MTRT: Motion Trajectory Reconstruction Transformer for EEG-Based BCI Decoding

Pengpai Wang, Zhongnian Li, Peiliang Gong, Yueying Zhou, Fang Chen, and Daoqiang Zhang, *Senior Member, IEEE*

***Abstract*— Brain computer interface (BCI) is a system that directly uses brain neural activities to communicate with the outside world. Recently, the decoding of the human upper limb based on electroencephalogram (EEG) signals has become an important research branch of BCI. Even though existing research models are capable of decoding upper limb trajectories, the performance needs to be improved to make them more practical for real-world applications. This study is attempt to reconstruct the continuous and nonlinear multi-directional upper limb trajectory based on Chinese sign language. Here, to reconstruct the upper limb motion trajectory effectively, we propose a novel Motion Trajectory Reconstruction Transformer (MTRT) neural network that utilizes the geometric information of human joint points and EEG neural activity signals to decode the upper limb trajectory. Specifically, we use human upper limb bone geometry properties as reconstruction constraints to obtain more accurate trajectory information of the human upper limbs. Furthermore, we propose a MTRT neural network based on this constraint, which uses the shoulder, elbow, and wrist joint point information and EEG signals of brain neural activity during upper limb movement to train its parameters. To validate the model, we collected the synchronization information of EEG signals and upper limb motion joint points of 20 subjects. The experimental results show that the reconstruction model can accurately reconstruct the motion trajectory of the shoulder, elbow, and wrist of the upper limb, achieving superior performance than the compared methods. This research is very meaningful to decode the limb motion parameters for BCI, and it is inspiring for the motion decoding of other limbs and other joints.**

Manuscript received 15 September 2022; revised 14 April 2023; accepted 8 May 2023. Date of publication 11 May 2023; date of current version 22 May 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62136004, Grant 61876082, and Grant 61732006; and in part by the National Key Research and Development Program of China under Grant 2018YFC2001600 and Grant 2018YFC2001602. *(Corresponding author: Daoqiang Zhang.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Nanjing University of Aeronautics and Astronautics, and performed in line with the Declaration of Helsinki.

Pengpai Wang, Peiliang Gong, Yueying Zhou, Fang Chen, and Daoqiang Zhang are with the Key Laboratory of Brain-Machine Intelligence Technology, Ministry of Education, MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China (e-mail: dqzhang@nuaa.edu.cn).

Zhongnian Li is with the School of Computer Science, China University of Mining Technology, Xuzhou 221116, China.

Digital Object Identifier 10.1109/TNSRE.2023.3275172

***Index Terms*— Brain computer interface, EEG, limb motion decoding, trajectory reconstruction, transformer, motor execution.**

I. Introduction

BRAIN computer interface (BCI) system can directly communicate with external equipment without relying on muscles and nerves tissue. BCI can be applied in many fields, such as helping patients with spinal cord injury, amyotrophic lateral sclerosis, atresia syndrome and other patients with brain-control equipment [1], [2], [3], [4], improving their ability to take care of themselves and communicate with people [5], [6], [7], and helping stroke patients with rehabilitation training [8], [9], [10]. It also has great potential in the game field [11], [12], [13]. There are information communication media involved in BCI, such as, magnetic resonance images, near-infrared, electroencephalogram (EEG) [14], [15], [16]. Among them, noninvasive EEG has the advantages of high time resolution, low price, high practicability, and convenient acquisition, and is widely used in the field of brain computer interface [17], [18], [19].

Recently, researches show that EEG signals contain various motion parameters of limb movement [20], [21], [22]. Decoding motion parameters directly from EEG signals can provide intuitive and natural control [23]. Therefore, we can decode various motion parameters from EEG signals, such as velocity, acceleration, displacement, position, angular velocity, etc. [24], [25], [26]. The reconstruction of upper limb kinematic parameters by EEG signals can not only promote the rehabilitation of patients with stroke or spinal cord injury, but also control the exoskeleton to enhance the strength and endurance of the ordinary human body.

EEG-based upper limb motion trajectory decoding is an important part of limb motion parameter decoding, which has been explored by researchers. Some researchers simply classify the two-dimensional plane movements extending from the center to the outside, and can only identify the limited and specific movement directions of the upper limbs. For example, Úbeda et al. [27] used multiple linear regression based on EEG to classify eight center-out movements of the arm. Zeng et al. [28] used EEG to reconstruct the four-direction movement of the hand centered and extended in two-dimensional (2D). Recent studies have decoded direction-specific motion trajectories of human joints. For example, Jeong et al. [29]

proposed a deep learning framework for the classification of upper limb movements for six-direction arm reaching tasks in three-dimensional (3D) space. Shakibaee et al. [30] used a nonlinear autoregressive network based on EEG to decode the angular change trajectory of the knee joint in extension and flexion of the right knee in the sitting position. Pancholi et al. [31] proposed a deep learning model based on EEG to predict the motion trajectory of one-handed hand grasping and trial lifting. To sum up, these works only use EEG signals for simple direction-specific classification or reconstruction.

However, these literature are far from meeting the requirements of decoding when the upper limbs perform continuous and multidirectional nonlinear motions in 3D space. Therefore, we propose a Motion Trajectory Reconstruction Transformer (MTRT) model based on the geometric constraints of human joints, i.e., we keep the spatial distances between the shoulder and elbow, elbow and wrist joint points at a fixed length during the movement of the upper limbs as reconstruction constraints to obtain more accurate trajectory information of the human upper limbs. Consequently, we try to decode the continuous nonlinear upper limb joint point motion by studying and solving the unique geometric characteristics of upper limb joints. In summary, the main contributions of this paper are in three aspects:

(1) To our knowledge, we are the first to decode the Chinese sign language motion trajectory with continuous multidirectional nonlinear upper limb movements in 3D space using EEG signals.

(2) We introduce a MTRT model by using constraints on the geometric features of the joints to reconstruct the motion trajectories of the joint points of the upper limbs.

(3) The MTRT model has achieved good accuracy and precision in decoding the spatial trajectory of human upper limb skeleton points.

II. Related Works

Extracting motion trajectory features from 3D nonlinear motion EEG signals is often complicated, and the linear trajectory reconstruction model can hardly meet this task. In this paper, inspired by the success of the Transformer model in the fields of natural language processing and sequence signals, we will investigate introducing the Transformer model to the task of motion trajectory reconstruction.

Compared to motion direction classification, motion trajectory reconstruction is more challenging and raises high demands on reconstruction model. To solve this task, Little et al. [32] explored a neural network based on a regularization algorithm to predict the trajectory of the elbow flexion angle to create an upper limb motion prediction model. Kim et al. [33] studied the prediction of 3D hand trajectories by multiple linear regression (MLR). The average correlation coefficient between the predicted trajectory and the actual trajectory is 0.684. Robinson et al. [20] reconstructed the position of 2D hand motion trajectories. Their task involved right-hand movements from the center outward in a random sequence in four different directions. Using a Kalman filter to estimate motion trajectories, they obtained a correlation of 0.60 between recorded and estimated data. Sosnik and Zheng [25] used a MLR model to predict the trajectories of the hands, elbows, and shoulders of seven subjects in a time-series 3D space. The mean Pearson correlation coefficients between the predicted and actual trajectories for the hand, elbow, and shoulder ranged to the highest of 0.49, 0.48, and 0.40, respectively. Mondini et al. [34] reconstructed hand trajectories from low-frequency EEG signals. Motion parameters (2D positions) were regressed from the EEG using a regression method combining partial least squares (PLS) and Kalman filtering. An overall significant online correlation between hand motion trajectories and decoded trajectories was obtained with an average of 0.32. However, we found that during the trajectory decoding process, normal human motion is nonlinear and directionally random, which brings challenges to motion trajectory decoding. Therefore, the focus of this work is to use the EEG signal information during 3D nonlinear motion to capture the position of joint points in space using deep neural networks to decode the motion trajectories of human upper limbs.

