Revisiting Noise Resilience Strategies in Gesture Recognition: Short-Term Enhancement in Surface Electromyographic Signal Analysis

Weiyu Guo[1] , **Ziyue Qiao**[2] , **Ying Sun**[1] and **Hui Xiong**[1]
[1]The Hong Kong University of Science and Technology, Guangzhou
[2]Great Bay University, Dongguan
{guoweiyu96, ziyuejoe}@gmail.com, {yings, xionghui}@ust.hk

Abstract

Gesture recognition based on surface electromyography (sEMG) has been gaining importance in many 3D Interactive Scenes. However, sEMG is easily influenced by various forms of noise in real-world environments, leading to challenges in providing long-term stable interactions through sEMG. Existing methods often struggle to enhance model noise resilience through various predefined data augmentation techniques. In this work, we revisit the problem from a short term enhancement perspective to improve precision and robustness against various common noisy scenarios with learnable denoise using sEMG intrinsic pattern information and sliding-window attention. We propose a Short Term Enhancement Module(STEM) which can be easily integrated with various models. STEM offers several benefits: 1) Learnable denoise, enabling noise reduction without manual data augmentation; 2) Scalability, adaptable to various models; and 3) Cost-effectiveness, achieving short-term enhancement through minimal weight-sharing in an efficient attention mechanism. In particular, we incorporate STEM into a transformer, creating the Short Term Enhanced Transformer (STET). Compared with best-competing approaches, the impact of noise on STET is reduced by more than 20%. We also report promising results on both classification and regression datasets and demonstrate that STEM generalizes across different gesture recognition tasks.

1 Introduction

Surface Electromyographic (sEMG) is a non-invasive technique for motoring muscle neurons firing, which is an effective way to capture human motion intention and has shown great application potential in the field of human-computer interaction (HCI) [Xiong *et al.*, 2021; Liu *et al.*, 2021b; Liu *et al.*, 2020]. A schematic diagram of the EMG-based HCI System is shown in Figure 1. Compared to traditional HCI channels, sEMG has the advantages of being generated prior to actual motion (50-150 ms), containing rich motion intention information, and being easy to collect [Sun *et al.*, 2020]. Therefore, there has been increasing interest in exploring EMG-based motion track [Liu *et al.*, 2021a] and pathological analysis.

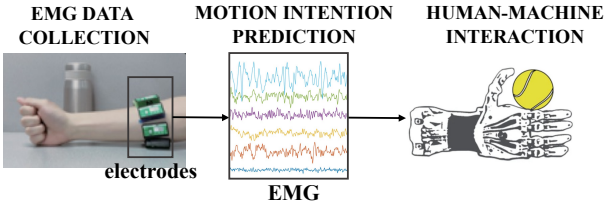


Figure 1: Schematic Diagram of EMG-based Human-Computer Interaction System.

By treating sEMG as time series, deep sequential models [Bi *et al.*, 2019; Tsinganos *et al.*, 2019; Becker *et al.*, 2018; Li *et al.*, 2021; Du *et al.*, 2017] have been applied to sEMG modeling. For example, Zhang *et al.* employs a multi-task encoder-decoder framework improve the robustness of sEMG-based Sign Language Translation (SLT) [Zhang *et al.*, 2022]. Rahimian *et al.* employs Vision Transformer (ViT)-based architecture (TEMGNet) to improve the accuracy of sEMG-based myocontrol of prosthetics [Rahimian *et al.*, 2021]. Although these methods demonstrate enhanced performance compared to traditional approaches, they process sEMG signals as generic time-series data, without specifically tailoring their design to address the unique characteristics of sEMG, such as their high variability and sensitive to external noise and interference. This oversight leads to issues such as difficulty in handling low signal-to-noise ratios due to changes in skin surface conditions and signal interference, and failing to capture subtle but important motion information in sEMG signals. As a result, the robustness and accuracy of existing models remain significantly challenged.

Processing sEMG signals is challenging due to the complex noise mixed in the skin's surface and the presence of patterns across various time scales. Existing works, mainly focusing on long-term sequences, have used transformers to treat sEMG as a typical time series, aiming to enhance long-term dependencies. These approaches overlook the critical features which present in short-time scales. Short-time scale features are important in sEMG analysis, as they aid in distinguishing subtle movements and facilitate the removal of variable noise. For example, gestures like Index Finger Extension (IFE) and Middle Extension (ME), while similar in

global sEMG patterns, can be differentiated through localized short-term signal variations.

To this end, in this paper, we present a lightweight but powerful Module called Short-Term Enhanced Module (STEM) which utilizes sliding window attention with weight sharing to capture short-term features. Building on STEM, we further propose the Short-Term Enhanced Transformer (STET). STET leverages STEM to capture local signal changes, enhancing noise resistance, and then combines STEM with long-term features, further improving accuracy for downstream predictions. Furthermore, to enhance model robustness with minimal annotation, we propose a self-supervised paradigm based on sEMG Signal Masking to leverage the inherent variability in sEMG signals.

Finally, we conducted extensive experiments on the largest public sEMG datasets. The experimental results show that STET surpasses existing methods by a significant margin in both gesture classification and joint angle regression tasks for single-finger, multi-finger, wrist, and rest gestures. Meanwhile, STET achieves strong robustness even when trained on pure data and tested on noisy data. Compared with best-competing approaches, the impact of noise on STET is reduced by more than 20%. Moreover, through visualizations, we show that the long-term and short-term features are complementary in sEMG-based gesture recognition tasks, and the fusion of the two features can make the classification boundary more obvious. This clearly demonstrates that short-term information is critical for sEMG-based gesture recognition and will provide a new design paradigm for future sEMG model design. In particular, we have deployed STET as an important functional component in our HCI system, which can offer a more intuitive and effective experience. Our real-world deployment is shown in the appendix.

