This article has been accepted for publication in IEEE Journal of Selected Topics in Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2024.3402340

ViT-MDHGR: Cross-day Reliability and Agility in Dynamic Hand Gesture Prediction via HD-sEMG Signal Decoding

Qin Hu†, *Student Member, IEEE*, Golara Ahmadi Azar†, Alyson Fletcher, *Member, IEEE*, Sundeep Rangan, *Fellow, IEEE*, S. Farokh Atashzar*, *Senior member, IEEE*

***Abstract*— Surface electromyography (sEMG) and high-density sEMG (HD-sEMG) biosignals have been extensively investigated for myoelectric control of prosthetic devices, neurorobotics, and more recently human-computer interfaces because of their capability for hand gesture recognition/prediction in a wearable and non-invasive manner. High intraday (same-day) performance has been reported. However, the interday performance (separating training and testing days) is substantially degraded due to the poor generalizability of conventional approaches over time, hindering the application of such techniques in real-life practices. There are limited recent studies on the feasibility of multi-day hand gesture recognition. The existing studies face a major challenge: the need for long sEMG epochs makes the corresponding neural interfaces impractical due to the induced delay in myoelectric control. This paper proposes a compact ViT-based network for multi-day dynamic hand gesture prediction. We tackle the main challenge as the proposed model only relies on very short HD-sEMG signal windows (i.e., 50 ms, accounting for only one-sixth of the convention for real-time myoelectric implementation), boosting agility and responsiveness. Our proposed model can predict 11 dynamic gestures for 20 subjects with an average accuracy of over 71% on the testing day, 3-25 days after training. Moreover, when calibrated on just a small portion of data from the testing day, the proposed model can achieve over 92% accuracy by retraining less than 10% of the parameters for computational efficiency.**

***Index Terms*— Human-robot Interactions, Surface Electromyography (sEMG), Vision Transformer, Hand Gesture Recognition (HGR), Cross-day HGR, Minimal Calibration**

I. Introduction

The initiation of voluntary human motion starts from the central nervous system, which sends synaptic inputs to the motor neurons in the spinal cord [1]. These inputs are then transduced into forces and transmitted through action potentials by motor units that comprise muscle fibers. Surface electromyography (sEMG) is the cumulative sum of motor unit action potentials [2]. It is measured and recorded from the skin surface during muscle activities. The early research of sEMG has the main interest in clinical diagnosis and biomedical applications, such as rehabilitation and assistive technologies (e.g., prosthetic control) and ergonomics [3], [4]. Nowadays, sEMG, which is the control signal in human-computer interaction, has a wide range of applications, such as in augmented reality and virtual reality (for hands-free control to counterbalance network delay) [5] and sports science (for performance measurement and optimization) [6]. The increasing research interest in sEMG is due to its non-invasiveness, wearability, and potential for real-time control. Over the past two decades, high-density sEMG (HD-sEMG) has enhanced the performance of these applications by capturing signals with high spatial resolution information on muscle activities.

sEMG-based hand gesture recognition (HGR) is achieved through pattern recognition (PR), which mainly includes two progressive steps: feature extraction and classification. Machine learning (ML) and deep learning (DL) are the two basic processing methods in sEMG PR [7]. Classical ML relies on researchers' expertise and feature engineering for feature extraction and models such as Support Vector Machine (SVM) for classification [8]. DL methods such as Convolutional Neural Networks (CNNs) automatically extract various temporal, spectral, and spatial features from sEMG through their hierarchical architectures [9]. Conventionally, these methods trained and tested on sEMG signals collected on the same day only provide short-term performance but underestimate long-term practicality due to the disregard for the effect of sensor misplacement, sensor displacement, and day-to-day variation in human neurophysiology and skin conductivity [10]. The performance of a previously trained model can drop by 55% if used on another day [11]. To enhance the scalability and feasibility of such wearable neural interfaces for human-robot interaction and control, temporal generalizability should be addressed by securing high interday reliability.

Researchers have started investigating multi-day HGR based on sEMG to improve the longitudinal performance of these control systems within the last five years. These works can

This material is based upon work supported by the US National Science Foundation under Grant No #2229697, #2208189, #2121391.

Hu, Rangan, and Atashzar are with the Department of Electrical and Computer Engineering, New York University (NYU), New York, NY, 11201 USA. Rangan is also the Director of NYU WIRELESS. Atashzar is also with the Department of Mechanical and Aerospace Engineering, Biomedical Engineering, NYU WIRELESS, and NYU Center for Urban Science and Progress (CUSP).

Azar and Fletcher are with the Department of Electrical and Computer Engineering, University of California Los Angeles (UCLA), Los Angeles, CA, 90095 USA. Fletcher is also with the Department of Statistics, Mathematics, and Computer Science, UCLA.

† Hu and Azar share the first authorship.

* Corresponding author: f.atashzar@nyu.edu.

also be categorized into traditional ML and DL methods. Traditional ML relies on training on the extracted features from multiple days to capture sEMG variations across days and secure high interday performance without calibration. Adaptive ML approaches such as Adaptive LDA (ALDA) can adapt to new data from the testing day by updating its parameters to improve interday performance. However, at root, the main hurdles of these traditional ML methods are (1) the simplicity of the extracted features that are incompetent to capture cross-day characteristics in sEMG and generalize the use of multi-day HGR systems over a large number of subjects and (2) the computational inefficiency of the feature extraction processes, such that extracting 4th-order autoregression from highly spatiotemporal HD-sEMG collected from 256 electrodes can be more than half a second; these processes will be even less efficient on wearable devices with limited power. Researchers favor DL approaches in sEMG-based research due to their strong feature extraction power, which relaxes the need for expertise in sEMG compared to feature-engineering-based conventional ML. Adaptive DL methods have used domain adaptation (DA) to calibrate their previously trained models from one day to another to enhance the interday performance. Although most of the existing DL efforts have reduced the pre-processing time by directly feeding their deepnets with raw sEMG signals, their methods did not minimize the induced myoelectric control delay by using window sizes from 150 ms to 300 ms and sEMG signals from the plateau phase.

Regardless of the approaches, the limitations of the current literature can be summarized into (1) time-consuming feature extraction, (2) large window sizes ranging from 150 to 750 ms, and (3) the use of plateau-phase sEMG, neglecting the rich information during the dynamic transient phase when gesture-associated motor unit recruitment occurs. Other constraints of existing literature include but are not limited to (1) the low number of subjects (i.e., $<$20), which makes a control system not representative enough to general users, (2) the exclusion of dynamic gestures, which makes the system unrealistic since hand gestures are naturally dynamic rather than static in real-world applications, and (3) training on data only from multiple consecutive days (where the potential changes in sEMG can be minimized), which is impractical due to the progressive performance degradation when the interval between training and testing days is large.

In this paper, we propose a multi-day dynamic hand gesture prediction method through decoding HD-sEMG signals to overcome the above limitations. We aim for a seamless (i.e., real-time-ready) and realistic hand gesture prediction system based on a deep vision transform structure. The restriction of control delay is lifted by investigating 11 dynamic gestures and using a small window size of 50 ms. Compared to static gestures, dynamic gestures are more natural and realistic, thus more complex to detect due to the dynamicity of neurophysiology during such tasks. It should be noted that the 50 ms window size is only one-sixth of the maximum window size qualified for real-time implementation [12] and the smallest size, to the best of our knowledge, in multi-day HGR research. Leveraging the power of the Multi-headed Self-attention (MSA) mechanism in a vision transformer (ViT), the proposed model can predict dynamic 11 gestures with an average accuracy of over 71% across 20 subjects on the testing day, 3-25 days after training. When calibrated on minimum data from the testing day (i.e., one or two repetitions), the proposed model can achieve over 92% accuracy by retraining only 8.8% of the parameters, almost regaining the intraday accuracy, for which the proposed model is trained and tested on the same day.

The *main contributions* of this paper to the research of multi-day HGR are as follows:

- This paper proposes the first ViT-based network to improve the multi-day performance of dynamic hand gesture prediction by capturing the temporal relationships within each HD-sEMG window through attention mechanisms.
- 50 ms is the shortest window size used in multi-day HGR, minimizing the latency of myoelectric control to the best of our knowledge. This window size is only one-sixth of the requirement for real-time implementation, preparing the proposed model for seamless real-life practices.
- This paper investigates the minimum calibration data needed from the testing day to maintain the performance of the proposed model trained on a previous day and reused on the testing day. This investigation, which falls under the umbrella of few-shot learning, is for the first time in the research of sEMG-based multi-day HGR when only two days of data are available.