The Transformer model [35] was proposed in 2017, and it was an encoder-decoder architecture that generates global dependencies between input and output based on a multi-head attention mechanism. Compared with general deep learning [36], [37], the advantages of the Transformer model lie in the feature learning representation and attention mechanism [38]. Furthermore, the masking mechanism in the Transformer model prevents the Transformer from shadow learning [39]. When this mechanism is used in conjunction with the attention mechanism, it can produce better feature representations in the spatial dimension and learn the tiny features of the dataset when generalizing it. The architectures of models such as classic Long Short-Term Memory (LSTM) require more time to train when dealing with sequential data and lack parallel processing capabilities [40]. The design of the architecture of models such as Convolutional Neural Network (CNN) is not suitable for computing the temporal features of time series data [41]. So we introduce Transformer model to fully utilize the computing power, allow parallelization with attention and feature representation learning, and reduce training time.

III. Data Collection and Pre-Processing

Twenty subjects (25×14-year-old) were recruited according to the experimental setup, including 9 females and 11 males [42]. All subjects were required to be in good health and full of energy without brain surgery or brain-related diseases. The subjects were informed of all experimental procedures and relevant precautions, signed a written informed consent form after expressing their consent, and were given corresponding cash rewards according to the duration of the experiment. The acquisition equipment required in this paper includes an EEG acquisition instrument, Kinect V2, and a computer. To reduce the influence of motion on the EEG acquisition process, a portable wireless EEG acquisition device (NeuSen.W64, Neuracle) was used. The device has 64 channels, including 59 EEG channels, 4 electrooculography, and 1 electrocardiograph channels. Fifty-nine EEG channels were arranged according to the international 10-20 standard.

A. Experimental Setup

One week before the experiment, the subjects were asked to learn relevant Chinese sign language movements. During the experiment, except for the movements of sign language, the lower limbs and trunk were kept still and movements such as eye movements and swallowing were reduced.

The experimental process includes two runs, and the time interval of each run is 15 minutes. Each run includes 22 executions of 30 sign language sentences. The signed sentences are selected from the common sign language database. When the subjects excuted the selected sign language, the upper limbs were greatly expanded, which was convenient for the Kinect to collect movement data. The course of each trial consisted of 2 seconds of preparation time, 3 seconds of execution time, and 3 seconds of rest time. The subjects performed sign language according to the prompt tone and the computer screen cross prompt. At the same time as EEG acquisition, Kinect collects the motion trajectories of the subjects' upper limb joint points, and the sampling rate is 30Hz. We reduce the EEG acquisition device with a sampling rate of 1000Hz to 900Hz, and the impedance of the electrodes is less than 5KΩ. The subjects were relaxed throughout the experiment and were not allowed to open their mouth, swallow or chew. Except for the sign language movement of upper limbs, the movement of other body parts is not allowed. The above measures are used to avoid electromyography (EMG) artifact in EEG.

To facilitate subsequent analysis and reconstruction tasks, we selected 59 channels with EEG signals from 64 channels. The EEG signals were frequency filtered and band-pass filtered at 0.1-100 Hz. Then, we remove the 50 Hz power frequency and electrooculography (EOG) interference signals in the EEG. Next we performed a whole-brain re-reference to the EEG data. The sentence process of the subjects executing the sign language lasted 3s, and most of them were simple sentences with three or four sign language words. To facilitate the training of the joint point motion trajectory reconstruction model, each EEG sequence is cut into trials with a length of 4s (-0.5s∼3.5s).

B. Joint Point Data Pre-Process

We use the Kinect instrument to collect the spatial position of the human upper limb joint points. We found that the collected data is offset. In the process of arm movement, the length of upper arm and lower arm will change due to the limitation of environmental conditions such as clothes or light. However, in practice the length of the arm don't change in movement. Therefore, we need to correct the obtained joint spatial position data to further decode the joint point position more accurately using EEG data later.

To correct the joint points position data, we use the Euler angle and inverse kinematics equation to correct the collected data [43], [44], [45]. Specifically, we first extract the spatial position data of joint points in the preparation stage of the experiment, to calculate the length of the upper arm and lower arm of the subject. Then the origin of the 3D coordinate system is positioned at the shoulder joint point, and the position data of the elbow and wrist joints are updated. In the process of arm movement, different angles in the process of movement are calculated through the coordinates of the arm. Then the spatial position of the new joint point is determined according to the angle and the fixed arm length. Through the same method, the spatial position coordinates of elbow and wrist joint are calculated in turn, as shown in Fig. 1.

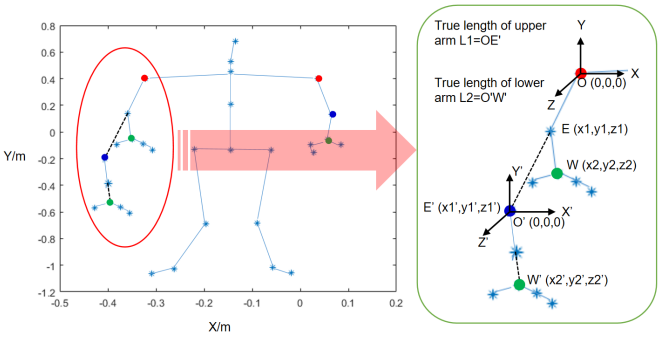

Fig. 1. The 3D coordinates of elbow and wrist joints are updated through the inverse motion equation. The position of the whole joint points collected by Kinect is displayed in the 2D plane. The blue asterisk refers to all human joint points collected. The red, blue and green dots are shoulder, elbow, and wrist nodes, respectively. Taking the right arm as an example, the length of the upper arm collected by the device is the distance of *OE*. and the length of the lower arm is the distance of *EW*. Calculate the angle of *OE* with *X*, *Y*, and *Z* axes in the new coordinate system (3D coordinate system with *O* as the origin) through Euler angle. The 3D coordinates of *E'* are solved by using the inverse motion equation through the angle and the actual arm length. The distance of *OE'* is the real upper arm length *L*1. Similarly, the distance of *O'W'* is the real lower arm length *L*2.

C. Trajectory Reconstruction Data

In our experimant, each subject executed 30 sign language sentences, each sign language sentence included 44 samples, and all subjects had a total of $20\times30\times44 = 26400$ samples. To remove residual noise in EEG electrodes, the standard deviation (SD) of EEG amplitude was calculated for each electrode in all samples, and if the amplitude of a channel exceeded 6 SD, the electrode was marked as a noisy channel. Finally, 9 channels are marked as noise channels. The corresponding electrode names are Fp1, AF7, AF8, F1, F8, T7, T8, TP8, and O2. Next, we performed a residual noise test on all samples and deleted the trial if the EEG amplitude of any one of the residual electrodes in the trial exceeded 6 SD [25], [46]. A total of 4440 samples were removed, and the remaining 21960 samples participated in the training and testing experiments of the MTRT model. The pre-processed EEG data format is samples×timesteps×channels=21960×3600×50, and the pre-processed joint points data format is samples×timesteps×joints=21960×120×18.