To the best of our knowledge, we are the first to highlight the short-term features in sEMG-based gesture recognition. Our contributions can be summarized as follows:

1) From the perspective of enhancing short-term features, we propose STEM, a learnable, scalable, and low-cost noise-resistant module. The integration of STEM into various neural networks has resulted in a marked improvement in their performance;
2) we introduce sEMG Signal Masking to self-supervised sEMG Intrinsic Pattern Capture Module to leverage the inherent variability in sEMG;
3) we conduct experiments on the largest wrist sEMG dataset, showing that our proposed method outperforms existing approaches in terms of accuracy and robustness. And short-term enhancement can be extended to other models like Informer.

2 RELATED WORK

The EMG-based Intention Prediction of Human Motion can be broadly divided into model-based and data-driven methods. Model-based methods typically combine disciplines such as kinesiology, biomechanics, and human dynamics to explicitly model the relationship between EMG and outputs (such as joint angles and forces). The model often includes specific parameters, such as joint positions and bone-on-bone friction, that need to be repeatedly experimented with and adjusted until the desired performance is achieved. In terms of parameter selection and determination, model-based methods can be further divided into kinematic models [Borbély and Szolgay, 2017], dynamic models [Koike and Kawato, 1995; Koirala *et al.*, 2015; Liu *et al.*, 2015], and muscle-bone models [Wang and Buchanan, 2002; Zhao *et al.*, 2020; Yao *et al.*, 2018]. Clancy *et al.* used a nonlinear dynamics model to identify the relationship between constant posture electromyography and torque at the elbow joint [Clancy *et al.*, 2012]. Hashemi *et al.* used the Parallel Cascade Identification method to establish a mapping between forearm muscles and wrist forces [Hashemi *et al.*, 2012]. However, model-based methods have a large number of parameters that are difficult to measure directly. Currently, only simple motion estimation with a limited number of joints and degrees of freedom is possible. In contrast to model-based approaches, data-driven methods do not require the measurement of various parameters. Recently, some researchers have begun to use temporal deep learning models to extract motion information from sEMG [Lin *et al.*, 2022; Zhang *et al.*, 2022; Guo *et al.*, 2021]. Lin *et al.* proposed a method based on the BERT structure to predict hand movement from the Root Mean Square (RMS) feature of the sEMG signal [Lin *et al.*, 2022]. Rahimian *et al.* proposed a novel Vision Transformer (ViT)-based neural network architecture to classify and recognize upper-limb hand gestures from sEMG for use in myocontrol of prostheses [Rahimian *et al.*, 2021]. However, these methods have neglected the modeling of short-term dependencies and have not considered the inherent variability in sEMG signals.

3 Preliminaries

3.1 Dataset

We conduct the experiments on Gesture Recognition and Biometrics ElectroMyogram (GRABMyo) Dataset [Pradhan *et al.*, 2022], which is the largest known open-source wrist EMG dataset with 43 subjects and has great potential for developing new generation human-machine interaction based on sEMG.

Data processing. The subjects performed 17 gestures of hand and wrist (including a rest period sEMG) according to the prompts on the computer screen. Each gesture was repeated 7 times, each lasting 5 seconds. In order to improve the convergence speed of the model, we use two methods (Max-Min normalization, μ-law normalization) to normalize the data [Rahimian *et al.*, 2020; Recommendation, 1988]. After normalization, we use a time-sliding window to split samples. In this paper, we set the window size as 200ms, and the overlap of adjacent windows is 10ms. μ-law normalization can logarithmically amplify the outputs of sensors with small magnitudes, which results in better performance than linear normalization.

Definition 1 (sEMG Signal Sequence)**.** *An sEMG signal sequence is defined as a temporal signal sequence sampled by multiple sensors from a human wrist, which can be formulated as* $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, .., \mathbf{x}_t]$*, where* t *is the time window.* $\mathbf{x}_i = [x_{i,1}, x_{i,2}, ...x_{i,c}]$ *represents the signal vector of* c *sen-*

sors, where $x_{i,j}$ is the signal value of the j-th sensor in the i-th time step.

4 Technology Detail

4.1 Model Overview

Figure 2 illustrates the overview of our proposed framework for gesture recognition, which contains three components: (1) The *sEMG Intrinsic Pattern Capture* module encodes the sEMG signal sequence into the hidden sEMG representations. A pre-training model with a segment masking strategy and MSE reconstructing loss is proposed to learn inherent variability from the sEMG signals into the model's parameters. (2) The *Long-term and Short-term Enhanced* module uses two decoupling heads to extract the long-term and short-term context information separately, which improves the sEMG representations in preserving both the global sEMG structure and multiple local signal changes of the sEMG. (3) The *Asymmetric Optimization* strategy addresses the problems of sample biases and imbalance in gesture recognition via an asymmetric classification loss, which can make the model focus on hard and positive samples to improve the recognition.

4.2 sEMG Intrinsic Pattern Capture Module

sEMG Signal Encoding

Given the sEMG signal sequence $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, .., \mathbf{x}_t]$, we first project each signal $\mathbf{x}_i \in \mathbb{R}^c$ into a hidden embedding via a transformation matrix and add each signal embedding with an absolute position embedding. Then, we feed the output sequence into a L-layer Transformer and obtain the output signal embeddings $\mathbf{X}^{(L)} = [\mathbf{x}_1^{(L)}, \mathbf{x}_2^{(L)}, .., \mathbf{x}_t^{(L)}]$, which incorporate temporal context signal information for each position in the sequence.

sEMG Signal Masking

After the sEMG signal-extracting module is constructed, we aim to use pre-training to exploit the intrinsic pattern and temporal semantics disclosed by the unlabeled sEMG signals (labeling sEMG is time-consuming and labor-intensive) and give a good initialization for the model parameters, then avoid the model focusing on some noisy features in the supervised learning task so as to over-fitting on some local minimums. Thus, we propose a sEMG Intrinsic Pattern Capture based on a signal masking strategy.