The rest of the paper is organized as follows: the Prior Works section summarizes the previous studies in single-day and multi-day HGR. The Dataset and Pre-processing section describes the acquisition and pre-processing of the datasets. The model architecture, the model evaluation protocols, and the statistical test are introduced in the Methods section. The Results section presents the model performance, followed by the comparative study section that compares our proposed method with the state-of-the-art approaches when solving the same multi-day HGR task. Finally, the Conclusion section summarizes this paper.

II. Prior Works

Single-day HGR: For single-day hand gesture recognition/prediction, a model is trained and tested on sEMG signals collected from the same day. Classical ML relies on researchers' expertise and feature engineering for feature extraction and models, such as LDA [24], [25], SVM [26], and K-nearest Neighbor (KNN) [27], for classification. CNNs [28], Recurrent Neural Networks (RNNs) [29], and other DL methods [30] automatically extract various temporal, spectral, and spatial features from sEMG through their hierarchical architectures.

Although researchers have reported high performance on a large number of classified gestures, the short-term (intraday) performance cannot be translated into long-term (interday) performance due to the changes in the characteristics of sEMG signals over time caused by electrode artifacts, electrode displacement, and electrode misplacement. Such artifacts mainly include (a) stochastic electromagnetic noises (such as fluorescent noise, power-line noise, and those by nearby electronics),

TABLE I
COMPARISON BETWEEN THE PROPOSED MODEL WITH THE STATE-OF-THE-ART EFFORTS IN SEMG-BASED MULTI-DAY HGR.

Paper	# Subs	# Moves	# Reps	# Days	Rep Len	Win Len	Signal Type	Method	Interday Acc
[13]	10	10+rest	4	7c	3 s	200 ms	plateau	feature engineering+AE	66-80%
[14]	6	12+rest	1	10c	5 s	256 ms	plateau	feature engineering+LDA	72.3%
[15]	10	11+rest	5	10	4 s	200 ms	complete	feature engineering+(KNN,LDA)	86.61%
[16]	20	10+rest	6	2	1 s	500 ms	dynamic	feature engineering+SVM	92.2%
[17]	10	10+rest	4	7c	3 s	160 ms	plateau	feature engineering+ANN	76.2-85.6%
[18]	6	12+rest	1	10c	5 s	256 ms	plateau	(feature engineering,raw)+CNN	83.42%
[19]	10	7	12	5c	6 s	150 ms	complete	raw+TCN	49.4%
[20]	8	10+rest	4	7c	5 s	200 ms	complete	mel spectrogram+CNN	65.88-88.73%
[21]	10	8	10	2	1 s	300 ms	plateau	raw+RNN	54.6% (w/o cal); 83.8 (w/ cal)
[22]	10	8	10	2	1 s	150 ms	plateau	raw+CNN+AdaBN	63.3%
[23]	20	10+rest	3	3	5 s	150 ms	dynamic	feature engineering+SCADANN	53.08% (w/o cal); 55.69% (w/ cal)
ours	**20**	**11**	**6**	**2**	**1 s**	**50 ms**	**dynamic**	**ViT**	**71.34% (w/o cal); 88.87% (w/ 1 rep cal); 92.25% (w/ 2 reps cal)**

Note: #: Number; Subs: Subjects; Rep: Repetition; Len: Length; Win: Window; Acc: Accuracy; s: Second; ms: Millisecond; c: Consecutive; w/o: Without; w/: With; cal: Calibration; KNN: K-Nearest Neighbors; ANN: Artificial Neural Network; AdaBN: Adaptive Batch Normalization; SCADANN: Self-Calibrating Asynchronous Domain Adversarial Neural Network.

(b) signal deterioration due to degraded electrode skin contact impedance (due to hair blockage and sweat), and (c) changes in capacitive coupling [31]. Electrode displacement results from electrode shift on the skin surface [32], while electrode misplacement happens due to imprecise electrode positioning [33]. Other factors include but are not limited to natural sEMG variation over time and the muscle contraction effort of subjects [34]. Due to the lack of robustness of the existing models to the sources of signal variation, commercial sEMG-PR-based control systems (e.g., myoelectric prostheses) are currently limited on the market [35], [36]. Therefore, it is crucial to address the day-to-day reliability by proposing algorithms that can generalized to sEMG collected from multiple days, enhancing the scalability and feasibility of such wearable neural interfaces for human-robot interaction and control.

Multi-day HGR: Researchers have investigated multi-day HGR, where a model is trained on the previous day(s) and reused on the testing day with or without being calibrated on the data from the testing day. These works can also be categorized into conventional ML and DL methods. Existing research based on conventional ML [13]–[17] relies on manually extracting commonly used temporal (e.g., Hudgin's time-domain features [37] and autoregression coefficients), spectral (e.g., median frequency and spectral entropy) features. The extracted features are often optimized using dimensionality reduction techniques (e.g., principal component analysis) to achieve computational efficiency and prevent overfitting. A traditional ML classifier (e.g., SVM) is then trained on the features extracted from sEMG signals collected from the previous day(s) and tested on the new day. One of the problems of the traditional ML methods based on feature engineering is the lack of adaptability to the data from the new day. Adaptive LDA (ALDA) is one of the most commonly used adaptive ML approaches in this category [38]–[41]. ALDA can be calibrated on the new day's data by updating its mean and covariance matrices based on the same parameters from previous and new days to capture the interday sEMG variation to improve interday performance.

Existing DL works in multi-day HGR have developed deepnets (e.g., CNNs and autoencoders or AEs) to automatically extract HGR-related features from raw sEMG signals in the time domain [18], [19], [42], or from spectrograms in the time-and-frequency domain [20]. Domain adaptation (DA) is the adaptive technique for DL methods to calibrate their previously trained models from one day to another to enhance the interday performance. DA aims to develop a discriminative predictor on the data from the source domain and then to adapt the predictor to the data from the target domain, which is different but related to the source domain, possibly achieving high performance on the testing day. One way to apply DA is through transfer learning using labeled data from the testing day [21]. The proposed deepnet consists of a DA layer followed by a classifier. In the pre-training stage, only the classifier is trained on sEMG collected before the testing day (source domain) while the DA layer is frozen. During the DA stage, the classifier is frozen at its pre-trained stage while the DA layer is trained on sEMG collected on the testing day (target domain). The other way is to reduce the domain divergence between the labeled sEMG (in the source domain) and the unlabeled sEMG (in the target domain) by progressively updating their proposed deepnets using the unlabeled sEMG based on techniques such as pseudo-labels generating heuristic and Adaptive Batch Normalization [22], [23], [43]. The comparison between our paper and the existing state-of-the-art efforts in sEMG-based multi-day HGR is summarized in Table I.

III. DATASET AND PRE-PROCESSING

A. Dataset

We employ the PR dataset in a publicly available HD-sEMG database, referred to as "Hyser" [44]. This dataset includes 34

This article has been accepted for publication in IEEE Journal of Selected Topics in Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2024.3402340

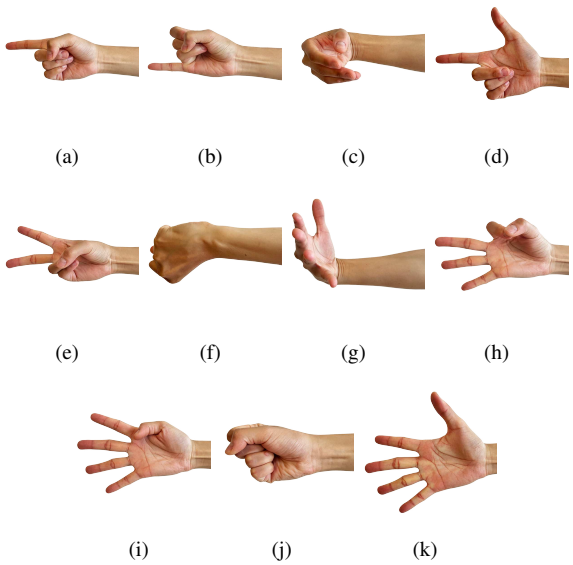

Fig. 1. 11 hand gestures that are common to all 20 subjects in the Hyser PR dataset are demonstrated. These gestures are the following: (a) index finger extension (IFE), (b) little finger extension (LFE), (c) wrist flexion (WF), (d) extension of thumb and index fingers (ETIF), (e) extension of index and middle fingers (EIMF), (f) wrist supination combined with hand close (WSHC), (g) wrist flexion combined with hand open (WFHO), (h) extension of middle, ring and little fingers (EMRLF), (i) extension of index, middle, ring and little fingers (EIMRLF), (j) hand close (HC), and (k) hand open (HO).

hand and finger gestures recorded from 20 intact subjects (12 male and 8 female with an average age of 26.5). In order to incorporate two-day data from all 20 subjects for statistical analysis, 11 gestures that are common to all subjects have been included in this study. Fig. 1 shows these gestures.