IV. Motion Trajectory Reconstruction Transformer

In this section, we introduce our proposed Motion Trajectory Reconstruction Transformer (MTRT) model to reconstruct joint points trajectory of upper limb. The general framework of our model is shown in Fig. 2. First, we performed the acquisition of EEG data and Kinect joint point data. Then, two

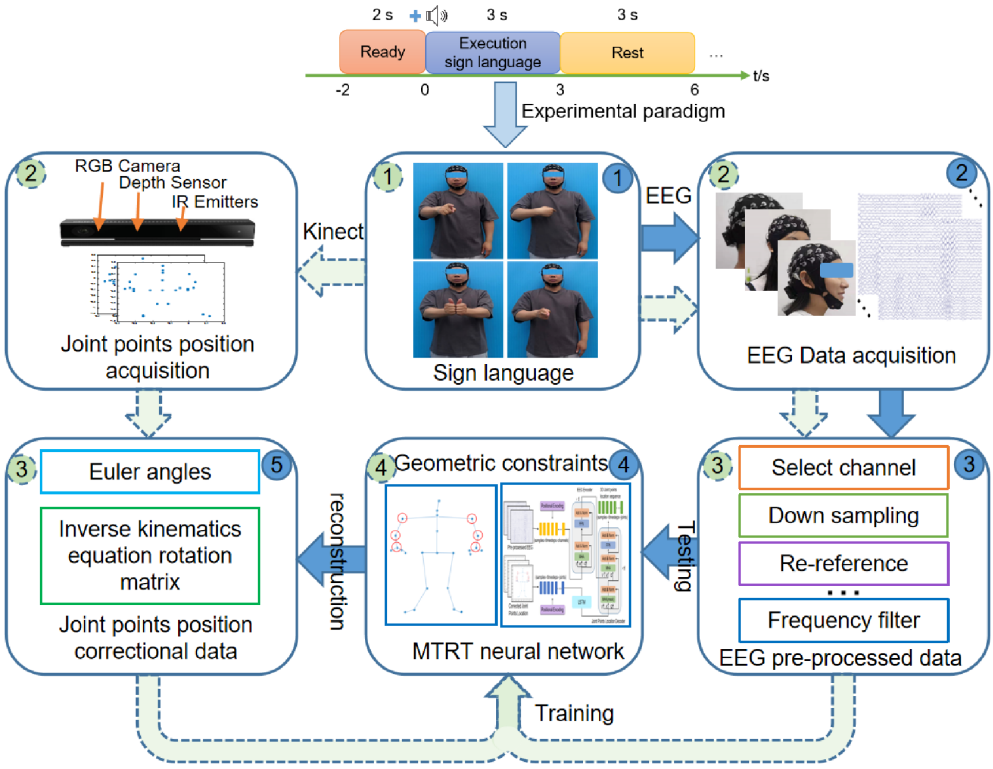

Fig. 2. The architecture of joint point trajectory reconstruction based on Chinese sign language. During the training phase, when the subjects performed sign language follow the experimental paradigm, EEG acquisition equipment and Kinect were used to collect joint point trajectory data of EEG and upper limb movement at the same time. Then we pre-process the EEG data, including channel selection, down sampling, re-reference, and frequency band filtering. We select the data of 6 joint points of shoulder, elbow, and wrist in the left and right arms from the trajectory data of 25 joint points. Further, the Euler angle and inverse motion equation are used to correct the data. Finally, the pre-processed EEG and corrected joint point data are input into the MTRT for model training. During the testing phase, the pre-processed EEG data is reconstructed to the motion trajectory of joint points through the trained MTRT.

kinds of data were pre-processed, and the MTRT model was trained with the pre-processed data. Finally, the trajectory of the joint points is reconstructed by EEG based on the trained model.

In this section, we introduce the EEG encoder and trajectory decoder used to reconstruct sign language motion trajectories. The model framework used in this paper consists of an EEG encoder and a joint point spatial location decoder. The EEG encoder learns the depth representation of the EEG in a self-attention manner, and the joint point spatial location decoder uses the participating EEG representations to generate continuous joint point 3D spatial coordinates.

The MTRT model architecture is shown in Fig. 3. Next we describe the core modules of the MTRT in the following sections. The architecture of the MTRT model encodes the EEG features and decodes the sequence of joint point spatial positions. The model is divided into two parts, including the EEG encoder and the trajectory decoder.

A. EEG Encoder

The model uses 6 encoder layers and a multi-head self-attention module to obtain attention weights for the pre-processed input EEG features. Since the Transformer encoder and decoder are permutation-invariant, we add a fixed sinusoidal spatial positional encoding and object query at the input as the learned positional embedding. The first sub-layer is a complex attention layer, and the second sub-layer is a complex-valued feed-forward network. Both sub-layers have residual connections and normalization layers [47], [48]. Each layer consists of two sub-layers: a multi-head self-attention and a fully connected feed-forward network. Each sub-layer starts with layer normalization to mitigate internal covariate shifts. There is a residual connection around each sub-layer, which preserves the information of the input features and enhances the stability of the model. The output of the Transformer layer is passed to a feed-forward network (FFN) module, which consists of a three-layer perceptron with rectified linear unit (ReLU) activation, and then the final detection predictions. Layer normalization is performed on the remaining connections in the encoder, which is called Norm & Add in the encoder.

To obtain information from different representation subspaces of different modalities at different locations, we combine multiple attention functions to achieve multi-head attention. Multi-head attention is the core module of Transformer, which allows the model to jointly focus on information in different representation sub-spaces at different locations [49]. Multi-Head Attention calculates h (= 8) Scaled Dot-Product Attention, and each Scaled Dot-Product Attention can calculate the corresponding head. Scaled Dot-Product Attention is a mechanism for learning action dependencies in sign language EEG and capturing the internal structure

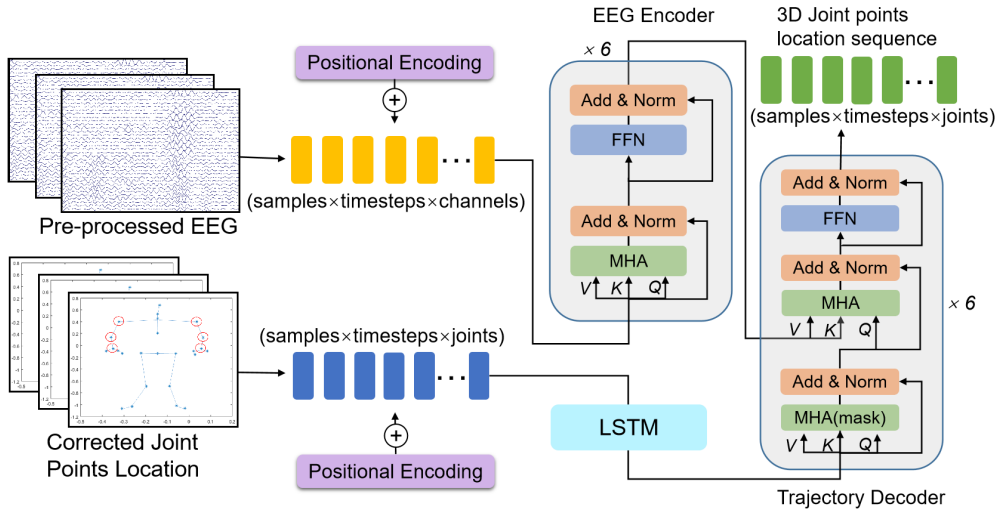

Fig. 3. MTRT model structure for the reconstruction of joint points space location. The model is divided into two parts: EEG encoder and trajectory decoder. The MTRT model has a total of six encoder layers and six decoder layers. Specifically, in the EEG encoder, the preprocessed EEG data and the positional encoding are input to the encoder. It is converted into a representation through six encoder layers, each encoder layer includes a multi-head self-attention layer, a feed-forward neural network layer and a summation and normalization layer (Add & Norm layer). In the trajectory decoder, the output of the EEG encoder is input to the decoder in the form of key and value. During the model training process, the joint point position information and position encoding data are processed through multi-head attention and the Add & Norm layer. The result is a form of a query and the encoder output is fed into the next multi-head attention and Add & Norm layer. Then the next data is predicted through the feedforward neural network and the Add & Norm layer to form an encoder layer.