Specifically, given a transformed signal embedding sequence $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, .., \mathbf{x}_t]$, instead of adding masks on the sequence in terms of time steps like BERT, we add sensor-wise masks for the signal sequence of each sensor similar with [Zerveas *et al.*, 2021], which can encourage the model to learn more fine-grained temporal context dependency on the signal sequence of multiple electrodes. For the signal sequence of the i-th sensor, formulated as $[x_{1,i}, x_{2,i}, ..., x_{t,i}]$, i.e, the i-th column of $\mathbf{X}$, we generate a binary mask vector $\mathbf{m}_i \in \mathbb{R}^t$, where average r radio of elements in $\mathbf{m}_i$ should be 0.15. Randomly generating $\mathbf{m}_i$ may cause a lot of isolated-masked signals, meaning one masked signal whose adjacent signals are unmasked. However, a single signal can be easily predicted by its immediately preceding or succeeding signals, making self-supervised learning easy to fitting on ineffective patterns and poor for learning temporal semantic information. In consideration of this, we introduce a more complex masking strategy that aims to generate multiple masked segments on the sequence with an average length l_m, which means m_i is composed of contiguous masked segments and unmasked segments. The length of masked segments follows a geometric distribution with mean l_m, and the length of unmasked segments follows a geometric distribution with mean l_u. Also, the $\frac{l_m}{l_u} = \frac{r}{1-r}$ so that the number of masked elements would follow the proportion r. The pseudocode of the masking algorithm is presented in Algorithm 1.

Algorithm 1: The Algorithm of sEMG Signal Masking

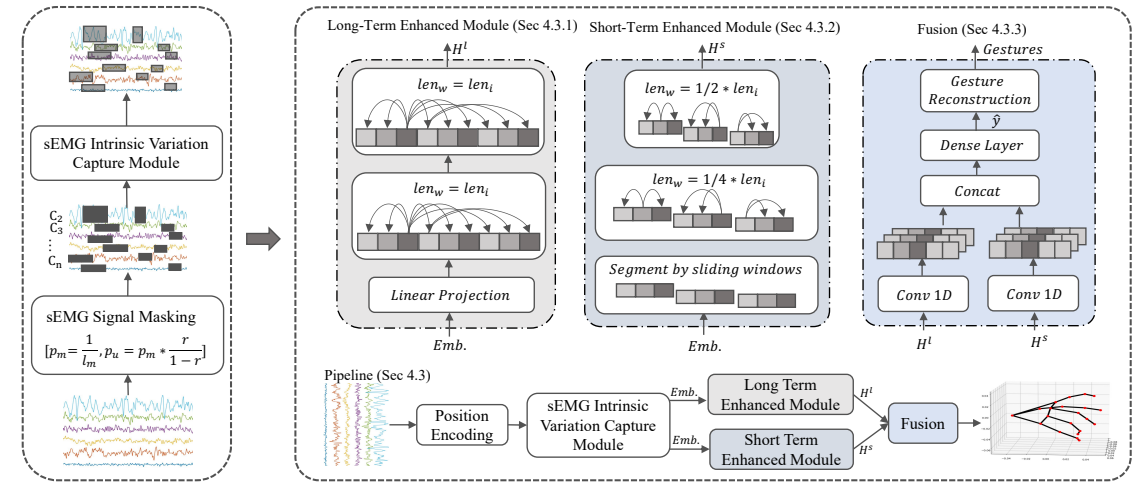
Input: The length of the input signal sequence t,
The number of signal sensors c,
The average length of masked segments l_m,
The masked radio r.
Output: The mask matrix $\mathbf{M}$.

```
for i = 1, ..., c do
    m_i = [True] * h;
    p_m = 1/l_m; // probability of each masking segment stopping.
    p_u = p_m * r/(1 - r); // probability of each unmasked segment stopping.
    p = [p_m, p_u];
    state = Bool(random(0, 1) > r); // the first state.
    for j = 1, ..., t do
        m_{i,j} = state;
        if random(0, 1) < p[state] then
            state = ¬state;
Return M = ([m_i]_{i=0}^c)^T;
```

Then, we can mask the input sEMG signal sequence $\mathbf{X}$ by $\widehat{\mathbf{X}} = \mathbf{X} \odot \mathbf{M}$, where $\odot$ is elementwise multiplication and $\widehat{\mathbf{X}}$ is the masked input. With the proposed Transformer-based sEMG signal encoder, we can obtain the output $\widehat{\mathbf{X}}^{(L)} = [\widehat{\mathbf{x}}_1^{(L)}, \widehat{\mathbf{x}}_2^{(L)}, ..., \widehat{\mathbf{x}}_t^{(L)}]$. For self-supervised learning, we add a linear layer on the top of masked output to reconstruct each sEMG signal $\widehat{\mathbf{x}}_i^{(L)}$ as $\widetilde{\mathbf{x}}_i \in \mathbb{R}^c$, which is the reconstructed sEMG signal in the i-th time step generated from the masked input. Then, we minimize the Mean Squared Error (MSE) of the reconstructed signals and original signals on the masked positions for each sample:

$$\min \frac{1}{|\mathbf{M}|} \sum_{i=0}^{t} \sum_{j=0}^{c} \mathbb{1}(\mathbf{M}_{i,j} = 0)(\widetilde{x}_{i,j} - x_{i,j})^2, \qquad (1)$$

where $\mathbb{1}(\cdot)$ is the indicator function, $\widetilde{x}_{i,j}$ and $x_{i,j}$ are the reconstructed value and original value of j-th sensor in $\widetilde{\mathbf{x}}_i$ and $\mathbf{x}_i$ respectively, and $\mathbf{M}_{i,j}$ is the element in the i-th row and



(a) Self-Supervised Pre-training based on sEMG Signal Masking

(b) Short-Term Enhanced Transformer

Figure 2: Overview of STET. The sEMG signal is encoded using the sEMG Intrinsic Pattern Capture module, which is first pre-trained via sEMG signal Masking. A long-term and short-term enhanced module improves sEMG representations. An asymmetric optimization strategy addresses biases and imbalances in gesture recognition through an asymmetric classification loss.

the j-th column of $\mathbf{M}$. Thus, we can pre-train the sEMG Intrinsic Pattern Capture via the above strategy to obtain well-initialized model parameters for the downstream task. In practice, we empirically set the masking proportion r as 0.15 and the average length of masked segments as 3. The illustration of the pre-training procedure is in Figure 2 (a).