HD-sEMG signals were recorded using four 8×8 electrode array patches (256 sensors in total) by attaching two patches of sensors at each of the forearm sides (extensor and flexor) with a sampling rate of 2048 Hz. Each subject was instructed to perform two trials for each gesture, each trial containing three dynamic tasks of one-second duration from the resting state to the desired gesture and one maintenance task of holding that gesture for four seconds. In this study, we focus only on the dynamic tasks, resulting in six repetitions of one-second duration per gesture for each individual subject. Data were recorded from two different days with a between-day interval of 3 to 25 days, hereafter referred to as Day 1 (first day of recording) and Day 2 (second day of recording).

Remark 1: It should be noted that this paper implicitly tackles electrode placement, signal degradation over time, and the impact of environmental conditions on signal quality, which are factors that can result in day-to-day variations of sEMG. Our proposed method has inherent encoded robustness due to the loosely controlled environment of data acquisition imposed by the utilized database.•

B. Data Pre-processing

"Hyser" pre-processed their PR dataset by applying a 10-500 Hz 8th-order Butterworth band-pass filter, followed by a notch filter that attenuates power-line interference at 1st-8th harmonics of 50 Hz. These pre-processed signals are further filtered by an 8th-order Butterworth low-pass filter at 200 Hz. Then, the first 250 ms of the signals corresponding to the reaction time are removed before windowing. Next, each signal is segmented into 50 ms windows with a stride of 10 ms, complying with the real-time requirements. More specifically, this window length corresponds to one-sixth of the maximum window length allowed for real-time analysis. To the best of our knowledge, this is the shortest window length used for multi-day gesture detection tasks using sEMG signals. In order to minimize all possible delays, no further pre-processing is performed on the windowed data. The input data shape is adjusted according to the model at hand. For the proposed ViT model, we use the input shape of $100 \times 4 \times 8 \times 8$ corresponding to window length $\times$ # electrode grids $\times$ electrode width $\times$ electrode length. For RNN models (described in section VI), we use the input shape of 100×256 corresponding to window length $\times$ # electrodes.

IV. Methods

A. Model Architecture

ViTs have attracted extensive attention in image classification, challenging the dominance of CNNs. They introduce fewer inductive biases into the architectures than CNNs, which allows them to capture global contextual information by dividing each image into non-overlapping patches. Attention mechanisms enable ViTs to focus on relevant patches and establish long-range dependencies to learn complex patch relationships, capturing important visual patterns. Because of the patch extraction, ViTs are also scalable for high-resolution tasks that require fine-grained details. Position embedding preserves the spatial arrangement of the patches, whereas the spatial information may be lost due to the pooling layers of CNNs [45], [46]. sEMG-based HGR can benefit from the above advantages of ViTs, but ViT-based HGR methods using sEMG or HD-sEMG have been rarely studied in the literature, neglecting the temporal relationship between signals at any two timestamps. Montazerin et al. [47] just published the first ViT-based model for HD-sEMG-based HGR with intraday evaluation. In this paper, for the first time, we propose a compact deep neural network with a ViT backbone in multi-day dynamic hand gesture prediction from HD-sEMG signals, named ViT-MDHGR. The proposed model captures cross-day features by learning the relationships between HD-sEMG signals at any two timestamps within a window. The four components of the proposed model (patch embedding, position embedding, transformer encoder, and Multi-Layer Perceptron or MLP head) are introduced below. We follow most of the notations from the original ViT paper [45].

Patch Embedding: The raw input of each HD-sEMG window can be represented as a tensor $x \in \mathbb{R}^{T \times N_g \times H \times W}$, where T is the window length, N_g is the number of electrode grids, and $H \times W$ is the shape of each electrode grid. In our experiment, $T = 100$, $N_g = 4$, and $H \times W = 8 \times 8$. As we aim to learn the relationships between the signals at any two timestamps of an HD-sEMG window, we consider the signals from the electrode grids at timestamp, $x[i, :, :, :]$, $i = 1, \ldots, T$,

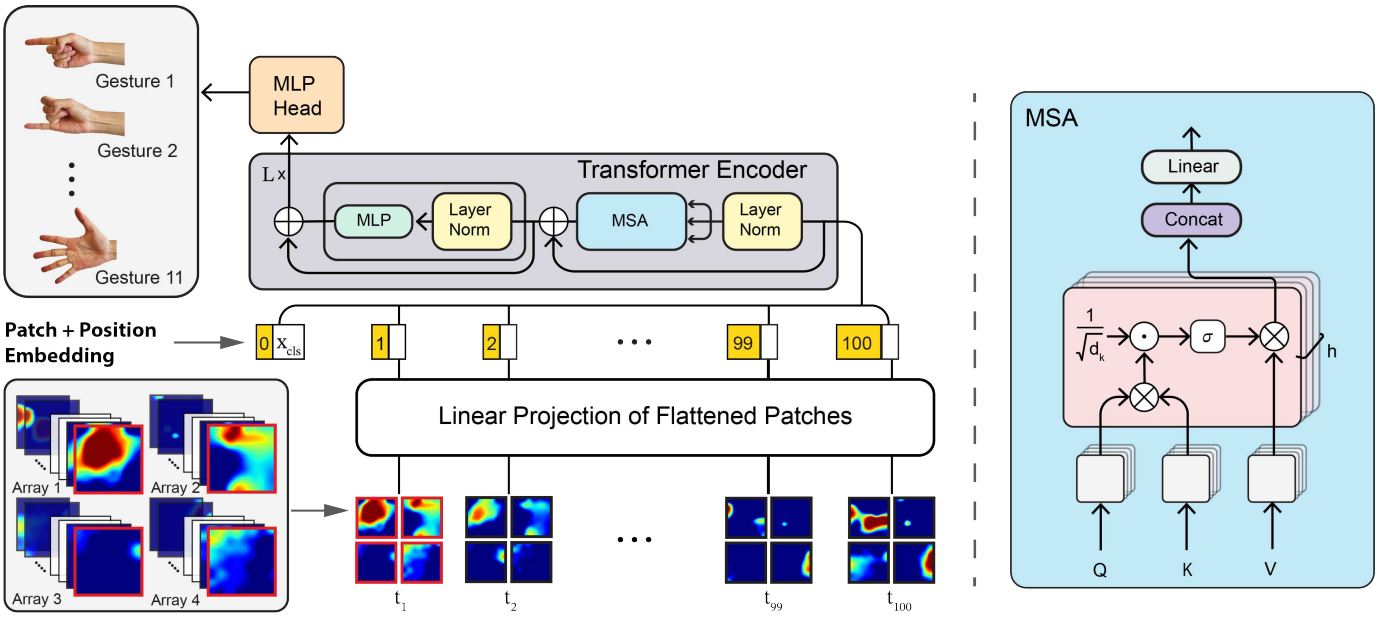

Fig. 2. Model architecture. The entire model is trained during the pre-training stage, while only the Linear Projection layer (the white rectangle connected to HD-sEMG heatmaps), which accounts for only 8.8% model complexity, is retrained. σ denotes the $\mathbf{softmax}$ function in MSA.

as a patch. Therefore, we reshape an HD-sEMG window input x into a sequence of flattened patches $x_p \in \mathbb{R}^{T \times N_g HW}$. A trainable linear projection layer maps these flattened patches at each time stamp to size D latent vectors. The output of the projection is the patch embeddings. Then we prepend a class token, a vector of learnable embeddings $x_{cls} \in \mathbb{R}^D$, to the patch embeddings, turning the shape of patch embeddings to $\mathbb{R}^{(T+1)\times D}$. The state of the class token at the output of the transformer encoder is referred to as the classification head, which serves as the representation of an HD-sEMG window to be classified as different gestures.

Position Embedding: Learnable position embeddings, $p \in \mathbb{R}^{(T+1)\times D}$, are added to the patch embeddings (including the class token) to retain the position information of the patches. Since we are interested in the patch relationships along the time axis and consider the input as a sequence rather than a grid of patches, we choose one-dimensional position embeddings. The position embeddings are initialized from a standard normal distribution. The input of the transformer encoder is:

$$z_0 = [x_{cls}, x_p^1 E, x_p^2 E, \ldots, x_p^T E] + p, \tag{1}$$

where $E \in \mathbb{R}^{N_g HW \times D}$ is the matrix for the linear projection.