of EEG. The attention function maps a query and a set of key-value pairs to an output, where the output is computed as a weighted sum of values. The weight assigned to each value is calculated by the query and the corresponding key. The attention function maps a query and a set of key-value pairs to an output, where the output is computed as a weighted sum of the values [50]. The weight assigned to each value is calculated from the query and the corresponding key. We adopt scaled dot product attention because the scaling factor d_v avoids extremely small gradients after softmax. The query for all keys takes the dot product and divides by $\sqrt{d_k}$ [35]. Then apply the softmax function to get the weights for these values. The formula for calculating attention can be written as:

$$\text{Attention}\ (Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

where query $Q \in R^{t\times d_q}$, key $K \in R^{t\times d_k}$, value $V \in R^{t\times d_v}$ and output $O \in R^{t\times d_{model}}$. They are all matrices, t is the sequence length, d_k is the dimension of the query key, d_v is the value dimension, and d_{model} is the output dimension of the encoder which value is set to 256.

The query, key, and value are projected h times onto the d_q, d_k, and d_v dimensions using the learned linearity, where h represents the number of heads. The attention function is then executed in parallel on the projected query, key, and value, computing the d_v dimensional output. The outputs of all heads are connected and linearly projected to deliver the d_{model} dimensional results to the next feed-forward sub-layer. Each head is then concatenated and fed to another linear projection to obtain the final output of multi-head attention. The formula for multi-head attention is as follows:

$$\text{MultiHead}(Q, K, V) = \text{Concat}\left(O_{h_1}, \ldots, O_{h_h}\right)W^o \qquad (2)$$

where $O_{h_i} = \text{Attention}\left(QW_i^Q, K\ W_i^K, V\ W_i^V\right)$, and learnable projection matrices $W_i^Q \in R^{d_{\text{model}}\times d_k}$, $W_i^K \in R^{d_{\text{model}}\times d_k}$, and $W_i^V \in R^{d_{\text{model}}\times d_v}$.

The input features are processed in the Transformer network and their order / position is not preserved, so positional encoding is employed to preserve the temporal order of the features [51]. Therefore, the EEG signal data is encoded with sequence information of sine and cosine functions according to the equation. The input vector and the position encoding vectorby the element-wise addition input into the encoder layer.

The output vector of the embedding layer needs to be computed through six encoder layers. There are two sub-layers in each coding layer. These are self-attention and a fully connected feed-forward layer. In self-attention, the attention function is used for a set of queries, keys, and values, respectively. The calculation of the output matrix is done using equations. Also, normalization is added at the end of each sub-layer. Generate an r-dimensional vector at the end of the encoder and pass it to the decoder.

B. Trajectory Decoder

The structure of the decoder consists of six identical decoder layers and an output layer. The output of the encoder and the decoder content serve as the input for the training of the decoder. Each decoder layer has three sub-layers, namely attention layers, complex FFN and another attention layer. The first attention layer is masked by additional diagonal lines to prevent attention to subsequent positions, to ensure that previous joint point data does not depend on later joint space position data points used for prediction. The second attention will be performed on the encoding representation X_{enc} and the decoder input X_{dec}.

Input a y-dimensional zero vector into the input layer of the decoder. The y is a hyperparameter and its value is set to 1024. The decoder adopts the same multi-head attention as the encoder. Finally, by mapping the output of the last decoder layer to a two-dimensional vector, a vector containing the sequence of spatial coordinates of the three joints shoulder, elbow, and wrist of the human body is obtained. A positional offset is used between the input of the decoder and the target output. We use LSTM as the timing processing module to perform timing processing and looping on the encoder input.

Set the value of the parameter d_{model} to 256. The d_{model} parameter determines the output dimension of the sub-layer output and the embedding layer [35]. Considering the large difference of independent individuals in EEG, each subject is trained separately in this paper, and the input format of the encoder is samples×timesteps×channels. The output format of the decoder is samples×timesteps×joints, where joints represents the sequence representation in the 3D space of the six joints of the left and right shoulders, elbows and wrists (18=2×3). The embedding layers in the encoder and decoder share the same set of weights. The value of parameter h in the calculation of self-attention and multi-head attention is 8. The parameter h refers to the number of parallel attention layers or attention heads. The number of hidden units value in the fully connected sub-layer in the decoder layer is set to 1024. Set the parameters d_k (queries, key vector dimension) and d_v (value vector dimension) to 8. To implement the positional encoding block, the method is the same as in [35], using sine and cosine to implement the positional encoding. The key and value vectors of the final encoding layer are input to the third multi-head attention sub-layer of the decoder layer. The multi-head attention layer in the decoder takes the query vector value from the layer below it.

C. Geometric Information Constraints

To train the MTRT model, we apply loss functions on top of the trajectory decoder outputs, and minimize the errors between reconstruction and groundtruth joint points position. The geometric constraints of skeleton points are only used as loss functions to train the model, not as data. When the training of the model is completed, the model can be directly used to reconstruct the motion trajectory of EEG data.

It is generally known that the lengths of the upper and lower arms are fixed when subjects perform sign language actions, the lengths of the left upper arm and the right upper arm of each healthy subject are equal, and the length of the left lower arm and the right lower arm are equal. Therefore, to reconstruct accurate joint point position data, we set the upper and lower arm lengths to be fixed and the left and right arms to be equal in length as geometric constraints for the MTRT model. The calculation of the geometric features of the upper limbs of the human body is also very simple and convenient. The spatial distances between the shoulder joint point and the elbow joint point, the elbow joint point and the wrist joint point are set as constants $L1$ and $L2$ in the whole movement process. The lengths of the left and right arms are equal to obtain, i.e. $L1_{left} = L1_{right}$, $L2_{left} = L2_{right}$. The geometric characteristics of human joints are used as the constraints of the reconstructed model to obtain more accurate trajectory information of human upper limbs. Therefore, the joint geometry constraint limb distance chanless loss ($LDCL$) and left-right distance equal loss ($LRDE$) set as:

$$LDCL = \frac{1}{m}\sum_{i=1}^{m}\left(\text{len}_i' - \text{len}_i\right)^2 \tag{3}$$

$$LRDE = \frac{1}{m}\sum_{i=1}^{m}\Big(\left(L1_{left_i} - L1_{right_i}\right)^2 + \left(L2_{left_i} - L2_{right_i}\right)^2\Big) \tag{4}$$

where len_i' and len_i are the predicted and actual length values of arm of sample i. $L1_{left_i}$ and $L1_{right_i}$ are the left and right upper limb length of sample i, and $L2_{left_i}$ and $L2_{right_i}$ are the left and right lower limb length of sample i in reconstruction joint points position.

We use the mean squared error (MSE) and the distance loss between the shoulder, elbow, and wrist joint point (the length of the forearm or rear arm is always equal during the movement) to train the MTRT model. The MSE $Loss$ is:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i' - y_i\right)^2 \tag{5}$$

where y_i and y_i' are the actual and predicted values of sample i, respectively. Overall, the loss of the MTRT model, named as trajectory reconstruction loss (TRL), consists of three parts: LDCL, LRDE and Mean Squared Error (MSE), SRL can be written as:

$$TRL = \frac{1}{2\sigma_1^2}LDCL + \frac{1}{2\sigma_2^2}LRDE + \frac{1}{2\sigma_3^2}MSE + \log\sigma_1\sigma_2\sigma_3 \tag{6}$$

where σ_1, σ_2, and σ_3 are the trainable weights of the regression model. It is randomly initialized and iteratively optimized during training.

