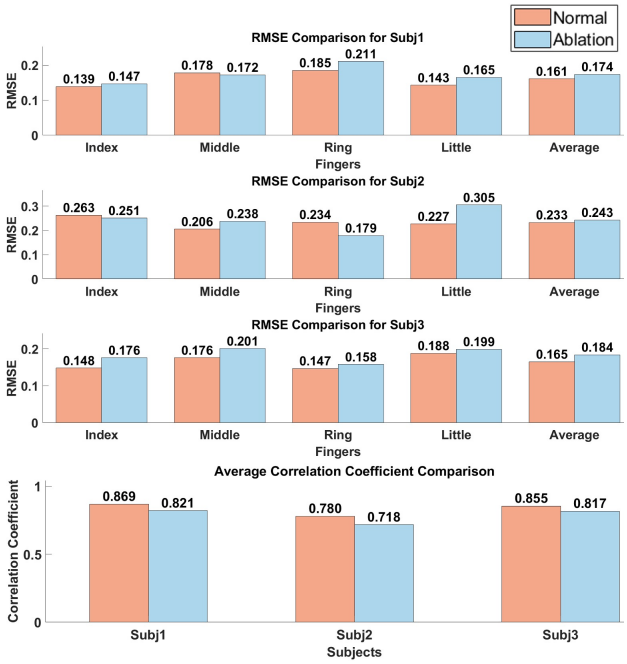

Fig. 6: Prediction RMSE and correlation coefficient of ablation experiments.

TL by using 20% of another subject's training set to further train the model, while keeping some parameters frozen. The performance of the TL models and the train-from-start models on the validation set is shown in Table III. Applying a pre-trained model from one subject directly to another subject's data, without TL, results in an average correlation coefficient of 0.140 between predicted and actual values, indicating very low correlation and almost no effectiveness. As observed in Table III, the models obtained through TL perform as well as the train-from-start models. Specifically, the model transferred to Subject 1 shows a global RMSE that is on average 0.034 higher than the train-from-start model. For models transferred to Subjects 2 and 3, the performance is slightly superior to the train-from-start models.

C. Unimodal And Multimodal Model Performance

We trained models using four-channel sEMG, US images, and a fusion dataset across three subjects to compare the impact of different sensor data and sensor fusion on model prediction performance. Additionally, we experimented with three-channel and two-channel sEMG data, selecting the optimal results from each combination for analysis. The inference results for all models are presented in Table IV.

On the validation set, the US image consistently demonstrated superior performance. On average, across the three subjects, the US image and sensor fusion improved the RMSE by 0.018 and 0.015 ($p<0.05$), respectively, compared to sEMG. However, on the prediction set, the four-channel sEMG outperformed the others, followed by sensor fusion, while the US image performed worse. Specifically, the US image and sensor fusion resulted in a decrease in average RMSE of 0.033 and 0.017 ($p<0.05$), respectively, compared

TABLE III: Validation RMSE of TL and train-from-start model

Validation	subj2to1	subj3to1	subj1	Subj1Avg.Improv.	subj1to2	subj3to2	subj2	Subj2Avg.Improv.	subj1to3	subj2to3	subj3	Subj3Avg.Improv.
Index	0.148	0.121	0.161	0.026	0.132	0.148	0.092	-0.048	0.129	0.141	0.122	-0.013
Middle	0.146	0.134	0.084	-0.056	0.114	0.110	0.191	0.079	0.099	0.099	0.159	0.060
Ring	0.137	0.111	0.093	-0.027	0.120	0.119	0.153	0.034	0.095	0.096	0.108	0.013
Little	0.234	0.237	0.150	-0.086	0.209	0.217	0.182	-0.031	0.141	0.148	0.098	-0.047
Average	0.166	0.151	0.122	-0.037	0.144	0.149	0.155	0.009	0.116	0.121	0.122	0.004