4.3 Long-term and Short-term Decoding

Then, based on the pre-trained sEMG Intrinsic Pattern Capture, we develop two decoder heads to further extract the long-term and short-term dependency on the signal sequences, respectively. Intuitively, both the long-term and short-term information on signal sequences are significant in the gesture recognition problem. Long-term information refers to the global context of an sEMG sequence, which provides the overall structure of a signal to help the interpretation of the gesture. Short-term information refers to the movement signal in a short time interval of the whole sequence, which can provide specific local characteristics for accurate recognition when the overall structures of sEMGs are ambiguous. For example, distinguishing between Index Finger Extension (IFE) and Middle Extension (ME) movements requires a closer examination of the local signal changes in sEMG, whereas differentiating gestures with large variations, such as hand gestures and wrist gestures, necessitates a focus on the global sEMG information.

Preserving Long-term sEMG Signal

Given the hidden output $\mathbf{X}^{(L)} = [\mathbf{x}_1^{(L)}, \mathbf{x}_2^{(L)}, ..., \mathbf{x}_t^{(L)}]$ of a sEMG signal sequence, we first build a long-term decoder to extract the long-term dependency on the complete output. Specifically, the long-term decoder is defined as a multi-head self-attention layer:

$$MultiHead\_L\left(\mathbf{X}^{(L)}\right) = Concat\left(h_1, \ldots, h_d\right)\mathbf{W}^O, \quad (2)$$

$$\{h_i\}_{i=0}^{d} = \{Attention\left(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i\right)\}_{i=0}^{d}, \quad (3)$$

$$Attention\left(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i\right) = Softmax\left(\frac{\mathbf{Q}_i\mathbf{K}_i^T}{\sqrt{h}}\right)\mathbf{V}_i, \quad (4)$$

$$where\ \mathbf{Q}_i = \mathbf{X}^{(L)}\mathbf{W}_i^Q,\ \mathbf{K}_i = \mathbf{X}^{(L)}\mathbf{W}_i^K,\ \mathbf{V}_i = \mathbf{X}^{(L)}\mathbf{W}_i^V, \quad (5)$$

where $\{\mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V\}_{i=0}^{d} \in \mathbb{R}^{h\times h}$ are parameter matrices and d is the number of attention heads. $Concat(\cdot)$ represents the concatenate operation. $\mathbf{W}^O \in \mathbb{R}^{dh\times h}$ is the output parameter matrix to transform the concatenated outputs of d attention heads. Then, the long-term sEMG embeddings $\mathbf{H}^l \in \mathbb{R}^{t\times h}$ is obtained by $\mathbf{H}^l = MultiHead\_L(\mathbf{X}^{(L)})$. Through the self-attention layer, the global context signal information is collected to the embeddings of t timesteps with different attention weights.

Preserving Short-term sEMG Signal

To model the local context information within a short time interval, we introduce a slide-window self-attention layer to extract the short-term dependency on the signal outputs. Similarly, we stack multiple attention heads and calculate the attention of context signals to weighted sum them up into the final representations. The difference is that, for each time step, we only calculate the attention of its nearest w context. Specifically, we can rewrite the $\mathrm{Attention}(\cdot)$ in Eq.4 as:

$$Attention\_S(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \left[Softmax\left(\frac{\mathbf{Q}_i\mathbf{K}_i^{wT}}{\sqrt{h}}\right)\mathbf{V}_i^w\right]_{i=1}^{t}, \quad (6)$$

$$[\mathbf{K}_i^w]_{i=1}^t = Unfold(\mathbf{K}, w),\ [\mathbf{V}_i^w]_{i=1}^t = Unfold(\mathbf{V}, w), \quad (7)$$

where w is the sliding windows size, $\mathbf{Q}_i \in \mathbb{R}^h$ as the i-th query is the i-th row of $\mathbf{Q}$, $\mathbf{K}_i^w \in \mathbb{R}^{w\times h}$ and $\mathbf{V}_i^w \in \mathbb{R}^{w\times h}$ are the keys and values in a window around the i-th query. For the key and value matrix, we utilize the $Unfold(\cdot)$ operation

to generate the slide windows for each timestep. Noted that to avoid confusion, we omit the index of attention head in the above equation.

Thus, by stacking multiple sets of parameters in $Attention\_S(\cdot)$ to constitute different attention heads, we can obtain the short-term sEMG embeddings $\mathbf{H}^s \in \mathbb{R}^{t \times h}$ by $\mathbf{H}^s = MultiHead\_S(\mathbf{X}^{(L)})$. Using slide windows, each row in $\mathbf{H}^s$ preserves the local context signal information of the corresponding timestep, representing the movement from the past $w/2$ timesteps to the next $w/2$ timesteps.

Unlike LST-EMG-Net [Zhang *et al.*, 2023] and Focal Transformer [Yang *et al.*, 2021], which process Long Term Feature and Short Term Feature sequentially, we handle them in parallel. This ensures that features processed earlier in a sequential manner are not overlooked.

Fusion

Obtained the long-term embeddings $\mathbf{H}^l$ and the short-term embeddings $\mathbf{H}^s$ of an sEMG signal sequence, we first concatenate them in terms of the hidden dimension, then introduce a 1-D convolution to summarize the t-step sEMG embedding sequence into the final sEMG representation, which is fed into a *Feed Forward Layer* with a *Sigmoid Layer* to obtain the final classification probability of which gesture the sEMG belonging to, which can be written as: $\widehat{\mathbf{y}} = \sigma(FC(\mathbf{u}^T \cdot [\mathbf{H}^l : \mathbf{H}^s]))$, where $FC(\cdot)$ is a two-layer fully connected layer, $\sigma(\cdot)$ is the activation function, and $\widehat{\mathbf{y}} \in \mathbb{R}^C$ is the output classification probability of the sEMG signals.