After the decoder is divided into blocks, the dense layer is used for mapping transformation, and the output of the dense layer is the spatial position sequence of the six joint points. The model was trained for 50 epochs using the Adam optimizer [52] and the batch size was set to 8. The model predicts the 3D position of a joint point in space every 30 timesteps.

V. Results and Discussions

To achieve the goal of sign language motion trajectory, we proposed the MTRT model to reconstruct the joint space trajectory of EEG signals to obtain effective feature acquisition and high reconstruction accuracy. We will describe the performance evaluation of joint motion trajectories reconstruction based on Chinese Sign Language EEG data. Furthermore, we compare the performance of our method and comparative methods. All scripts are written using the deep learning framework of PyTorch 1.10 and CUDA 11.3. The model runs on a Dell precision workstation devices configured with NVIDIA GeForce RTX 2080Ti GPU, 16 GB RAM and Intel i7-9700 CPU@3-GHz, without any special hardware optimization.

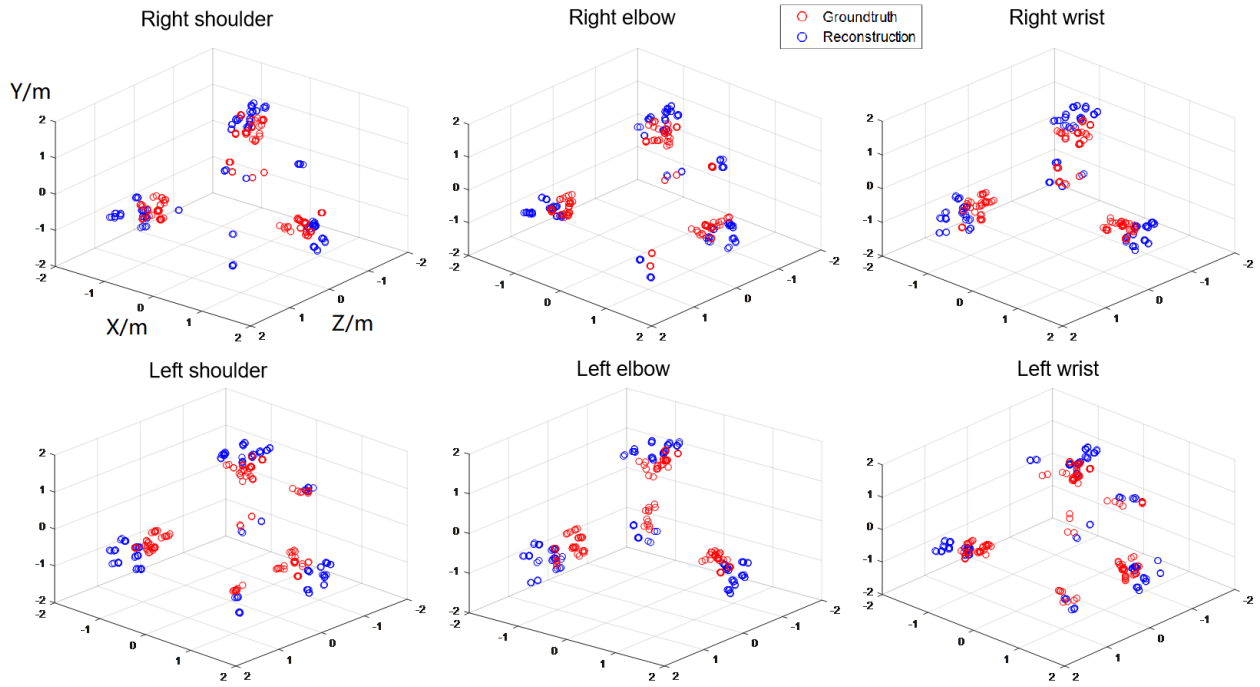

Fig. 4. Groundtruth trajectories and reconstructed trajectories of the left and right shoulders, elbows, and wrists of the 28th handed sentence from the third subject. Its trajectory contains a total of 120 trajectory points and has a duration of four seconds. The red circles are the trajectory points of the groundtruth, and the blue circles represent the trajectory points of the reconstructed joint points.

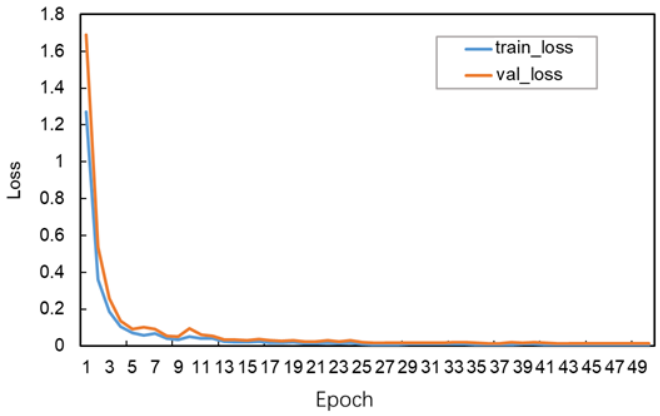

Fig. 5. Training and validation loss changes for 50 epochs of MTRT model training.

After 50 times of training, the MTRT model tends to converge, and the loss decreases steadily with the increase of training times, and finally stabilizes (Fig. 5). Next, we will summarize the experiment data based on Chinese Sign Language and discuss the results of the proposed method.

A. Comparison of Reconstruction Performance of Different Models

In trajectory decoding, a commonly used similarity measure between upper limb motion trajectories and decoded trajectories is the Pearson correlation coefficient ρ [29], [31], [34]. The Pearson correlation coefficient between true trajectories and decoded trajectories is defined as the quotient of their covariance and standard deviation, and its formula can be written as:

$$\rho_{(G,R)} = \frac{\mathrm{cov(G,R)}}{\sigma_G \sigma_R} = \frac{\sum_{i=1}^{n}\left(G_i - \bar{G}\right)\left(R_i - \bar{R}\right)}{\sqrt{\sum_{i=1}^{n}\left(G_i - \bar{G}\right)^2}\sqrt{\sum_{i=1}^{n}\left(R_i - \bar{R}\right)^2}} \tag{7}$$

where $\rho_{(G,R)}$ represents the Pearson correlation coefficient between the reconstructed joint point value G and the real value R, $\bar{G}$ and $\bar{R}$ represents the sample average of G and R, respectively.

In addition, we used the Normalized Root Mean Square Error (NRMSE) model performance measurement method as follows [53]. We measured the decoding performance of the 3D axis using NRMSE, and the results were the average values of the X-axis, Y-axis, and Z-axis. The total average NRMSE of the 3D axis for the six joint points in sign language using the proposed method is 0.159.

To measure the performance of our trained MTRT model, a Pearson correlation coefficient was calculated for the comparative model and our model. Comparing models in our nonlinear multi-directional motion trajectory detection task include MLR [25], LSTM [54] and CNN-LSTM [55], sequence-to-sequence (Seq2seq) [56], and Transformer models. The Transformer model is our MTRT model that only uses MSE loss function.