TABLE IV: Prediction and Validation Performance for Subj1 & Subj2 & Subj3

Subj1 Val/Pred	4channels	3channels	2channels	US	fusion
Index RMSE	0.135/0.174	0.150/0.178	0.228/0.234	0.162/0.139	0.139/0.160
Middle RMSE	0.148/0.094	0.147/0.091	0.165/0.134	0.213/0.093	0.178/0.087
Ring RMSE	0.162/0.078	0.159/0.103	0.212/0.135	0.210/0.096	0.185/0.096
Little RMSE	0.140/0.142	0.156/0.139	0.171/0.166	0.154/0.153	0.143/0.140
Average RMSE	0.146/0.122	0.154/0.128	0.194/0.167	0.185/0.120	0.161/0.121
Average Correlation	0.905/0.907	0.891/0.891	0.816/0.786	0.865/0.902	0.869/0.910

Subj2 Val/Pred	4channels	3channels	2channels	US	fusion
Index RMSE	0.256/0.212	0.237/0.185	0.244/0.210	0.267/0.087	0.263/0.092
Middle RMSE	0.160/0.155	0.176/0.146	0.258/0.251	0.196/0.179	0.206/0.191
Ring RMSE	0.128/0.133	0.102/0.124	0.107/0.137	0.221/0.145	0.234/0.153
Little RMSE	0.243/0.190	0.231/0.184	0.245/0.214	0.226/0.171	0.227/0.182
Average RMSE	0.196/0.173	0.187/0.160	0.213/0.203	0.228/0.145	0.232/0.155
Average Correlation	0.812/0.820	0.839/0.849	0.763/0.761	0.746/0.863	0.781/0.862

Subj3 Val/Pred	4channels	3channels	2channels	US	fusion
Index RMSE	0.184/0.174	0.229/0.232	0.248/0.262	0.163/0.124	0.148/0.122
Middle RMSE	0.146/0.127	0.138/0.133	0.143/0.155	0.265/0.158	0.176/0.159
Ring RMSE	0.139/0.119	0.203/0.132	0.181/0.201	0.191/0.102	0.147/0.108
Little RMSE	0.198/0.170	0.212/0.194	0.259/0.243	0.162/0.102	0.188/0.098
Average RMSE	0.166/0.147	0.195/0.173	0.210/0.215	0.195/0.121	0.165/0.122
Average Correlation	0.851/0.876	0.830/0.871	0.835/0.830	0.841/0.884	0.855/0.890

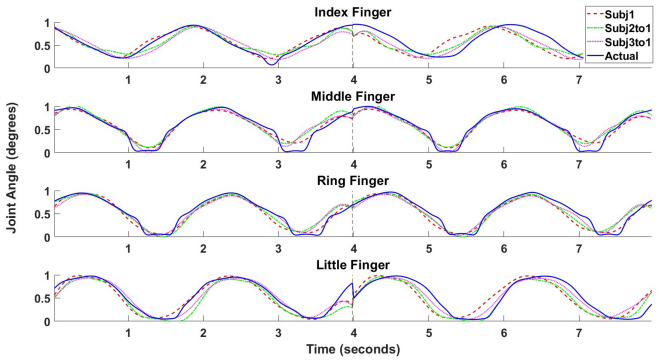

Fig. 7: Prediction performance of TL framework, the joint angle had been normalized to 0 to 1.

to sEMG. Furthermore, in terms of sensor performance, US image outperformed two-channel sEMG in prediction set but was inferior to the three- and four-channel sEMG signals.

D. Discussion

The TL framework based on parameter freezing is effective for inter-subject application scenarios. Directly applying a model trained on one subject to another results in a low correlation between predicted and actual values, highlighting significant differences in muscle signal distributions. By retraining the upsampling module with a small amount of new user data, the transpose convolution layers effectively map new data to suitable feature maps while retaining the original downsampling module's feature extraction. This TL framework, with parameter freezing, facilitates online fine-tuning for prosthetic hand control, reducing the time and cost of training from scratch.

In experiments comparing multimodal and unimodal models, sensor fusion did not achieve the anticipated accuracy gains. Due to its rich spatial features, the US image performed best on the validation set but worst on the prediction set, likely due to sensor drift during data collection. EMG-based models showed decreasing error with more channels, as each channel corresponds to different muscles, enhancing spatial complexity. Sensor fusion was performed between US images and EMG on both sets, indicating successful feature integration, though without a clear performance boost. This might be because the four-channel EMG already offers sufficient spatial complexity, making the additional US image features slightly outweighed by noise and computational complexity.