4.4 Asymmetric Optimization

As the common use in multi-label classification, we reduce the gesture recognition problem into a series of binary classification tasks. However, we consider two tricky problems that exist in the above model optimization. (1) As the sampled signals are usually unstable over time, making the samples of the sEMG signal sequence may be critically biased. Some samples with strong signals are easily predicted, while many samples with fuzzy signals are hard to predict. (2) As we set 17 classes for the sEMG signals, and each class contains a comparable number of samples. Thus, each class contains, on average, many more negative samples than positive ones. This imbalance may make the model eliminate the gradients from the positive samples in the optimization process, resulting in poor accuracy. Realizing this, we introduce the Asymmetric loss [Ridnik *et al.*, 2021] for the gesture classification task. Asymmetric loss is a variant of Focal loss. (1) It uses focusing parameters to reduce the contribution of easily predicted samples and make the model optimization focus on hard samples; (2) It further introduces asymmetric focusing parameters and asymmetric probability shifting to down-weight the contribution from massive easy negatives and emphasizes the contribution of positive samples. Thus, we define the loss function as follows:

$$\mathcal{L}_{STET} = -\sum_{i=1}^{N}\sum_{j=1}^{C}\Big(y_{i,j}\left(1-\widehat{y}_{i,j}\right)^{\gamma^+}\log\left(\widehat{y}_{i,j}\right) + \left(1-y_{i,j}\right)\left(\widehat{y}^m_{i,j}\right)^{\gamma^-}\log\left(1-\widehat{y}^m_{i,j}\right)\Big), \tag{8}$$

$$\widehat{y}^m_{i,j} = max\left(\widehat{y}_{i,j} - m, 0\right), \tag{9}$$

where $y_{i,j}$ and $\widehat{y}_{i,j}$ is the ground-truth and probability of the i-th sEMG signal sequence belonging to the j-th gesture. $\left(1-\widehat{y}_{i,j}\right)^{\gamma^+}$ and $\left(\widehat{y}^m_{i,j}\right)^{\gamma^-}$ are two terms to make the weights of hard predicted samples bigger than those easily predicted samples, γ^+, and γ^- are two focusing parameters and $\gamma^+ > \gamma^-$ lead to asymmetric focusing that help the optimization pay more focus on positive samples of each class. $\widehat{y}^m_{i,j}$ is the shifted probability and m is shifting margin. The probability shifting for negative samples encourages the optimizer to further reduces their contribution.

5 Experiment

5.1 Settings

Implementation Details STET is implemented in PyTorch [Paszke *et al.*, 2019] and is trained using one RTX 3090 GPU. During training, we use the RAdam [Liu *et al.*, 2019], which is a theoretically sound variant of the Adam optimizer with a weight decay of 1e-3. We pre-train on GRABMyo for 20 epochs using a fixed learning rate of 1e-4 for the backbone. In the decoder, we use two layers of full attention in the long-term decoder and two layers of sliding window attention in the short-term decoder. The short-term decoder's window size is 41 and 21, and the window's move step is set to 1. In both the pre-training and fine-tuning periods, we set the batch size to 16. To avoid overfitting, we set drop out to 0.2.

Evaluation Metrics

Following the prior works[Guo *et al.*, 2021; Wang *et al.*, 2020; Chen *et al.*, 2021; Rahimian *et al.*, 2021], we choose the below metrics to evaluate the model's performance. **Pearson Correlation Coefficient** (CC) is a widely used measure of the linear relationship between two variables. It ranges from -1 to 1, where a larger CC value indicates greater similarity between the predicted and estimated joint angles curve, indicating improved estimation. **Root Mean Square Error** (RMSE) is a common metric for evaluating the deviation between predicted and observed values. As the range of fluctuations in the curves of different joint angles can vary significantly, it is difficult to evaluate the performance of models using RMSE alone fairly. Normalization of RMSE addresses this issue, resulting in the Normalized RMSE (NRMSE). **Average curvature** (κ) of all points for each joint is used to measure the smoothness of an estimated curve. A smaller κ indicates a smoother curve.

5.2 Comparison with Baselines

We compare the accuracy (ACC) and Standard deviation (STD) between our proposed STET and previous popular sEMG-based gesture recognition methods. Specifically, we train the model on the GRABMyo dataset [Pradhan *et al.*, 2022] (the detail of data processing shown in section 3.1) and separately report the classification results on the categories of Single-finger gestures, Multi-finger gestures, Wrist gestures, Rest, and the overall results. The ratio of the training set to the test set for each gesture is 5 to 2.

From the experimental results, we can observe that STET consistently performs best on four categories of gestures and

Model	Single-finger		Multi-finger		Wrist		Rest		Overall	
	ACC	STD	ACC	STD	ACC	STD	ACC	STD	ACC	STD
Asif *et al.* [Asif *et al.*, 2020]	83.44%	0.015	83.58%	0.013	89.40%	0.009	90.86%	0.012	85.34%	0.014
TCN [Tsinganos *et al.*, 2019]	78.78%	0.017	79.10%	0.018	87.27%	0.011	88.57%	0.017	81.50%	0.016
GRU [Chen *et al.*, 2021]	84.45%	0.015	84.88%	0.013	90.06%	0.009	89.42%	0.019	86.30%	0.015
TEMGNet [Rahimian *et al.*, 2021]	77.70%	0.019	74.00%	0.029	84.04%	0.017	87.46%	0.014	78.02%	0.019
Zerveas *et al.* [Zerveas *et al.*, 2021]	78.45%	0.016	77.20%	0.020	87.28%	0.016	86.76%	0.017	80.43%	0.018
Informer [Zhou *et al.*, 2021]	86.88%	0.016	86.54%	0.017	91.90%	0.011	83.56%	0.024	87.71%	0.016
LST-EMG-Net [Zhang *et al.*, 2023]	87.21%	0.011	83.16%	0.012	88.36%	0.018	82.52%	0.021	85.31%	0.015
TEMGNET+STEM(ours)	84.57%	0.017	81.23%	0.022	88.12%	0.017	88.74%	0.013	84.07%	0.017
Informer+STEM(ours)	87.42%	0.015	88.39%	0.018	92.07%	0.013	90.33%	0.011	89.14%	0.015
STET	**88.27%**	0.014	**89.93%**	0.015	**93.77%**	0.010	**95.33%**	0.012	**90.76%**	0.012

Table 1: The gesture classification performance on the Single-finger, Multi-finger, Wrist, Rest, and Overall categories.

overall data. In particular, STET and [Zerveas *et al.*, 2021] both use Transformer-based encoders. While in the decoder part, STET introduces both the short-term and long-term decoder rather than the fully connected layers used in [Zerveas *et al.*, 2021]. As a result, the overall accuracy of STET is improved from 80.43% to 90.76% compared to [Zerveas *et al.*, 2021]. This is because the proposed long-term and short-term decoupling module can extract both the global and fine-grained dependency on the signal dependency and thus can learn better sEMG representations.