We randomize the data before starting training, excute a performance comparison experiment on the test data set of 20 subjects, and calculate the average value of ρ between the reconstructed trajectory and the real trajectory. Comparative experiments were carried out using the same hardware and software environment. The three dimensions X, Y, and Z of

TABLE I

PEARSON CORRELATION COEFFICIENTS BETWEEN THE RECONSTRUCTION OF THE MOVEMENT TRAJECTORIES OF THE SHOULDER, ELBOW, AND WRIST DURING THE EXECUTION OF SIGN LANGUAGE AND THE TRUTH SIGN LANGUAGE MOVEMENT TRAJECTORIES FOR THE REGRESSION MODELS

Model	Pearson Correlation Coefficient																		
	Right Shoulder			Right elbow			Right Wrist			Left Shoulder			Left elbow			Left Wrist			Mean
	X	Y	Z	X	Y	Z	X	Y	Z	X	Y	Z	X	Y	Z	X	Y	Z	
MLR	0.15±0.22	0.17±0.17	0.22±0.14	0.14±0.12	0.05±0.10	0.19±0.11	0.21±0.06	0.14±0.13	0.37±0.13	0.14±0.15	0.09±0.21	0.25±0.18	0.31±0.31	0.23±0.20	0.24±0.42	0.18±0.11	0.24±0.25	0.11±0.24	0.19±0.20
LSTM	0.21±0.19	0.29±0.16	0.23±0.31	0.4±0.24	0.31±0.10	0.36±0.23	0.48±0.25	0.26±0.23	0.44±0.15	0.35±0.34	0.27±0.21	0.24±0.14	0.43±0.32	0.36±0.0.22	0.41±0.17	0.49±0.16	0.51±0.23	0.44±0.16	0.36±0.27
CNN-LSTM	0.43±0.14	0.38±0.09	0.41±0.11	0.52±0.17	0.45±0.13	0.40±0.30	0.55±0.26	0.56±0.31	0.51±0.26	0.37±0.11	0.34±0.12	0.41±0.34	0.42±0.23	0.45±0.26	0.39±0.07	0.44±0.31	0.61±0.33	0.58±0.16	0.46±0.26
Seq2seq	0.33±0.19	0.51±0.12	0.43±0.28	0.42±0.16	0.37±0.28	0.3±0.12	0.45±0.18	0.35±0.31	0.57±0.36	0.29±0.17	0.35±0.33	0.28±0.15	0.36±0.26	0.5±0.16	0.33±0.21	0.61±0.16	0.41±0.17	0.47±0.18	0.41±0.24
Transformer	0.85±0.12	0.83±0.14	0.81±0.15	0.88±0.07	0.85±0.13	0.83±0.14	0.92±0.06	0.89±0.10	0.85±0.13	0.85±0.10	0.81±0.17	0.78±0.23	0.87±0.06	0.86±0.11	0.84±0.13	0.91±0.06	0.85±0.11	0.86±0.16	0.85±0.21
MTRT (Ours)	**0.89±0.10**	**0.93±0.15**	**0.92±0.14**	**0.95±0.08**	**0.94±0.13**	**0.96±0.07**	**0.97±0.08**	**0.98±0.12**	**0.96±0.11**	**0.95±0.12**	**0.87±0.17**	**0.86±0.20**	**0.98±0.07**	**0.97±0.08**	**0.91±0.15**	**0.98±0.12**	**0.97±0.08**	**0.95±0.14**	**0.94±0.13**

TABLE II

NRMSE OF RECONSTRUCTION PERFORMANCE COMPARISON WITH SIX METHODS

Model	NRMSE					
	Right Shoulder	Right Elbow	Right Wrist	Left Shoulder	Left Elbow	Left Wrist
MLR	0.326	0.353	0.284	0.351	0.262	0.336
LSTM	0.275	0.247	0.300	0.250	0.274	0.261
CNN-LSTM	0.235	0.331	0.213	0.207	0.274	0.264
Seq2seq	0.253	0.203	0.215	0.247	0.257	0.287
Transformer	0.197	0.168	0.206	0.211	0.166	0.221
MTRT(Ours)	0.182	0.165	0.147	0.177	0.130	0.155

the six joints of the shoulder, elbow, and wrist in the left and right directions are compared with ten-fold cross-validation. As shown in Table I, the Pearson correlation coefficient between the spatial position of each joint point reconstructed by the MTRT model and the real value is consistently higher than those of the comparison models, achieving a mean value of 0.94.

B. Comparison of Initial and Reconstructed Spatial Coordinates

To visualize the reconstruction performance of our proposed model for the left and right joint motion trajectories of the shoulder, elbow, and wrist, we selected the data of the third subject for visualization. Input the EEG data in the test set into the trained MTRT model to get the 3D position coordinates of the six joints. Six joint points trajectories of groundtruth and reconstruction as shown in Fig. 4. The groundtruth and reconstructed joint point trajectories distributions are similar, indicating the superiority of the performance of our MTRT reconstruction model. It shows a comparison diagram of the reconstructed 3D coordinates of the right shoulder, right elbow, and right wrist and the real coordinates. It can be seen from the visualized experimental results that the 3D coordinates of the real joints are basically similar to the spatial distribution of the reconstructed 3D coordinates of the joints. There are some points whose distribution is far from the true value, such as the right wrist and left elbow both predict isolated points that are quite different from the true value. Fig. 6 shows the reconstruction results and real value distribution of the 6 joint points in the three dimensions of X, Y, and Z. The distribution core densities in the figure are very similar, indicating that the reconstruction model is effective.

C. Activated Brain Regions for 3D Motion

During the whole process of sign language execution, we intercepted the 4 s segment, and the time range is from -0.5 s to 3.5 s. In this subsection, we selected the sentence of the 28th sign language, and averaged all the sentence samples

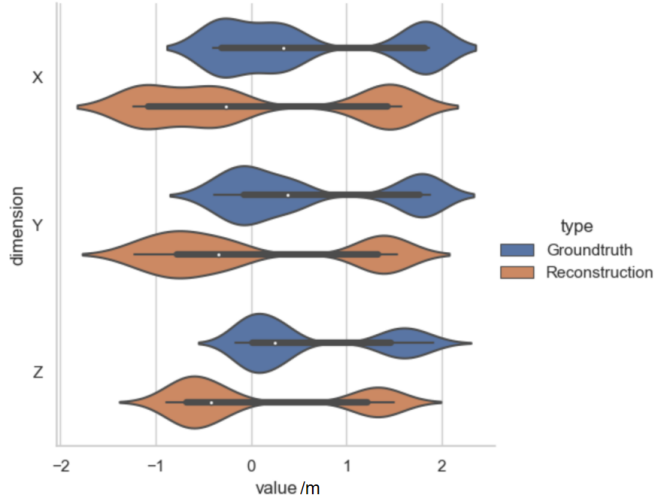

Fig. 6. Reconstruction results and true value distributions in three dimensions for all relevant nodes.

to obtain the EEG map showing the area of electrode activation as shown in Fig. 7. A total of nine pictures are obtained, and the interval between each picture is 400 ms. It can be seen from the figure that the main brain area activated by joint movements of the upper limbs during sign language is the parietal lobe area, and the main electrodes are FC2, FC3, C3, C4, CP3, CP4, P3, and P4. This provides reference value for later decoding trajectories with fewer channels.