Despite achieving good prediction performance for MCP joint angles, this work has certain limitations. The preprocessing of US signals uses frame-to-frame block correlation to reduce the dimensionality of raw time-series images and extract features. However, this method neglects the directional information of muscle movements, reducing temporal changes between frames to a single value and may be influenced by background noise.

V. Conclusions And Future Work

In conclusion, our work successfully estimates finger MCP joint angles using forearm muscle information by incorporating an upsampling module into the traditional CNN-LSTM framework. Additionally, we only need to train a single model to predict all finger movements, enabling the possibility of real-time prosthetic hand control in the future. We also conducted experiments combining DL and sensor fusion methods, drawing some preliminary conclusions.

In future works, we plan to incorporate more complex finger movements, such as simultaneous movements of multiple fingers, where sensor fusion may demonstrate its advantage in feature integration. Furthermore, the preprocessing of ultrasound images is a potential area for improvement. We plan to explore other methods suitable for capturing time-series image signals, such as optical flow, to obtain potentially more effective inputs for DL models.

References

[1] V. Mendez, F. Iberite, S. Shokur, and S. Micera, "Current solutions and future trends for robotic prosthetic hands," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 4, no. 1, pp. 595–627, 2021.

[2] R. H. Chowdhury, M. B. Reaz, M. A. B. M. Ali, A. A. Bakar, K. Chellappan, and T. G. Chang, "Surface electromyography signal processing and classification techniques," *Sensors*, vol. 13, no. 9, pp. 12431–12466, 2013.

[3] R. Merletti and G. Cerone, "Tutorial. surface emg detection, conditioning and pre-processing: Best practices," *Journal of Electromyography and Kinesiology*, vol. 54, p. 102440, 2020.

[4] J. E. Lara, L. K. Cheng, O. Röhrle, and N. Paskaranandavadivel, "Muscle-specific high-density electromyography arrays for hand gesture classification," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 5, pp. 1758–1766, 2021.

[5] J. J. A. M. Junior, M. L. Freitas, H. V. Siqueira, A. E. Lazzaretti, S. F. Pichorim, and S. L. Stevan Jr, "Feature selection and dimensionality reduction: An extensive comparison in hand gesture classification by semg in eight channels armband approach," *Biomedical Signal Processing and Control*, vol. 59, p. 101920, 2020.

[6] M. A. Ozdemir, D. H. Kisa, O. Guren, and A. Akan, "Hand gesture classification using time–frequency images and transfer learning based on cnn," *Biomedical Signal Processing and Control*, vol. 77, p. 103787, 2022.

[7] R. Crepin, C. L. Fall, Q. Mascret, C. Gosselin, A. Campeau-Lecours, and B. Gosselin, "Real-time hand motion recognition using semg patterns classification," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2655–2658, IEEE, 2018.

[8] L. Zhang, G. Liu, B. Han, Z. Wang, and T. Zhang, "semg based human motion intention recognition," *Journal of Robotics*, vol. 2019, no. 1, p. 3679174, 2019.

[9] J. Shi, Y.-P. Zheng, X. Chen, and Q.-H. Huang, "Assessment of muscle fatigue using sonomyography: Muscle thickness change detected from ultrasound images," *Medical engineering & physics*, vol. 29, no. 4, pp. 472–479, 2007.

[10] L. Mesin, "Crosstalk in surface electromyogram: literature review and some insights," *Physical and Engineering Sciences in Medicine*, vol. 43, pp. 481–492, 2020.

[11] S. Sikdar, H. Rangwala, E. B. Eastlake, I. A. Hunt, A. J. Nelson, J. Devanathan, A. Shin, and J. J. Pancrazio, "Novel method for predicting dexterous individual finger movements by imaging muscle activity using a wearable ultrasonic system," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 1, pp. 69–76, 2013.

[12] N. Li, Y. Yang, G. Li, T. Yang, Y. Wang, W. Chen, P. Yu, X. Xue, C. Zhang, W. Wang, N. Xi, and L. Liu, "Multi-sensor fusion-based mirror adaptive assist-as-needed control strategy of a soft exoskeleton for upper limb rehabilitation," *IEEE Transactions on Automation Science and Engineering*, vol. 21, no. 1, pp. 475–487, 2024.