Transformer Encoder: The transformer encoder consists of multiple layers of MSA and MLP blocks. Pre-norm happens before each block, where Layer Normalization (LN) is applied to the block input to estimate the normalization statistics and residual connections after (by adding the block input to the output) to improve model convergence. The outputs of an MSA block and an MLP block on layer ℓ are (2) and (3), respectively.

$$z'_\ell = MSA(LN(z'_{\ell-1})) + z'_{\ell-1} \tag{2}$$

$$z_\ell = MLP(LN(z'_\ell)) + z'_\ell \tag{3}$$

MSA has a building block of the standard **qkv** self-attention [48], where the sequence of patches z_0 is linearly projected separately into Q (queries), K (keys), and V (values), all with the same dimension of $h \times (T+1) \times d_k$, h denotes the number of heads and d_k is the head dimension. The attention weights A are the similarities between two patches of the sequence and are calculated as

$$A = \text{softmax}(\frac{QK^\intercal}{\sqrt{d_k}}). \tag{4}$$

The scaling factor $\sqrt{d_k}$ ensures the dot product of Q and K will not be a large number. The softmax function converts the scaled dot product to the range between 0 and 1, indicating the total attention paid to V sums up to 1. Therefore, the self-attention operation for each head is

$$MSA(Q, K, V) = \text{softmax}(\frac{QK^\intercal}{\sqrt{d_k}})V. \tag{5}$$

The self-attention operations of h heads are run in parallel. The outputs of all the heads are concatenated and projected to form the input of the MLP block after the residual connection.

An MLP block consists of two linear layers with Gaussian Error Linear Unit (GELU) activation and a dropout in between. The output of an MLP block matches the dimension of the latent space D.

MLP Head: The MLP head is a linear classifier that predicts the classification head z_L^0, the class token from the last layer L of the transformer encoder, into different gestures $\hat{y}$ as follows:

$$\hat{y} = \text{softmax}(FC(LN(z_L^0))), \tag{6}$$

where FC denotes "fully connected", mapping z_L^0 from dimension D to the total number of predicted gestures (i.e., 11 in this paper).

This article has been accepted for publication in IEEE Journal of Selected Topics in Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2024.3402340

The proposed model is trained for a maximum of 200 epochs with a batch size of 32 for all the experiments, including pre-training and calibration. We use Adam as the optimizer with an adaptive learning rate of 0.001, which will be reduced to half at Epoch 40 and 80. Early stopping is employed with patience 40, such that the proposed model will stop training if the validation accuracy does not improve for 40 consecutive epochs.

B. Model Evaluation Protocols

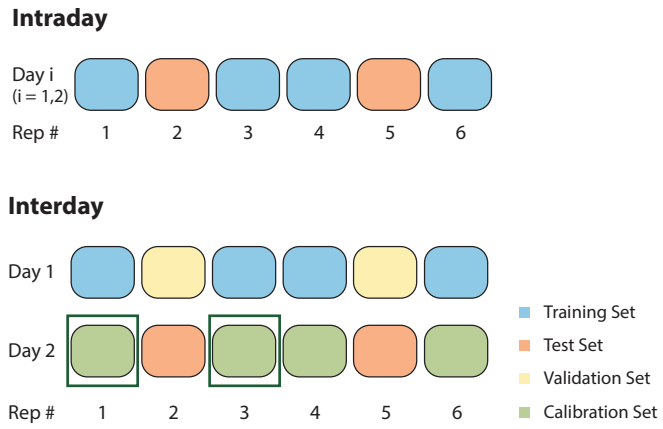

Fig. 3. Visualization of training/training-calibration strategy of model evaluation for intraday (upper figure) and interday (lower figure) performances. Each colored rectangle with rounded corners represents a repetition. The two dark green rectangles show an example of 2 Reps Calibration on the repetitions (1 and 3) from Day 2.

This paper evaluates the proposed model for intraday and interday performance. The intraday performance serves as a benchmark to be compared with the interday performance to see (1) the acceptable performance degradation of our model trained on one day and used on another without calibration and (2) the slim performance gap between the intraday performance and the interday performance when calibrating our model on limited data from the testing day.

Intraday Performance: To evaluate the intraday performance of our proposed model on Day 1 or Day 2, we use the four repetitions (1, 3, 4, and 6) and the remaining two repetitions (2 and 5) of the same day for training and testing, respectively.

Interday Performance: We use the signals from Day 1 for training and validation and Day 2 for calibration and testing. As this paper leverages transfer learning to learn the HGR-related pre-knowledge on Day 1 and transfer this knowledge to each subject to predict gestures on Day 2, the evaluation of interday performance consists of two stages: pre-training and calibration stages.

This paper investigates two strategies in the pre-training stage: (1) pre-training our proposed model on the data from individual subjects, named "Pre-trained on Individuals", and (2) pre-training the proposed model on the data pooled from all 20 subjects, named "Pre-trained on All". The first pre-trained strategy will result in 20 pre-trained models, one for each subject, while the second strategy will result in only one general pre-trained model. Under both strategies, we pre-train and validate our model on the Training and Validation Sets (i.e., repetitions (1, 3, 4, and 6) and repetitions (2 and 5) on Day 1), respectively. The calibration stage includes three strategies: "0 Rep" calibration, "1 Rep" calibration, and "2 Reps" calibration. 0 Rep calibration directly evaluates a pre-trained model (either Pre-trained on Individuals or Pre-trained on All) on the Test Set (i.e., repetitions (2 and 5) on Day 2). To improve the performance of a pre-trained model on Day 2, we calibrate it on limited data from Day 2 by only retraining the Linear Projection layer (which accounts for 8.8% of total trainable parameters) of the pre-trained model. We conduct 4-fold or 6-fold cross-validation for calibration by selecting one or two repetitions from the Calibration set (i.e., repetitions (1, 3, 4, and 6) on Day 2). The training/training-calibration strategy of model evaluation for intraday and interday performances can be visualized in Fig. 3.

C. Statistical Test

The distribution of the differences between two compared groups of results is not normal. Also, the compared groups of results are dependent. Thus, we use Wilcoxon signed-rank test with an alpha value of 0.05 in this study. Asterisks ($*$) are used to indicate statistical significance with respect to p values. ns or not significant denotes $p > 0.05$; $*$ denotes $p \leq 0.05$; $**$ denotes $p \leq 0.01$; $***$ denotes $p \leq 0.001$; and $****$ denotes $p \leq 0.0001$.

V. Results

A. Model Configuration

This paper particularly focuses on learning the relationships of signals from all the sensors at any timestamp of an HD-sEMG window to enhance the long-term, interday performance of HGR. Thus, the dimension of each patch ($patch\_size$) is the same as each electrode grid (i.e., 8×8). We consider $channels$ as the number of used electrode grids, which is four for the "Hyser". As a result of hyper-parameter tuning, the dimension of the latent space D is set to 128. The transformer encoder has eight layers ($L = 8$), each having an MSA with four size 16 heads ($h = 4$ and $d_k = 16$). mlp_dim, which is the dimension of the MLP block's hidden layer, is set to 32. The dropout rates after the position embedding and inside the transformer encoder are 0.1 and 0.5, respectively. The model configuration is summarized in Table II.

TABLE II
MODEL CONFIGURATION (NOTATION AND VALUE PAIRS).

Notation	Value	Notation	Value
$patch\_size$	8×8	$channels$	4
D	128	L	8
h	4	d_k	16
mlp_dim	32		

B. Results of Evaluation Protocols

This paper aims to achieve high reliability and performance for day-to-day HGR without or with minimal calibration.

We investigate the minimum amount of data needed for our proposed model pre-trained on Day 1 to maintain the performance on Day 2, lifting the data collection burden from users and enhancing the ease of use of such HGR systems. In this regard, we conduct 0 Rep calibration, 1 Rep calibration, and 2 Reps calibration strategies on sEMG signals collected on Day 2 using our pre-trained model (i.e., Pre-trained on Individuals or Pre-trained on All). The calibration data of the latter two strategies account for 17% and 33% of data collected on Day 2, respectively. As 1 Rep and 2 Reps calibrations are evaluated using 4-fold and 6-fold cross-validation, the average accuracies across folds are reported as the final results. Our proposed model's intraday and interday performances are shown in Table III and Fig. 4.

Intraday performance: When training our proposed model from scratch on four repetitions (1, 3, 4, and 6), we can achieve consistently high accuracies of 94.32%±2.66% for Day 1 and 94.33%±1.78% for Day 2 averaged across 20 subjects. The distributions of sEMG signals collected on Day 1 and Day 2 are different due to sEMG variation over time, though the data collection followed the same protocol. Thus, the performance consistency shows that our proposed model performs stably in single-day HGR.

TABLE III
AVERAGE INTERDAY PERFORMANCE BY PRE-TRAINING STRATEGY.