Among the transformer-based methods, Informer and STET performed best, with accuracy rates of 87.71% and 90.76%, respectively. Informer relies heavily on max pooling layers to aggregate features, leading to the relative weakness in extracting some short-term features. STET enhances accuracy and stability by strengthening the short-term feature extraction. The improvement is remarkable on the Rest gestures, where the accuracy improves from 83.56% to 95.33%. Furthermore, after incorporating our designed short-term encoder into Informer, its accuracy rate increased from 87.71% to 89.14%, and the classification accuracy for Rest gestures improved from 83.56% to 90.33%.

Note that the Rest category of gesture, as the resting state of devices such as interactive bracelets, is the most frequent gesture that appears in the signals. Thus, the stability of its prediction plays a significant role in the problem. STET achieves the highest and most stable results on the prediction of the Rest category compared with all baselines, indicating the robustness of STET.

5.3 Ablation Studies

To validate the effects of the unsupervised sEMG Intrinsic Pattern Capture (EIPC), Long-term decoder, Short-term decoder, Fuse strategy, and loss function. We designed variants of STET and reported their results in Table 2.

First, we can observe that, with the unsupervised EIPC, the accuracy of the transformer and STET is improved by 0.60% and 1.12% compared with training from scratch. This suggests that unsupervised EIPC can aid in discerning additional data features, such as the inherent variability in sEMG, without the requirement for new samples or extra annotations. Significantly, this process circumvents the need for external data, thus preserving user privacy in the context of data acquisition and processing. Replacing the fully connected layer

Transformer	EIPC	LT	ST	Fusing	CEL	ASL	ACC
✓					✓		85.73%
✓	✓				✓		86.33%
✓	✓	✓			✓		88.02%
✓	✓		✓		✓		87.72%
✓	✓	✓	✓	✓	✓		89.37%
✓		✓	✓	✓		✓	89.42%
✓	✓	✓	✓	✓		✓	90.54%

Table 2: The results of ablation study. EIPC: sEMG Intrinsic Pattern Capture; LT: Long-Term decoder; ST: Short-Term decoder; CEL: Cross Entropy Loss; ASL: ASymmetric Loss.

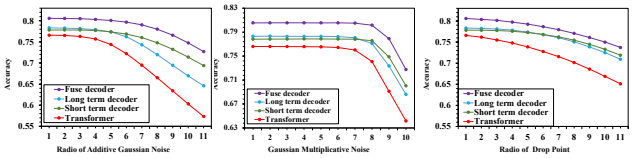


Figure 3: Accuracy versus Noise Intensity Curve.

of the transformer's decoder with the long-term decoder or short-term decoder, the performance is improved by 1.16% and 1.39%, respectively. Furthermore, the performance is comparable when using the long-term decoder or short-term decoder alone, indicating that the two kinds of features may play different significant roles in the sEMG signal recognition, and the short-term cannot be ignored. Most importantly, when we used our designed fuse module to combine long-term and short-term features, the accuracy further improved by 2.46%. This suggests that the two decoders are complementary and that enhancing short-term features is necessary on the basis of the long-term decoder. We employed Asymmetric Loss (ASL) to drive the model's focus on difficult samples. Compared to simply using the Cross-Entropy Loss (CEL), the accuracy improved by 0.73%, indicating the effectiveness of ASL.

Backbone	In STET framework	AG noise	MG noise	Signal loss
Transformer	No	25%	16%	14%
Transformer	Yes	10%	10%	8%
Informer	No	11%	9%	26%
Informer	Yes	9%	8%	17%

Table 3: Drop rates of accuracy calculated by $drop\_rate = \frac{ACC_{raw} - ACC_{noise}}{ACC_{raw}}$. AG: Additive Gaussian noise, MG: Multiplicative Gaussian noise

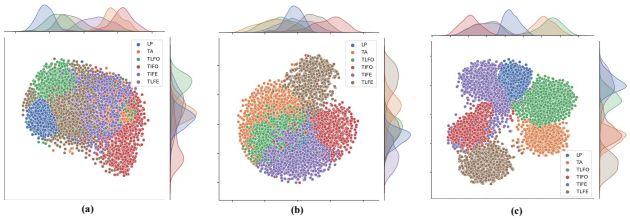


Figure 4: Visualization of (a) the long-term sEMG embeddings, (b) the short-term sEMG embeddings, and (c) the fused sEMG embeddings for gesture recognition. Note that we color each sample by its classes.

5.4 Robustness Analysis

To verify the robustness of the model, we only used high-quality data collected in the lab to train the model and added different types of noise (Additive Gaussian noise, Multiplicative Gaussian noise, and signal loss) during validation to simulate complex scenarios that might be encountered in real situations.

Additive noise typically refers to thermal noise, which is added to the original signal. This type of noise exists regardless of the presence of the original signal and is often considered the background noise of the system in sEMG acquisition. Multiplicative noise is generally caused by channel instability, and it has a multiplicative relationship with the original signal. Also, we simulated signal loss during transmission by randomly setting a portion of the signals to zero.

Figure 3 illustrates the influence exerted by three distinctive noise categories, namely additive noise, multiplicative noise, and signal loss, on the accuracy of the proposed model. The model using only the short-term decoder is less affected by noise compared to the long-term version. This relative robustness of the short-term decoder is potentially attributable to its unique capability to mitigate the global impact of noise by virtue of a sliding window multiple-sampling scheme, which effectively confines the sphere of noise impact. The model that integrates both long-term and short-term characteristics persistently outperforms models that rely on only one. This highlights the significant effectiveness of the integrating process in dealing with noise-induced interference. As depicted in Table 3, it is evident that both Transformer and Informer models demonstrate a notable enhancement in noise resistance when their decoders are replaced with the design from STET.