D. Limitation and Future Directions

In this paper, the MTRT model is used to reconstruct the 3D space position of the joints of the upper limbs of the human body and achieve good reconstruction results. The Pearson correlation coefficient with the true value is also high, indicating a strong correlation between them. Due to the use of Kinect equipment to collect motion trajectories of limb joint points and adding interference factors such as light and clothing, the collected joints motion trajectories data and the real joint motion trajectory data have a certain deviation. In the

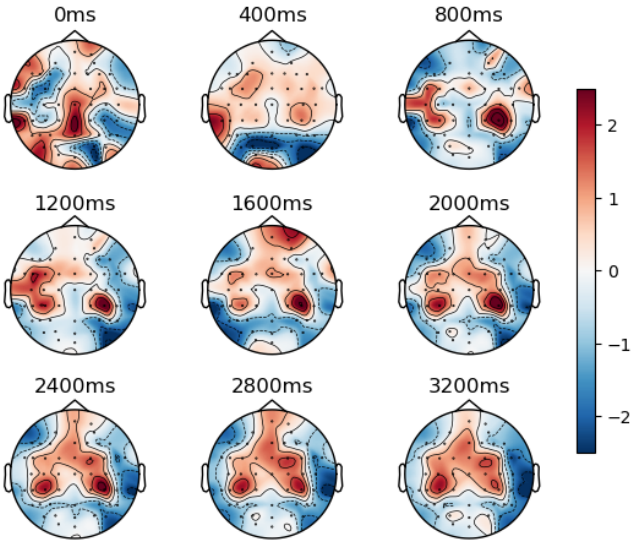

Fig. 7. Mean EEG electrode activity mapping at 400ms intervals during the execution of the third handed sentence by the tenth subject.

future, we will fix the spatial sensors at the joints of the limbs to obtain more accurate spatial motion trajectory data and make the reconstructed data more accurate. It is also possible to reconstruct the motion trajectories of finer joints, such as knuckles. Furthermore, the MTRT model has relatively high requirements on the amount of data, and more data samples can train a more accurate reconstruction model. In the next step, we plan to collect more joint motion data into the model for training, and obtain a model with better robustness and accuracy by training a large number of data samples.

VI. Conclusion

BCI can directly use brain neural activity to exchange information with external devices. It can help people with physical disabilities to interact with the outside world. The field of BCI research includes brain-controlled devices, speech, and text output, among which the direct acquisition of human body movement information based on EEG signals is an important direction. Existing studies have only explored the decoding of simple linear limb movements through EEG signals, and the decoding accuracy is not high. In this paper, the MTRT model is trained by the 3D spatial information of the joint points and the EEG signal, to decode the motion trajectory of the joint points of the upper limbs. We constrain the reconstructed model according to the geometric characteristics of the relative distance of human joint points to obtain more accurate joint point motion trajectories. To verify the performance of our MTRT model, we collected EEG signals and joint motion information of 20 subjects. The experimental results show that our proposed model can decode the complex multi-directional nonlinear upper limb motion trajectory based on Chinese Sign Language. Our research is meaningful to decode human motion information based on EEG signals, and provides a reference for decoding other joint points of the body. Our method can be used in the future for precise manipulation of external devices, such as robotic arms. In addition, it can also be used for remote control of special equipment.

References

[1] P. Ofner, A. Schwarz, J. Pereira, D. Wyss, R. Wildburger, and G. R. Müller-Putz, "Attempted arm and hand movements can be decoded from low-frequency EEG from persons with spinal cord injury," *Sci. Rep.*, vol. 9, no. 1, pp. 1–15, May 2019.

[2] V. Chamola, A. Vineet, A. Nayyar, and E. Hossain, "Brain-computer interface-based humanoid control: A review," *Sensors*, vol. 20, no. 13, p. 3620, Jun. 2020.

[3] X. Gu et al., "EEG-based brain-computer interfaces (BCIs): A survey of recent studies on signal sensing technologies and computational intelligence approaches and their applications," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 5, pp. 1645–1666, Sep. 2021.

[4] X. Gao, Y. Wang, X. Chen, and S. Gao, "Interface, interaction, and intelligence in generalized brain-computer interfaces," *Trends Cognit. Sci.*, vol. 25, no. 8, pp. 671–684, Aug. 2021.

[5] O. B. Guney, M. Oblokulov, and H. Ozkan, "A deep neural network for SSVEP-based brain-computer interfaces," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 2, pp. 932–944, Feb. 2022.

[6] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature*, vol. 568, no. 7753, pp. 493–498, Apr. 2019.

[7] R. Abiri, S. Borhani, E. W. Sellers, Y. Jiang, and X. Zhao, "A comprehensive review of EEG-based brain-computer interface paradigms," *J. Neural Eng.*, vol. 16, no. 1, Feb. 2019, Art. no. 011001.

[8] L. Zhou, X. Tao, F. He, P. Zhou, and H. Qi, "Reducing false triggering caused by irrelevant mental activities in brain-computer interface based on motor imagery," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 9, pp. 3638–3648, Sep. 2021.

[9] E. Mihelj, M. Bächinger, S. Kikkert, K. Ruddy, and N. Wenderoth, "Mental individuation of imagined finger movements can be achieved using TMS-based neurofeedback," *NeuroImage*, vol. 242, Nov. 2021, Art. no. 118463.

[10] Z. Yuan et al., "Effect of BCI-controlled pedaling training system with multiple modalities of feedback on motor and cognitive function rehabilitation of early subacute stroke patients," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 2569–2577, 2021.

[11] T. Mondéjar, R. Hervás, E. Johnson, C. Gutiérrez-López-Franca, and J. M. Latorre, "Analyzing EEG waves to support the design of serious games for cognitive training," *J. Ambient Intell. Humanized Comput.*, vol. 10, no. 6, pp. 2161–2174, Jun. 2019.

[12] A. E. Alchalabi, S. Shirmohammadi, A. N. Eddin, and M. Elsharnouby, "FOCUS: Detecting ADHD patients by an EEG-based serious game," *IEEE Trans. Instrum. Meas.*, vol. 67, no. 7, pp. 1512–1520, Jul. 2018.

[13] B. Kerous, F. Skola, and F. Liarokapis, "EEG-based BCI and video games: A progress report," *Virtual Reality*, vol. 22, no. 2, pp. 119–135, Jun. 2018.

[14] B. Du, X. Cheng, Y. Duan, and H. Ning, "FMRI brain decoding and its applications in brain-computer interface: A survey," *Brain Sci.*, vol. 12, no. 2, p. 228, Feb. 2022.

[15] Y. Kwak, W.-J. Song, and S.-E. Kim, "FGANet: FNIRS-guided attention network for hybrid EEG-fNIRS brain-computer interfaces," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 329–339, 2022.

[16] A. Kline, N. D. Forkert, B. Felfeliyan, D. Pittman, B. Goodyear, and J. Ronsky, "FMRI-informed EEG for brain mapping of imagined lower limb movement: Feasibility of a brain computer interface," *J. Neurosci. Methods*, vol. 363, Nov. 2021, Art. no. 109339.

[17] Y. Miao and V. J. Koomson, "A CMOS-based bidirectional brain machine interface system with integrated fdNIRS and tDCS for closed-loop brain stimulation," *IEEE Trans. Biomed. Circuits Syst.*, vol. 12, no. 3, pp. 554–563, Jun. 2018.

[18] X. Zhang, L. Yao, X. Wang, J. Monaghan, D. McAlpine, and Y. Zhang, "A survey on deep learning-based non-invasive brain signals: Recent advances and new frontiers," *J. Neural Eng.*, vol. 18, no. 3, Jun. 2021, Art. no. 031002.

[19] P. D. E. Baniqued et al., "Brain-computer interface robotics for hand rehabilitation after stroke: A systematic review," *J. Neuroeng. Rehabil.*, vol. 18, no. 1, pp. 1–25, 2021.

[20] N. Robinson, W. J. Chester, and S. Kg, "Use of mobile EEG in decoding hand movement speed and position," *IEEE Trans. Hum.-Mach. Syst.*, vol. 51, no. 2, pp. 120–129, Apr. 2021.

[21] C. Ieracitano, F. C. Morabito, A. Hussain, and N. Mammone, "A hybrid-domain deep learning-based BCI for discriminating hand motion planning from EEG sources," *Int. J. Neural Syst.*, vol. 31, no. 9, Sep. 2021, Art. no. 2150038.