Pre-training Strategy	0 Rep	1 Rep	2 Reps
Pre-trained on All	**71.34%**	**88.87%**	**92.25%**
Pre-trained on Individuals	62.84%	87.92%	91.38%

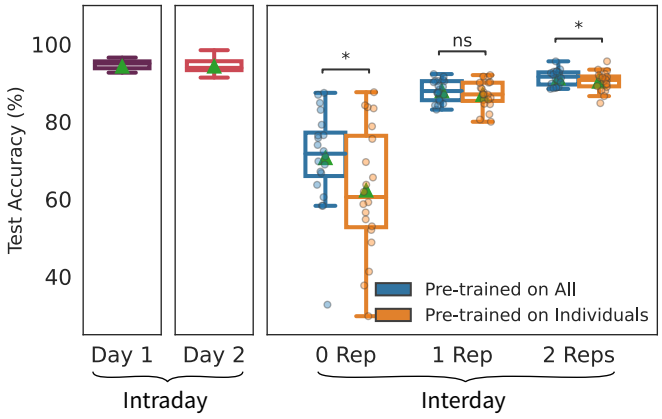

Fig. 4. Box plots of intraday and interday performance given different pre-training strategies. The number of data points in each box plot is 20, equal to the total subjects. Green triangles denote the averages.

Interday performance: Pre-trained on Individuals models can achieve 62.84%±16.67%, 87.92%±3.57%, and 91.38%±2.50% averaged across all subjects in 0 Rep, 1 Rep, and 2 Reps calibrations. In comparison, these performances are 8.5%±3.91%, 0.95%±0.76%, and 0.87%±0.5% higher when using the Pre-trained on All model. It should be noted that the average standard deviations are lower, indicating that our model pre-trained on all subjects is more confident in predicting gestures based on sEMG signals collected on Day 2 than the Pre-trained on Individuals models. Thus, pre-training on all subjects can help our proposed model learn time-invariant HGR-related knowledge from sEMG signals collected on Day 1. This knowledge can be more generalized, emphasized, and strengthened when learned from all subjects than captured from individuals. Furthermore, the interday performance of our proposed model can almost match the intraday one when using the Pre-trained on All and 2 Reps calibration strategies only with a 2.08% accuracy gap.

TABLE IV
AVERAGE INTERDAY PERFORMANCE BY WINDOW SIZE UNDER PRE-TRAINED ON ALL STRATEGY.

Window Size	0 Rep	1 Rep	2 Reps
30 ms	66.58%	86.19%	89.59%
40 ms	69.20%	86.81%	90.31%
50 ms	71.34%	88.87%	92.25%
100 ms	74.16%	89.65%	92.44%
200 ms	74.70%	91.92%	94.46%

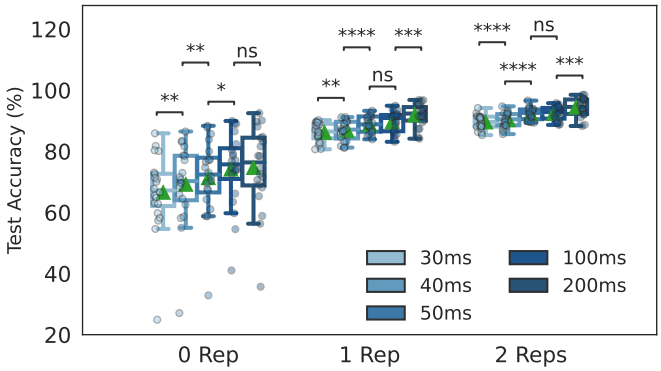

Fig. 5. Box plots of interday performance by window size under Pre-trained on All strategy. The number of data points in each box plot is 20, equal to the total subjects. Green triangles denote the averages. The shade of color darkens as the window size increases.

There is a trade-off between window size and the performance of an HGR system. Larger window sizes contain more information that can be important for the proposed model to learn to enhance HGR performance. However, larger window sizes can result in longer induced delays to myoelectric control and more complex model structures. This paper aims to find the shortest window size that can achieve the optimal interday HGR performance, preparing our proposed model for real-time cross-day use. Previous works [29], [47] achieved high intraday performance in classifying a large number of multi-degree-of-freedom gestures using short window sizes ranging from 30 ms to 250 ms. Hence, in this paper, we investigate the effect of window sizes on interday performance by pre-training (under the Pre-trained on All strategy) and calibrate our proposed model on five window sizes: 30 ms, 40 ms, 50 ms, 100 ms, and 200 ms. The results are shown in Table IV and Fig. 5. We statistically analyze and compare the interday performances on (30 ms, 40 ms), (40 ms, 50 ms), (50 ms, 100 ms), and (100 ms, 200 ms) pairs of window sizes on each calibration strategy. As shown in Fig. 5, the performance on 50 ms windows is significantly higher than on 30 ms and 40 ms windows but similar to the performance on 100 ms windows.

This article has been accepted for publication in IEEE Journal of Selected Topics in Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2024.3402340

As a result, our proposed model achieves the most optimistic interday performance on 50 ms windows.

This paper also investigates gesture-wise reliability across days by analyzing the interday performance of our proposed model without calibration (0 Rep calibration) for each gesture performed by the top-5-performing subjects (subjects 1, 7, 9, 18, and 19). As a result, Fig. 6 shows that our proposed model can robustly and reliably predict almost all gestures on two different days. Our model achieves more than 90% accuracy in predicting five gestures (IFE, WF, EIMF, HO, and EIMRLF). Only one gesture (WSCHC) has an accuracy of less than 75%.

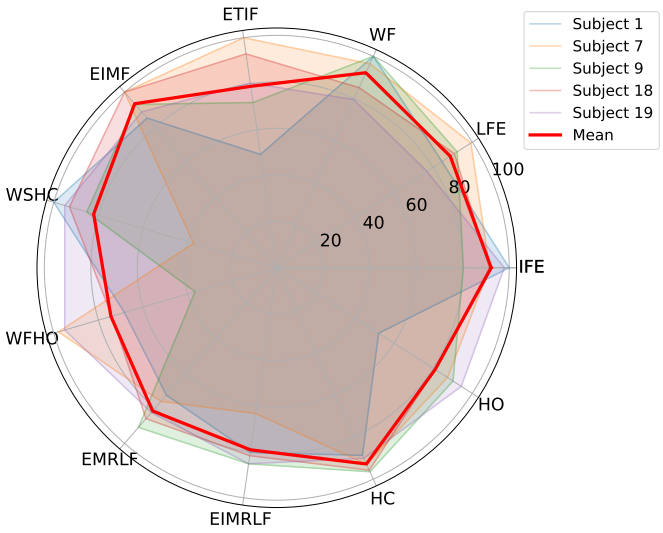

Fig. 6. Gesture-wise average interday performance without calibration (0 Rep calibration) on the top-5-performing subjects. The bold red line denotes the interday performance averaged across the five subjects.

Remark 2: This paper uses the 50 ms window length, which accounts for only one-sixth of the real-time requirement [12] and less than 66% of the smallest window length (i.e., 150 ms) used in the existing literature on multi-day HGR. Using such a short window length is already a challenge in single-day HGR, where a proposed method needs to distinguish different gestures given limited temporal information. This challenge will be amplified in multi-day HGR, where the method not only needs to differentiate one gesture from another but also recognize the same gesture by acknowledging the variations in the sEMG signals collected from varying days.•

Remark 3: Due to the cross-day variations in sEMG signals, the interday reliability of an HGR system depends on calibrating the system on the data from the testing day. There is a trade-off between the amount of data needed for calibration from the testing day and the practicality of such HGR systems; generally, the more data required for calibrating an HGR system on another day, the more tedious the calibration process will be, the less likely the users will use such a system frequently in daily living so less practicality and translational value. This paper investigates the minimum calibration data needed from the testing day to maintain the performance of an HGR system trained on a previous day, enhancing the feasibility of such a system in real-life practices. As a result, our proposed ViT-based network secures ≈60% higher performance than the general literature [11] when trained and tested on different days without calibration. Furthermore, calibrated on minimum data (i.e., up to two repetitions data) from the testing day, the proposed model can almost match the interday performance with the intraday. •

Remark 4: ViT models have usually been used to capture global spatial contextual information. In this paper, for multi-day HGR, we propose the first ViT-based deep-learning model to capture the global "temporal contextual" information between signals at any two timestamps within an HD-sEMG window. The innovation of temporally rather than spatially segmenting each input into patches enables the proposed model to be best suited for multi-day HGR, achieving high performance with the shortest window length and minimum calibration mentioned in the previous two remarks. The compactness of the novel design of the transformer structure in addition to the need for only retraining 8.8% of the model on a new day further enhances the real-life practicality.•

VI. Comparative Study

In this study, we present a novel strategy to generalize HGR across days while requiring minimal calibration. In order to evaluate the performance of our model compared to the state of the art, we comprehensively conduct three sets of comparative experiments. In one experiment, we compare our model with one of the most common approaches toward generalizable gesture detection using sEMG signals: feature engineering combined with ALDA. We show that our model significantly outperforms ALDA under similar training and calibration configurations. To further highlight the superiority of our proposed model over other state-of-the-art models, in another experiment, we compare the performance of our ViT-MDHGR with three well-known RNN models which have proven to be effective when addressing time-series data [49]. Finally, as three-dimensional CNNs (3D CNNs) [50] become increasingly popular in HGR for extracting temporal and spatial information from HD-sEMG using 3D kernels, we compare our proposed model with a 3D CNN model. These experiments follow the same pre-training (i.e., Pre-trained on All) and calibration strategies as our ViT-MDHGR. The experimental details are described as follows.