5.5 Visualizations

To demonstrate the distinction, we first obtained STET's long-term, short-term, and fuse embeddings. The embeddings with dimensions (N, T, H) were then flattened to $(N, T * H)$ and separately projected in 2D by t-SNE, the result shown in Figure 4. We colored each node by category for the illustration. As shown in Figure 4, The classification boundary generated by the long-term feature and the short-term feature is a significant difference, indicating that the long-term and short-term features are capable of recognizing different types of gestures. This further suggests that the two features are complementary in data representation. For example, short-term embedding can distinguish TA gesture and TIFO gesture very well, but TA gesture and TIFO gesture will be confused in long-term embedding. Meanwhile, long-term embedding can distinguish LP gesture and TIFE gesture very well, but short-term embedding will confuse them. As shown in Figure 4(c), after the fusion of the two types of features, the classification interface is wider, and the confusion points are significantly reduced, which indicates that the fusion module can effectively complement the strengths of the two types of features.

Model	PCC	NRMSE	κ	Time Cost/epoch(s)
LSTM	0.779	0.096	0.581	26.36
TCN	0.833	0.088	1.533	**3.62**
BERT	0.867	0.077	1.571	4.95
sBERT-OHME	0.869	0.076	0.532	4.96
STET	**0.877**	**0.073**	**0.522**	6.83

Table 4: Comparison of STET with Other models in Predicting Joint Angle for Fingers.

5.6 Regression: Hand Joint Angles Prediction

STET can conveniently handle regression tasks by changing the loss function to mean squared error (MSE) loss. Continuous motion estimation extracts continuous motion information, such as joint angles and torques, from sEMG signals. Since continuous motion estimation requires outputting subtle variations of the movement at each time instant, the local signal variations are particularly important for this type of estimation. In this section, we have re-selected the most competitive models known for sEMG-based joint angle prediction as the baseline and tested the performance of STET on the regression task of predicting the main 10 joint angles for fingers using the Ninapro DB2 [Atzori *et al.*, 2014] dataset. As shown in Table 4, STET achieved the best performance in PCC, NRMSE, and κ, indicating that the joint angle curve predicted by STET is more in line with the real curve and has less abnormal fluctuations, which will significantly improve the user's interactive experience. In terms of training time, due to the addition of the short-term decoder, its training speed is slightly slower than BERT but still within an acceptable range.

6 Conclusion

Current sEMG-based gesture recognition models usually fail to handle various noisy and distinguish similar gestures, especially in non-laboratory settings. In this paper, we found using short-term information and self-supervised EIPC mitigates this issue. Therefore, we proposed STEM for capturing local signal changes and enhancing noise resistance. The STEM is easily deployable and serves as a plug-in that can potentially be applied to most time series deep learning models. According to our experimental results, our method significantly improved performance for both classification and regression tasks in sEMG, and the model's ability to resist signal loss, Gaussian additive noise, and Gaussian multiplicative noise was clearly improved. This will further drive the practical application of sEMG in VR, AR, and other human-computer interaction scenarios.

References

[Asif *et al.*, 2020] Ali Raza Asif, Asim Waris, Syed Omer Gilani, Mohsin Jamil, Hassan Ashraf, Muhammad Shafique, and Imran Khan Niazi. Performance evaluation of convolutional neural network for hand gesture recognition using emg. *Sensors*, 20(6):1642, 2020.

[Atzori *et al.*, 2014] Manfredo Atzori, Arjan Gijsberts, Claudio Castellini, Barbara Caputo, Anne-Gabrielle Mittaz Hager, Simone Elsig, Giorgio Giatsidis, Franco Bassetto, and Henning Müller. Electromyography data for non-invasive naturally-controlled robotic hand prostheses. *Scientific data*, 1(1):1–13, 2014.

[Becker *et al.*, 2018] Vincent Becker, Pietro Oldrati, Liliana Barrios, and Gábor Sörös. Touchsense: classifying finger touches and measuring their force with an electromyography armband. In *Proceedings of the 2018 ACM International Symposium on Wearable Computers*, pages 1–8, 2018.

[Bi *et al.*, 2019] Luzheng Bi, A. Feleke, and Cuntai Guan. A review on emg-based motor intention prediction of continuous human upper limb motion for human-robot collaboration. *Biomedical Signal Processing and Control*, 51:113–127, 2019.

[Borbély and Szolgay, 2017] Bence J Borbély and Péter Szolgay. Real-time inverse kinematics for the upper limb: a model-based algorithm using segment orientations. *Biomedical engineering online*, 16(1):1–29, 2017.

[Chen *et al.*, 2021] Rui Chen, YuanZhi Chen, Weiyu Guo, Chao Chen, Zheng Wang, and Yongkui Yang. semg-based gesture recognition using gru with strong robustness against forearm posture. In *2021 IEEE International Conference on Real-time Computing and Robotics (RCAR)*, pages 275–280. IEEE, 2021.

[Clancy *et al.*, 2012] Edward A. Clancy, Lukai Liu, Pu Liu, and Daniel V.Zandt Moyer. Identification of constant-posture emg-torque relationship about the elbow using nonlinear dynamic models. *IEEE Transactions on Biomedical Engineering*, 59:205–212, 2012.

[Du *et al.*, 2017] Yu Du, Yongkang Wong, Wenguang Jin, Wentao Wei, Yu Hu, Mohan S Kankanhalli, and Weidong Geng. Semi-supervised learning for surface emg-based gesture recognition. In *IJCAI*, pages 1624–1630, 2017.

[Guo *et al.*, 2021] Weiyu Guo, Chenfei Ma, Zheng Wang, Hang Zhang, Dario Farina, Ning Jiang, and Chuang Lin. Long exposure convolutional memory network for accurate estimation of finger kinematics from surface electromyographic signals. *Journal of Neural Engineering*, 18, 2021.