[22] L. Mercado et al., "Decoding the torque of lower limb joints from EEG recordings of pre-gait movements using a machine learning scheme," *Neurocomputing*, vol. 446, pp. 118–129, Jul. 2021.

[23] A. Stroh and J. Desai, "Hand gesture-based artificial neural network trained hybrid human–machine interface system to navigate a powered wheelchair," *J. Bionic Eng.*, vol. 18, no. 5, pp. 1045–1058, Sep. 2021.

[24] T. Blom, S. Bode, and H. Hogendoorn, "The time-course of prediction formation and revision in human visual motion processing," *Cortex*, vol. 138, pp. 191–202, May 2021.

[25] R. Sosnik and L. Zheng, "Reconstruction of hand, elbow and shoulder actual and imagined trajectories in 3D space using EEG current source dipoles," *J. Neural Eng.*, vol. 18, no. 5, Oct. 2021, Art. no. 056011.

[26] A. Buerkle, W. Eaton, N. Lohse, T. Bamber, and P. Ferreira, "EEG based arm movement intention recognition towards enhanced safety in symbiotic human–robot collaboration," *Robot. Comput.-Integr. Manuf.*, vol. 70, Aug. 2021, Art. no. 102137.

[27] A. Úbeda, J. M. Azorín, R. Chavarriaga, and J. D. R. Millán, "Classification of upper limb center-out reaching tasks by means of EEG-based continuous decoding techniques," *J. NeuroEng. Rehabil.*, vol. 14, no. 1, pp. 1–14, Dec. 2017.

[28] H. Zeng et al., "The advantage of low-delta electroencephalogram phase feature for reconstructing the center-out reaching hand movements," *Frontiers Neurosci.*, vol. 13, p. 480, May 2019.

[29] J. Jeong, K. Shim, D. Kim, and S. Lee, "Brain-controlled robotic arm system based on multi-directional CNN-BiLSTM network using EEG signals," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 5, pp. 1226–1238, May 2020.

[30] F. Shakibaee, E. Mottaghi, H. R. Kobravi, and M. Ghoshuni, "Decoding knee angle trajectory from electroencephalogram signal using NARX neural network and a new channel selection algorithm," *Biomed. Phys. Eng. Exp.*, vol. 5, no. 2, Jan. 2019, Art. no. 025024.

[31] S. Pancholi, A. Giri, A. Jain, L. Kumar, and S. Roy, "Source aware deep learning framework for hand kinematic reconstruction using EEG signal," 2021, *arXiv:2103.13862*.

[32] K. Little, B. K. Pappachan, S. Yang, B. Noronha, D. Campolo, and D. Accoto, "Elbow motion trajectory prediction using a multi-modal wearable system: A comparative analysis of machine learning techniques," *Sensors*, vol. 21, no. 2, p. 498, Jan. 2021.

[33] Y. J. Kim et al., "A study on a robot arm driven by three-dimensional trajectories predicted from non-invasive neural signals," *Biomed. Eng. OnLine*, vol. 14, no. 1, pp. 1–19, Dec. 2015.

[34] V. Mondini, R. J. Kobler, A. I. Sburlea, and G. R. Müller-Putz, "Continuous low-frequency EEG decoding of arm movement for closed-loop, natural control of a robotic arm," *J. Neural Eng.*, vol. 17, no. 4, Aug. 2020, Art. no. 046031.

[35] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[36] X. Deng, J. Zhu, and S. Yang, "SFE-Net: EEG-based emotion recognition with symmetrical spatial feature extraction," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 2391–2400.

[37] R. Li, Y. Wang, and B.-L. Lu, "A multi-domain adaptive graph convolutional network for EEG-based emotion recognition," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 5565–5573.

[38] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.

[39] D. Kostas, S. Aroca-Ouellette, and F. Rudzicz, "BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data," *Frontiers Hum. Neurosci.*, vol. 15, Jun. 2021, Art. no. 653659.

[40] P. Nagabushanam, S. Thomas George, and S. Radha, "EEG signal classification using LSTM and improved neural network algorithms," *Soft Comput.*, vol. 24, no. 13, pp. 9981–10003, Jul. 2020.

[41] N. Mammone, C. Ieracitano, and F. C. Morabito, "A deep CNN approach to decode motor preparation of upper limbs from time-frequency maps of EEG signals at source level," *Neural Netw.*, vol. 124, pp. 357–372, Apr. 2020.

[42] P. Wang, Y. Zhou, Z. Li, S. Huang, and D. Zhang, "Neural decoding of Chinese sign language with machine learning for brain-computer interfaces," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 2721–2732, 2021.

[43] J. Li, C. Xu, Z. Chen, S. Bian, L. Yang, and C. Lu, "HybrIK: A hybrid analytical-neural inverse kinematics solution for 3D human pose and shape estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3383–3393.

[44] D. Pavllo, C. Feichtenhofer, M. Auli, and D. Grangier, "Modeling human motion with quaternion-based neural networks," *Int. J. Comput. Vis.*, vol. 128, no. 4, pp. 855–872, Apr. 2020.

[45] Z. Wang, J. Chang, B. Li, C. Wang, and C. Liu, "Kinematics solution of snake-like manipulator based on improved backbone mode method," in *Proc. IEEE Int. Conf. Mechatronics Autom. (ICMA)*, Oct. 2020, pp. 1774–1779.

[46] R. Sosnik and O. B. Zur, "Reconstruction of hand, elbow and shoulder actual and imagined trajectories in 3D space using EEG slow cortical potentials," *J. Neural Eng.*, vol. 17, no. 1, Feb. 2020, Art. no. 016065.

[47] A. Bhattacharya, T. Baweja, and S. P. K. Karri, "Epileptic seizure prediction using deep transformer model," *Int. J. Neural Syst.*, vol. 32, no. 2, Feb. 2022, Art. no. 2150058.

[48] J. Zhao, Y. Zhao, and J. Li, "M3TR: Multi-modal multi-label recognition with transformer," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 469–477.

[49] Y. Huang, H. Xue, B. Liu, and Y. Lu, "Unifying multimodal transformer for bi-directional image and text generation," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1138–1147.

[50] Z. Jia, Y. Lin, X. Cai, H. Chen, H. Gou, and J. Wang, "SST-EmotionNet: Spatial–spectral–temporal based attention 3D dense network for EEG emotion recognition," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2909–2917.

[51] Y. Zhang, B. Wu, W. Li, L. Duan, and C. Gan, "STST: Spatial–temporal specialized transformer for skeleton-based action recognition," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 3229–3237.

[52] Q. Tong, G. Liang, and J. Bi, "Calibrating the adaptive learning rate to improve convergence of Adam," *Neurocomputing*, vol. 481, pp. 333–356, Apr. 2022.

[53] M. Spuler, A. Sarasola-Sanz, N. Birbaumer, W. Rosenstiel, and A. Ramos-Murguialday, "Comparing metrics to evaluate performance of regression methods for decoding of neural signals," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2015, pp. 1083–1086.

[54] Y.-F. Chen et al., "Continuous bimanual trajectory decoding of coordinated movement from EEG signals," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 12, pp. 6012–6023, Dec. 2022.

[55] S. Pancholi, A. Giri, A. Jain, L. Kumar, and S. Roy, "Source aware deep learning framework for hand kinematic reconstruction using EEG signal," *IEEE Trans. Cybern.*, early access, May 9, 2022, doi: 10.1109/TCYB.2022.3166604.

[56] Y.-E. Lee and S.-H. Lee, "EEG-transformer: Self-attention from transformer architecture for decoding EEG of imagined speech," in *Proc. 10th Int. Winter Conf. Brain-Comput. Interface (BCI)*, Feb. 2022, pp. 1–4.