Comparison to ALDA: ALDA based on feature engineering has been subject to extensive studies in the gesture recognition literature [39]–[41]. In order to demonstrate the superiority of our ViT-MDHGR over ALDA, we adopt a feature engineering + ALDA model to the same dataset ("Hyser") following the general methodology of [39]. More specifically, we create a pipeline that extracts a number of features from each 50 ms signal window and classifies the feature vectors into gestures using an LDA classifier. The features we considered are Mean Absolute Value, Zero Crossing, Slope Sign Changes, and Waveform Length similar to [39]. The training and calibration stages of ALDA are described as follows.

- **Training:** Features are extracted from the Training Set of all 20 subjects. Next, an LDA with the following discriminant function is fit to the extracted features following

[39], [41]:

$$g_c(x) = x^\intercal \Sigma^{-1} \mu_c - \frac{1}{2} \mu_c^\intercal \Sigma^{-1} \mu_c + \log \pi_c \quad (7)$$

where μ_c is the mean value for class c, Σ is the covariance matrix of classes (assumed to be equal for all classes), and π_c is the prior probability for class c.

- **Calibration:** Given new samples on Day 2 (Any one or two repetitions selected from the Calibration Set) for each subject, we adapt the LDA by readjusting the mean vectors and covariance matrix similar to [41] as follows:

$$\tilde{\mu}_c = (1-\lambda)\mu_{tr_c} + \lambda \mu_{cal_c} \quad (8a)$$

$$\tilde{\Sigma}_c = (1-\lambda)\Sigma_{tr_c} + \lambda \Sigma_{cal_c} \quad (8b)$$

where $(\mu_{tr_c}, \Sigma_{tr_c})$ refer to training (mean, covariance) and $(\mu_{cal_c}, \Sigma_{cal_c})$ refer to calibration data (mean, covariance). $\lambda = 0.5$ is a hyperparameter.

Limitations of ALDA: Approaches based on ALDA rely on feature extraction which can cause control delays in the gesture detection pipeline. In our comparative study, we have selected common features that are theoretically suitable to real-time applications (e.g. they are causal and do not rely on future windows). However, our experiments show that extracting these features for each signal window takes about 12.1-14.9 ms across various subjects. Noting that the window increments used in this study are 10 ms, this time-consuming feature extraction step can cause delays in the gesture detection pipeline. An advantage of our ViT-MDHGR model is that it does not require such time consuming feature extraction steps and thus does not suffer additional delays. Another fundamental limitation of ALDA is the assumption that all gesture classes have the same covariance matrix Σ which might not be true in practice. Another advantage of our proposed model is that it does not require any assumptions on the data distribution of gesture classes. Finally, looking at Fig. 7 we note that ViT-MDHGR significantly outperforms ALDA when there is no calibration data available.

Comparison to RNNs: The compared models in this experiment set are Gated Recurrent Units (GRUs) [51], Long Short Term Memories (LSTMs) [52], and Bidirectional LSTMs (BILSTMs). Each model has three RNN layers followed by FC layer(s) with hyperbolic tangent activation and Dropout (rate=0.2) layers. The last layer is an FC layer with softmax activation that outputs a predicted gesture. In the calibration stage, only the first RNN layer is trainable. Table V shows the configurations and hyperparameters of RNN models.

Limitations of RNNs: Looking at Tables VII and VIII we note that RNN-based models that have similar model complexity and no calibration Day 2 performance as ViT-MDHGR require a larger percentage of their parameters to be retrained to yield satisfactory interday performance. We also note that despite the retraining of around $40-50\%$ of their parameters, RNN-based models perform significantly worse than ViT-MDHGR on Day 2. This implies that RNN-based models are more sensitive to interday variations and therefore need a more extensive retraining to improve their accuracy. The advantage of ViT-MDHGR over RNN-based models is that our proposed model outperforms the RNN-based models while requiring significantly fewer parameters to be retrained.

TABLE V
RNN MODEL CONFIGURATIONS

Hyperparameter	GRU	LSTM	BILSTM
layers	3	3	3
hidden units	134	115	75
dilation order	3	3	3
FC layers before classifier	2	1	1
Dropout	None	1	1

Comparison to 3D CNN: Similar to the previous experiments, we compare the ViT-MDHGR model with a 3D CNN model structured as shown in Table VI. For calibration, we freeze everything except the second 3D convolution and batch normalization layers.

Limitations of 3D CNN: Looking at Table VIII and Fig. 7, we note that the 3D CNN model is more sensitive to interday variations. Learning local information within the sliding kernels, instead of capturing the global contextual information through the multi-headed self-attention mechanism (ViT-MDHGR), can be the reason for the inefficacy of the interday performance of 3D CNN. However, we note that a 3D CNN model in general is lightweight compared to ViT-MDHGR. Despite the comparatively light-weight architecture, 3D CNN in general performs poorly compared to the other state-of-the-art models and ViT-MDHGR with or without calibration.

TABLE VI
MODEL STRUCTURE FOR 3D CNN. C: CHANNELS/UNITS, K: KERNEL SIZE, D: DILATION RATE, A: ACTIVATION.

Layer	Configuration
3D convolution	c = 16, k = (8, 2, 2), d=(2, 1, 1), a='relu'
Batch normalization	N/A
3D convolution	c=32, k=(8, 2, 2), d=(2, 1, 1), a='relu'
Batch normalization	N/A
3D Max pooling	N/A
3D convolution	c=64, k=(8, 2, 2), d=(4, 1, 1), a='relu'
Batch normalization	N/A
3D convolution	c=64, k=(8, 2, 2), d=N/A, a='relu'
Batch normalization	N/A
flatten	N/A
FC	c=32, a='tanh'
Dropout	rate=0.2
FC (classifier)	c=11, a='softmax'

TABLE VII
TRAINABLE PARAMETERS IN PRE-TRAINING AND CALIBRATION STAGES.

Model	Pre-training	Calibration	Percentage
ViT-MDHGR	382,475	33,664	8.8%
GRU	385,747	157,584	40.8%
LSTM	387,715	171,120	44.1%
BILSTM	385,595	199,200	51.6%
3D CNN	218,011	16,480	7.5%

Table VII shows the total number of trainable parameters in the pre-training and calibration stages for each of the models. The ratio of the trainable parameters of the calibration stage to the pre-training stage is also shown. We note that

TABLE VIII
PERFORMANCE EVALUATION OF DIFFERENT MODELS. MEAN VALUES ARE REPORTED FOR CALIBRATION RESULTS.

Model	0 Rep	1 Rep	2 Reps
ViT-MDHGR	**71.34%**	**88.87%**	**92.25%**
GRU	69.02%	78.07%	81.20%
LSTM	70.26%	81.41%	84.59%
BILSTM	70.61%	82.11%	84.80%
ALDA	50.84%	74.76%	81.41%
3D CNN	67.49%	79.47%	83.05%

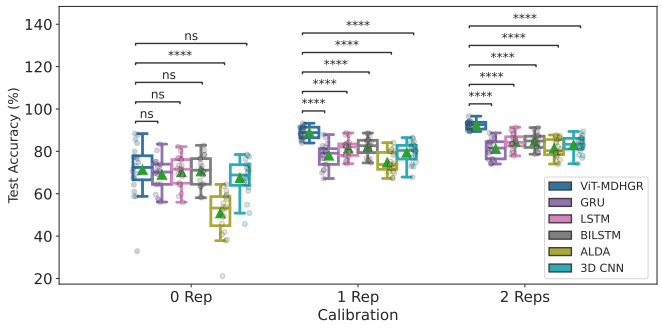

Fig. 7. Day 2 Accuracy comparison between ViT-MDHGR and various state of the art models. The number of data points in each box plot is 20, equal to the total subjects. Green triangles show mean values across 20 subjects.