[13] X. Yang, X. Sun, D. Zhou, Y. Li, and H. Liu, "Towards wearable a-mode ultrasound sensing for real-time finger motion recognition," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 6, pp. 1199–1208, 2018.

[14] C. M. Moran and A. J. Thomson, "Preclinical ultrasound imaging—a review of techniques and imaging applications," *Frontiers in Physics*, vol. 8, p. 124, 2020.

[15] M. V. Arteaga, J. C. Castiblanco, I. F. Mondragon, J. D. Colorado, and C. Alvarado-Rojas, "Emg-driven hand model based on the classification of individual finger movements," *Biomedical Signal Processing and Control*, vol. 58, p. 101834, 2020.

[16] N. Akhlaghi, C. A. Baker, M. Lahlou, H. Zafar, K. G. Murthy, H. S. Rangwala, J. Kosecka, W. M. Joiner, J. J. Pancrazio, and S. Sikdar, "Real-time classification of hand motions using ultrasound imaging of forearm muscles," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 8, pp. 1687–1698, 2015.

[17] S. Sikdar, H. Rangwala, E. B. Eastlake, I. A. Hunt, A. J. Nelson, J. Devanathan, A. Shin, and J. J. Pancrazio, "Novel method for predicting dexterous individual finger movements by imaging muscle activity using a wearable ultrasonic system," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 1, pp. 69–76, 2014.

[18] R. C. Sîmpetru, V. Cnejevici, D. Farina, and A. Del Vecchio, "Influence of spatio-temporal filtering on hand kinematics estimation from high-density emg signals," *Journal of Neural Engineering*, vol. 21, no. 2, p. 026014, 2024.

[19] Y. Huang, X. Yang, Y. Li, D. Zhou, K. He, and H. Liu, "Ultrasound-based sensing models for finger motion classification," *IEEE journal of biomedical and health informatics*, vol. 22, no. 5, pp. 1395–1405, 2017.

[20] W. Li, P. Shi, and H. Yu, "Gesture recognition using surface electromyography and deep learning for prostheses hand: state-of-the-art, challenges, and future," *Frontiers in neuroscience*, vol. 15, p. 621885, 2021.

[21] J. He, H. Luo, J. Jia, J. T. Yeow, and N. Jiang, "Wrist and finger gesture recognition with single-element ultrasound signals: A comparison with single-channel surface electromyogram," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 5, pp. 1277–1284, 2018.

[22] X. Yang, J. Yan, Y. Fang, D. Zhou, and H. Liu, "Simultaneous prediction of wrist/hand motion via wearable ultrasound sensing," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 4, pp. 970–977, 2020.

[23] B. Fang, C. Wang, F. Sun, Z. Chen, J. Shan, H. Liu, W. Ding, and W. Liang, "Simultaneous semg recognition of gestures and force levels for interaction with prosthetic hand," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 2426–2436, 2022.

[24] X. Zhang, Z. Wang, R. Fu, D. Wang, X. Chen, X. Guo, and H. Wang, "V-shaped dense denoising convolutional neural network for electrical impedance tomography," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–14, 2022.

[25] H. Li, J. Tan, and H. He, "Magichand: Context-aware dexterous grasping using an anthropomorphic robotic hand," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9895–9901, IEEE, 2020.

[26] R. Fu, Z. Wang, X. Zhang, D. Wang, X. Chen, and H. Wang, "A regularization-guided deep imaging method for electrical impedance tomography," *IEEE Sensors Journal*, vol. 22, no. 9, pp. 8760–8771, 2022.

[27] J. Zeng, Y. Zhou, Y. Yang, J. Yan, and H. Liu, "Fatigue-sensitivity comparison of semg and a-mode ultrasound based hand gesture recognition," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 4, pp. 1718–1725, 2022.

[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 Transactions on Automation Science and Engineering*, vol. 21, no. 1, pp. 475–487, 2022.

[29] H. Zhou and G. Alici, "Non-invasive human-machine interface (hmi) systems with hybrid on-body sensors for controlling upper-limb prosthesis: A review," *IEEE Sensors Journal*, vol. 22, no. 11, pp. 10292–10307, 2022.