[Hashemi *et al.*, 2012] Javad Hashemi, Evelyn Morin, Parvin Mousavi, Katherine Mountjoy, and Keyvan Hashtrudi-Zaad. Emg–force modeling using parallel cascade identification. *Journal of Electromyography and Kinesiology*, 22:469–477, 6 2012.

[Koike and Kawato, 1995] Yasuharu Koike and Mitsuo Kawato. Estimation of dynamic joint torques and trajectory formation from surface electromyography signals using a neural network model. *Biological cybernetics*, 73(4):291–300, 1995.

[Koirala *et al.*, 2015] Kishor Koirala, Meera Dasog, Pu Liu, and Edward A. Clancy. Using the electromyogram to anticipate torques about the elbow. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 23(3):396–402, 2015.

[Li *et al.*, 2021] Wei Li, Ping Shi, and Hongliu Yu. Gesture recognition using surface electromyography and deep learning for prostheses hand: state-of-the-art, challenges, and future. *Frontiers in neuroscience*, 15:621885, 2021.

[Lin *et al.*, 2022] Chuang Lin, Xingjian Chen, Weiyu Guo, Ning Jiang, Dario Farina, and Jingyong Su. A bert based method for continuous estimation of cross-subject hand kinematics from surface electromyographic signals. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2022.

[Liu *et al.*, 2015] Pu Liu, Lukai Liu, and Edward A. Clancy. Influence of joint angle on emg-torque model during constant-posture, torque-varying contractions. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 23(6):1039–1046, 2015.

[Liu *et al.*, 2019] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*, 2019.

[Liu *et al.*, 2020] Yilin Liu, Fengyang Jiang, and Mahanth Gowda. Finger gesture tracking for interactive applications: A pilot study with sign languages. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(3):1–21, 2020.

[Liu *et al.*, 2021a] Yang Liu, Chengdong Lin, and Zhenjiang Li. Wr-hand: Wearable armband can track user's hand. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(3):1–27, 2021.

[Liu *et al.*, 2021b] Yilin Liu, Shijia Zhang, and Mahanth Gowda. Neuropose: 3d hand pose tracking using emg wearables. In *Proceedings of the Web Conference 2021*, pages 1471–1482, 2021.

[Paszke *et al.*, 2019] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[Pradhan *et al.*, 2022] Ashirbad Pradhan, Jiayuan He, and Ning Jiang. Multi-day dataset of forearm and wrist electromyogram for hand gesture recognition and biometrics. *Scientific Data*, 9(1):1–10, 2022.

[Rahimian *et al.*, 2020] Elahe Rahimian, Soheil Zabihi, Seyed Farokh Atashzar, Amir Asif, and Arash Mohammadi. Xceptiontime: Independent time-window xceptiontime architecture for hand gesture classification. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1304–1308. IEEE, 2020.

[Rahimian *et al.*, 2021] Elahe Rahimian, Soheil Zabihi, Amir Asif, Dario Farina, S Farokh Atashzar, and Arash Mohammadi. Temgnet: Deep transformer-based decoding of upperlimb semg for hand gestures recognition. *arXiv preprint arXiv:2109.12379*, 2021.

[Recommendation, 1988] CCITT Recommendation. Pulse code modulation (pcm) of voice frequencies. In *ITU*, 1988.

[Ridnik *et al.*, 2021] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 82–91, 2021.

[Sun *et al.*, 2020] Ying Sun, Chao Xu, Gongfa Li, Wanfen Xu, Jianyi Kong, Du Jiang, Bo Tao, and Disi Chen. Intelligent human computer interaction based on non redundant emg signal. *Alexandria Engineering Journal*, 59(3):1149–1157, 2020.

[Tsinganos *et al.*, 2019] Panagiotis Tsinganos, Bruno Cornelis, Jan Cornelis, Bart Jansen, and Athanassios Skodras. Improved gesture recognition based on semg signals and tcn. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1169–1173, 2019.

[Wang and Buchanan, 2002] Lin Wang and T.S. Buchanan. Prediction of joint moments using a neural network model of muscle activations from emg signals. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 10(1):30–37, 2002.

[Wang *et al.*, 2020] Chao Wang, Weiyu Guo, Hang Zhang, Linlin Guo, Changcheng Huang, and Chuang Lin. semg-based continuous estimation of grasp movements by long-short term memory network. *Biomedical Signal Processing and Control*, 59:101774, 2020.

[Xiong *et al.*, 2021] Dezhen Xiong, Daohui Zhang, Xingang Zhao, and Yiwen Zhao. Deep learning for emg-based human-machine interaction: A review. *IEEE/CAA Journal of Automatica Sinica*, 8(3):512–533, 2021.

[Yang *et al.*, 2021] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal self-attention for local-global interactions in vision transformers. *arXiv preprint arXiv:2107.00641*, 2021.

[Yao *et al.*, 2018] Shaowei Yao, Yu Zhuang, Zhijun Li, and Rong Song. Adaptive admittance control for an ankle exoskeleton using an emg-driven musculoskeletal model. *Frontiers in neurorobotics*, 12:16, 2018.

[Zerveas *et al.*, 2021] George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2114–2124, 2021.

[Zhang *et al.*, 2022] Qian Zhang, JiaZhen Jing, Dong Wang, and Run Zhao. Wearsign: Pushing the limit of sign language translation using inertial and emg wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(1):1–27, 2022.

[Zhang *et al.*, 2023] Wenli Zhang, Tingsong Zhao, Jianyi Zhang, and Yufei Wang. Lst-emg-net: Long short-term transformer feature fusion network for semg gesture recognition. *Frontiers in Neurorobotics*, 17:1127338, 2023.

[Zhao *et al.*, 2020] Yihui Zhao, Zhiqiang Zhang, Zhenhong Li, Zhixin Yang, Abbas A Dehghani-Sanij, and Shengquan Xie. An emg-driven musculoskeletal model for estimating continuous wrist motion. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(12):3113–3120, 2020.

[Zhou *et al.*, 2021] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11106–11115, 2021.