ViT-MDHGR and 3D CNN require significantly lower ratios of trainable parameters during calibration. Table VIII shows how these models compare to each other in various interday calibration scenarios. Mean values are shown for various Calibration percentages and the models are calibrated on individual subjects separately. Fig. 7 gives a more detailed view of Calibration results for the two models. We make the following observations:

Observation 1: ViT-MDHGR slightly outperforms the RNN models when evaluated on Day 2 data with 0 Rep calibration. However, by calibrating over one repetition, the performance gap increases to more than 6%. Calibrating on two repetitions increases this gap to around 8% compared to the best RNN model. We emphasize that in all scenarios, our proposed model outperforms RNN models while requiring calibration over significantly fewer parameters.

Observation 2: ViT-MDHGR significantly outperforms ALDA in all scenarios. When calibrating on two repetitions, this gap is around 10.84%. In comparison to ALDA, We emphasize that our proposed model does not require any additional pre-processing or feature extraction. ViT-MDHGR also outperforms 3D CNN, more significantly when two repetitions for calibration are available.

Observation 3: When calibrated on one or two repetitions, ViT-MDHGR is more reliable in the sense that it has a lower variance among 20 subjects. We also note that when calibrated on two repetitions, the worst ViT-MDHGR accuracy is still higher than the third quartile values of all the other models.

In general, these results confirm the potential and reliability of our proposed model for accurate and agile gesture recognition across days.

VII. CONCLUSION

In this study, we propose a ViT-based network (ViT-MDHGR) that can be deployed for real-time HGR from HD-sEMG signals. Our proposed model addresses the challenge of inter-day gesture recognition relying on 50 ms HD-sEMG signal windows. This innovation enhances the agility and responsiveness of the system, making it suitable for practical applications. Retraining only 8.8% of the model parameters on a different day, we show that this model can achieve an inter-day accuracy of 92.25% for detecting 11 gestures by calibration on only two repetitions of each gesture. We demonstrate that our proposed model significantly outperforms various sequential, CNN, and LDA based state of the art networks. This research highlights a significant step towards making multi-day hand gesture recognition a reality, with potential applications in myoelectric control of prosthetic devices, neurorobotics, and human-computer interfaces. The ViT-MDHGR not only improves the generalizability of hand gesture recognition but also offers a promising avenue for enhancing the usability and practicality of such systems in real-world contexts.

In section II, we enumerate electrode noise, electrode displacement, and electrode misplacement as the hurdles of reliable and practical HGR systems for multi-day uses. We recently discussed these issues separately and achieved high performance, providing promising solutions [53]–[57]. For future work, to understand the model's capability in real-world implementation thoroughly, we will conduct comprehensive studies on the effect of each challenge. In addition, real-time multi-day HGR is one of the future lines of research that bridges the gap between offline and online model performance evaluation. Other possible directions of future work include extending the number of gestures, incorporating subject-generalization, and increasing the length of inter-day intervals. Focusing on refining and expanding the capabilities of this model will bring us closer to seamless and efficient myoelectric control solutions for a wide range of applications.

ACKNOWLEDGMENT

We would like to acknowledge Mohammedali Roowala's contributions to this paper. Roowala is with the Department of Mechanical and Aerospace Engineering, New York University (NYU), New York, NY, 11201 USA.

REFERENCES

[1] J. Feher, "3.6 - the neuromuscular junction and Excitation-Contraction coupling," in *Quantitative Human Physiology*, J. Feher, Ed. Boston: Academic Press, Jan. 2012, pp. 259–269.
[2] A. Del Vecchio, A. Holobar, D. Falla, F. Felici, R. M. Enoka, and D. Farina, "Tutorial: Analysis of motor unit discharge characteristics from high-density surface EMG signals," *J. Electromyogr. Kinesiol.*, vol. 53, p. 102426, Aug. 2020.
[3] M. B. I. Raez, M. S. Hussain, and F. Mohd-Yasin, "Techniques of EMG signal analysis: detection, processing, classification and applications," *Biol. Proced. Online*, vol. 8, pp. 11–35, Mar. 2006.

[4] L. Bi, A.-g. Feleke, and C. Guan, "A review on EMG-based motor intention prediction of continuous human upper limb motion for human-robot collaboration," *Biomed. Signal Process. Control*, vol. 51, pp. 113–127, May 2019.

[5] Y. Zhang, B. Liang, B. Chen, P. M. Torrens, S. F. Atashzar, D. Lin, and Q. Sun, "Force-Aware interface via electromyography for natural VR/AR interaction," *ACM Trans. Graph.*, vol. 41, no. 6, pp. 1–18, Nov. 2022.

[6] M. Ergeneci, D. Carter, and P. Kosmas, "sEMG onset detection via bidirectional recurrent neural networks with applications to sports science," *IEEE Sens. J.*, vol. 22, no. 19, pp. 18 751–18 761, Oct. 2022.

[7] D. Kumar and A. Ganesh, "A critical review on hand gesture recognition using sEMG: Challenges, application, process and techniques," *J. Phys. Conf. Ser.*, vol. 2327, no. 1, p. 012075, Aug. 2022.

[8] A. W. Shehata, H. E. Williams, J. S. Hebert, and P. M. Pilarski, "Machine learning for the control of prosthetic arms: Using electromyographic signals for improved performance," *IEEE Signal Process. Mag.*, vol. 38, no. 4, pp. 46–53, Jul. 2021.

[9] A. Holobar and D. Farina, "Noninvasive neural interfacing with wearable muscle sensors: Combining convolutive blind source separation methods and deep learning techniques for neural decoding," *IEEE Signal Process. Mag.*, vol. 38, no. 4, pp. 103–118, Jul. 2021.

[10] R. H. Chowdhury, M. B. I. Reaz, M. A. B. M. Ali, A. A. A. Bakar, K. Chellappan, and T. G. Chang, "Surface electromyography signal processing and classification techniques," *Sensors*, vol. 13, no. 9, pp. 12 431–12 466, Sep. 2013.

[11] F. S. Botros, A. Phinyomark, and E. J. Scheme, "Day-to-Day stability of wrist EMG for Wearable-Based hand gesture recognition," *IEEE Access*, vol. 10, pp. 125 942–125 954, 2022.

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

[13] M. Zia ur Rehman, S. O. Gilani, A. Waris, I. K. Niazi, G. Slabaugh, D. Farina, and E. N. Kamavuako, "Stacked sparse autoencoders for EMG-Based classification of hand motions: A comparative multi day analyses between surface and intramuscular EMG," *NATO Adv. Sci. Inst. Ser. E Appl. Sci.*, vol. 8, no. 7, p. 1126, Jul. 2018.

[14] Z. Wang, Y. Fang, G. Li, and H. Liu, "Facilitate sEMG-Based Human–Machine interaction through channel optimization," *Int. J. Humanoid Rob.*, vol. 16, no. 04, p. 1941001, Aug. 2019.

[15] L. Meng, X. Jiang, X. Liu, J. Fan, H. Ren, Y. Guo, H. Diao, Z. Wang, C. Chen, C. Dai, and W. Chen, "User-Tailored hand gesture recognition system for wearable prosthesis and armband based on surface electromyogram," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–16, 2022.

[16] X. Jiang, X. Liu, J. Fan, X. Ye, C. Dai, E. A. Clancy, D. Farina, and W. Chen, "Optimization of HD-sEMG-Based Cross-Day hand gesture classification by optimal feature extraction and data augmentation," *IEEE Transactions on Human-Machine Systems*, vol. 52, no. 6, pp. 1281–1291, Dec. 2022.

[17] A. Waris, I. K. Niazi, M. Jamil, K. Englehart, W. Jensen, and E. N. Kamavuako, "Multiday evaluation of techniques for EMG-Based classification of hand motions," *IEEE J Biomed Health Inform*, vol. 23, no. 4, pp. 1526–1534, Jul. 2019.

[18] Y. Fang, X. Zhang, D. Zhou, and H. Liu, "Improve inter-day hand gesture recognition via convolutional neural network-based feature fusion," *Int. J. Humanoid Rob.*, vol. 18, no. 02, p. 2050025, Apr. 2021.

[19] M. Zanghieri, S. Benatti, F. Conti, A. Burrello, and L. Benini, "Temporal variability analysis in sEMG hand grasp recognition using temporal convolutional networks," in *2020 2nd IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, Aug. 2020, pp. 228–232.