[30] Q. Zhang, K. Kim, and N. Sharma, "Prediction of ankle dorsiflexion moment by combined ultrasound sonography and electromyography," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 1, pp. 318–327, 2020.

[31] K. G. Rabe, T. Lenzi, and N. P. Fey, "Performance of sonomyographic and electromyographic sensing for continuous estimation of joint torque during ambulation on multiple terrains," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 2635–2644, 2021.

[32] Q. Zhang, A. Iyer, Z. Sun, K. Kim, and N. Sharma, "A dual-modal approach using electromyography and sonomyography improves prediction of dynamic ankle movement: A case study," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 1944–1954, 2021.

[33] Q. Zhang, K. Lambeth, Z. Sun, A. Dodson, X. Bao, and N. Sharma, "Evaluation of a fused sonomyography and electromyography-based control on a cable-driven ankle exoskeleton," *IEEE Transactions on Robotics*, vol. 39, no. 3, pp. 2183–2202, 2023.

[34] S. Wei, Y. Zhang, and H. Liu, "A multimodal multilevel converged attention network for hand gesture recognition with hybrid semg and a-mode ultrasound sensing," *IEEE transactions on cybernetics*, vol. 53, no. 12, pp. 7723–7734, 2022.

[35] A. Gijsberts, M. Atzori, C. Castellini, H. Müller, and B. Caputo, "Movement error rate for evaluation of machine learning methods for semg-based hand movement classification," *IEEE transactions on neural systems and rehabilitation engineering*, vol. 22, no. 4, pp. 735–744, 2014.

[36] J. McIntosh, A. Marzo, M. Fraser, and C. Phillips, "Echoflex: Hand gesture recognition using ultrasound imaging," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 1923–1934, 2017.

[37] J. Qi, G. Jiang, G. Li, Y. Sun, and B. Tao, "Surface emg hand gesture recognition system based on pca and grnn," *Neural Computing and Applications*, vol. 32, pp. 6343–6351, 2020.
[38] Z. Lu, S. Cai, B. Chen, Z. Liu, L. Guo, and L. Yao, "Wearable real-time gesture recognition scheme based on a-mode ultrasound," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 2623–2629, 2022.
[39] C. Wang, W. Guo, H. Zhang, L. Guo, C. Huang, and C. Lin, "semg-based continuous estimation of grasp movements by long-short term memory network," *Biomedical Signal Processing and Control*, vol. 59, p. 101774, 2020.
[40] J. G. Ngeo, T. Tamei, and T. Shibata, "Continuous and simultaneous estimation of finger kinematics using inputs from an emg-to-muscle activation model," *Journal of neuroengineering and rehabilitation*, vol. 11, pp. 1–14, 2014.
[41] Z. Taghizadeh, S. Rashidi, and A. Shalbaf, "Finger movements classification based on fractional fourier transform coefficients extracted from surface emg signals," *Biomedical Signal Processing and Control*, vol. 68, p. 102573, 2021.
[42] H. Lee, D. Kim, and Y.-L. Park, "Explainable deep learning model for emg-based finger angle estimation using attention," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 1877–1886, 2022.
[43] D. Zadok, O. Salzman, A. Wolf, and A. M. Bronstein, "Towards predicting fine finger motions from ultrasound images via kinematic representation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 12645–12651, IEEE, 2023.
[44] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2020.
[45] K. Bimbraw, C. J. Nycz, M. J. Schueler, Z. Zhang, and H. K. Zhang, "Prediction of metacarpophalangeal joint angles and classification of hand configurations based on ultrasound imaging of the forearm," in *2022 International Conference on Robotics and Automation (ICRA)*, pp. 91–97, 2022.
[46] D. Xiong, D. Zhang, X. Zhao, and Y. Zhao, "Deep learning for emg-based human-machine interaction: A review," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 3, pp. 512–533, 2021.
[47] M. C. Fiorentino, F. P. Villani, M. Di Cosmo, E. Frontoni, and S. Moccia, "A review on deep-learning algorithms for fetal ultrasound-image analysis," *Medical image analysis*, vol. 83, p. 102629, 2023.