[20] M. F. Qureshi, Z. Mushtaq, M. Z. u. Rehman, and E. N. Kamavuako, "Spectral Image-Based multiday surface electromyography classification of hand motions using CNN for Human–Computer interaction," *IEEE Sens. J.*, vol. 22, no. 21, pp. 20 676–20 683, Nov. 2022.

[21] *Domain Adaptation for sEMG-based Gesture Recognition with Recurrent Neural Networks*, Jan. 2019.

[22] Y. Du, W. Jin, W. Wei, Y. Hu, and W. Geng, "Surface EMG-Based Inter-Session gesture recognition enhanced by deep domain adaptation," *Sensors*, vol. 17, no. 3, Feb. 2017.

[23] U. Côté-Allard, G. Gagnon-Turcotte, A. Phinyomark, K. Glette, E. J. Scheme, F. Laviolette, and B. Gosselin, "Unsupervised domain adversarial Self-Calibration for Electromyography-Based gesture recognition," *IEEE Access*, vol. 8, pp. 177 941–177 955, 2020.

[24] Y. Fang, D. Zhou, K. Li, and H. Liu, "Interface prostheses with Classifier-Feedback-Based user training," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 11, pp. 2575–2583, Nov. 2017.

[25] S. A. Raurale, J. McAllister, and J. M. del Rincon, "Real-Time embedded EMG signal analysis for Wrist-Hand pose identification," *IEEE Trans. Signal Process.*, vol. 68, pp. 2713–2723, 2020.

[26] M. Tavakoli, C. Benussi, P. Alhais Lopes, L. B. Osorio, and A. T. de Almeida, "Robust hand gesture recognition with a double channel surface EMG wearable armband and SVM classifier," *Biomed. Signal Process. Control*, vol. 46, pp. 121–130, Sep. 2018.

[27] S. Amsuess, I. Vujaklija, P. Goebel, A. D. Roche, B. Graimann, O. C. Aszmann, and D. Farina, "Context-Dependent upper limb prosthesis control for natural and robust use," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 24, no. 7, pp. 744–753, Jul. 2016.

[28] S. Lee, B. Jamil, S. Kim, and Y. Choi, "Fabric vest socket with embroidered electrodes for control of myoelectric prosthesis," *Sensors*, vol. 20, no. 4, Feb. 2020.

[29] T. Sun, Q. Hu, J. Libby, and S. F. Atashzar, "Deep heterogeneous dilation of LSTM for Transient-Phase gesture prediction through High-Density electromyography: Towards application in neurorobotics," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2851–2858, Apr. 2022.

[30] Y. Liu, X. Li, L. Yang, G. Bian, and H. Yu, "A CNN-Transformer hybrid recognition approach for sEMG-Based dynamic gesture prediction," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–16, 2023.

[31] M. Stachaczyk, S. F. Atashzar, and D. Farina, "Adaptive spatial filtering of High-Density EMG for reducing the influence of noise and artefacts in myoelectric control," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 7, pp. 1511–1517, Jul. 2020.

[32] A. Boschmann and M. Platzner, "Towards robust HD EMG pattern recognition: Reducing electrode displacement effect using structural similarity," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug. 2014, pp. 4547–4550.

[33] A. Moin, A. Zhou, A. Rahimi, S. Benatti, A. Menon, S. Tamakloe, J. Ting, N. Yamamoto, Y. Khan, F. Burghardt, L. Benini, A. C. Arias, and J. M. Rabaey, "An EMG gesture recognition system with flexible High-Density sensors and Brain-Inspired High-Dimensional classifier," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2018, pp. 1–5.

[34] A. Phinyomark, E. Campbell, and E. Scheme, "Surface electromyography (EMG) signal processing, classification, and practical considerations," in *Biomedical Signal Processing: Advances in Theory, Algorithms and Applications*, G. Naik, Ed. Singapore: Springer Singapore, 2020, pp. 3–29.

[35] N. Jiang, C. Chen, J. He, J. Meng, L. Pan, S. Su, and X. Zhu, "Bio-robotics research for non-invasive myoelectric neural interfaces for upper-limb prosthetic control: a 10-year perspective review," *Natl Sci Rev*, vol. 10, no. 5, p. nwad048, May 2023.

[36] A. Prakash and S. Sharma, "A low-cost transradial prosthesis controlled by the intention of muscular contraction," *Australas. Phys. Eng. Sci. Med.*, vol. 44, no. 1, pp. 229–241, Mar. 2021.

[37] B. Hudgins, P. Parker, and R. N. Scott, "A new strategy for multifunction myoelectric control," *IEEE Trans. Biomed. Eng.*, vol. 40, no. 1, pp. 82–94, Jan. 1993.

[38] Y. Gu, D. Yang, Q. Huang, W. Yang, and H. Liu, "Robust EMG pattern recognition in the presence of confounding factors," *Expert Syst. Appl.*, vol. 96, no. C, pp. 208–217, Apr. 2018.

[39] S. Jiang, B. Lv, W. Guo, C. Zhang, H. Wang, X. Sheng, and P. B. Shull, "Feasibility of wrist-worn, real-time hand, and surface gesture recognition via semg and imu sensing," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 8, pp. 3376–3385, 2018.

[40] F. S. Botros, A. Phinyomark, and E. J. Scheme, "Day-to-day stability of wrist emg for wearable-based hand gesture recognition," *IEEE Access*, vol. 10, pp. 125 942–125 954, 2022.

[41] M. Höhne, h.-j. Hwang, S. Amsuss, J. Hahne, D. Farina, and K.-R. Müller, "Improving the robustness of myoelectric pattern recognition for upper limb prostheses by covariate shift adaptation," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 24, pp. 1–1, 01 2015.

[42] M. Zia Ur Rehman, A. Waris, S. O. Gilani, M. Jochumsen, I. K. Niazi, M. Jamil, D. Farina, and E. N. Kamavuako, "Multiday EMG-Based classification of hand motions with deep learning techniques," *Sensors*, vol. 18, no. 8, Aug. 2018.

[43] D. Wu, J. Yang, and M. Sawan, "Transfer learning on electromyography (EMG) tasks: Approaches and beyond," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 3015–3034, Jul. 2023.

[44] X. Jiang, X. Liu, J. Fan, X. Ye, C. Dai, E. A. Clancy, M. Akay, and W. Chen, "Open access dataset, toolbox and benchmark processing

results of high-density surface electromyogram recordings," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 1035–1046, 2021.

[45] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv [cs.CV]*, Oct. 2020.

[46] *ViViT: A Video Vision Transformer*, Mar. 2021.

[47] M. Montazerin, E. Rahimian, F. Naderkhani, S. F. Atashzar, S. Yanushkevich, and A. Mohammadi, "Transformer-based hand gesture recognition from instantaneous to fused neural decomposition of high-density EMG signals," *Sci. Rep.*, vol. 13, no. 1, p. 11000, Jul. 2023.

[48] *Attention Is All You Need*, Jun. 2017.

[49] F. Quivira, T. Koike-Akino, Y. Wang, and D. Erdogmus, "Translating semg signals to continuous hand poses using recurrent neural networks," in *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, 2018, pp. 166–169.

[50] T. Sun, J. Libby, J. Rizzo, and S. F. Atashzar, "Deep augmentation for electrode shift compensation in transient high-density semg: Towards application in neurorobotics," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 6148–6153.

[51] J. Chung, Ç. Gülçehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *CoRR*, vol. abs/1412.3555, 2014. [Online]. Available: http://arxiv.org/abs/1412.3555

[52] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 11 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[53] C. S. M. Castillo, S. Wilson, R. Vaidyanathan, and S. F. Atashzar, "Wearable MMG-Plus-One armband: Evaluation of normal force on mechanomyography (MMG) to enhance Human-Machine interfacing," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 196–205, Feb. 2021.

[54] C. S. M. Castillo, R. Vaidyanathan, and S. F. Atashzar, "Synergistic Upper-Limb functional muscle connectivity using acoustic mechanomyography," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 8, pp. 2569–2580, Aug. 2022.

[55] C. S. Mancero Castillo, S. F. Atashzar, and R. Vaidyanathan, "3D muscle networks based on vibrational mechanomyography," *J. Neural Eng.*, vol. 20, no. 6, Nov. 2023.

[56] T. Sun, J. Libby, J. Rizzo, and S. Farokh Atashzar, "Deep augmentation for electrode shift compensation in transient high-density sEMG: Towards application in neurorobotics," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Oct. 2022, pp. 6148–6153.

[57] E. Tyacke, K. Gupta, J. Patel, R. Katoch, and S. Farokh Atashzar, "From unstable contacts to stable control: A deep learning paradigm for HD-sEMG in neurorobotics," Sep. 2023.