Alignment-Enhanced Interactive Fusion Model for Complete and Incomplete Multimodal Hand Gesture Recognition

Shengcai Duan, *Graduate Student Member, IEEE*, Le Wu, *Member, IEEE*, Aiping Liu, *Member, IEEE*, and Xun Chen, *Senior Member, IEEE*

***Abstract*— Hand gesture recognition (HGR) based on surface electromyogram (sEMG) and Accelerometer (ACC) signals is increasingly attractive where fusion strategies are crucial for performance and remain challenging. Currently, neural network-based fusion methods have gained superior performance. Nevertheless, these methods typically fuse sEMG and ACC either in the early or late stages, overlooking the integration of entire cross-modal hierarchical information within each individual hidden layer, thus inducing inefficient inter-modal fusion. To this end, we propose a novel Alignment-Enhanced Interactive Fusion (AiFusion) model, which achieves effective fusion via a progressive hierarchical fusion strategy. Notably, AiFusion can flexibly perform both complete and incomplete multimodal HGR. Specifically, AiFusion contains two unimodal branches and a cascaded transformer-based multimodal fusion branch. The fusion branch is first designed to adequately characterize modality-interactive knowledge by adaptively capturing inter-modal similarity and fusing hierarchical features from all branches layer by layer. Then, the modality-interactive knowledge is aligned with that of unimodality using cross-modal supervised contrastive learning and online distillation from embedding and probability spaces respectively. These alignments further promote fusion quality and refine modality-specific representations. Finally, the recognition outcomes are set to be determined by available modalities, thus contributing to handling the incomplete multimodal HGR problem, which is frequently encountered in real-world scenarios. Experimental results on five public datasets demonstrate that AiFusion outperforms most state-of-the-art benchmarks in complete multimodal HGR. Impressively, it also surpasses the unimodal baselines in the challenging incomplete multimodal HGR. The proposed AiFusion provides a promising solution to realize effective and robust multimodal HGR-based interfaces.**

***Index Terms*— Multimodal fusion, hand gesture recognition, myoelectric control, accelerometer, incomplete multimodal, alignment.**

I. Introduction

HAND gesture recognition (HGR) has become a vital role in an intuitive and practical human-machine interface (HMI). An increasing number of studies are focusing on pursuing the high effectiveness and quite robustness of HGR-based HMI, because these directly affect the acceptance of users and the efficiency of collaboration with external machines [1], [2], [3], [4], [5]. The HGR based on surface electromyogram (sEMG) and Accelerometer (ACC) signal has expressed enormous potential [6], [7], [8], [9] for two reasons. Firstly, the excellent characteristics of sEMG and ACC signal guarantee its feasibility as the medium of HMI [7], [8]. Specifically, sEMG is a non-invasive electrophysiological signal containing rich motor and physiological information [10], [11], [12], [13]. Therefore, sEMG can assist in capturing intrinsic differences among alike gestures. Meanwhile, the inertial measurement units (IMU) signals, represented by ACC signals [14], have the benefit of presenting the kinematic information of gestures and enhancing robustness [8], [15]. Furthermore, sensors' well wearability and low cost for collecting sEMG and ACC have accelerated the development of downstream application tasks [8]. Currently, the sEMG-ACC-based HGR has been applied to prosthetic control [6], sign language interaction [16], [17], and virtual interaction [18] and so on.

The multimodal fusion strategy significantly affects the performance of sEMG-ACC-based HGR. Early fusion and late fusion are two typical fusion strategies that perform modality fusion in the early and late stages, respectively [19]. However, these strategies are unable to sufficiently characterize intra-modal specificity and cross-modal association at the same time [20]. Furthermore, current methods rarely consider the effective fusion between fused features and unimodal features, incurring inadequate exploration of inter-modal knowledge, such as the interaction of cross-modal hierarchical features. It is widely reported that the correlation of modalities can promote the performance of multimodal methods [21]. The effectiveness of sEMG-ACC-based HGR

Manuscript received 21 June 2023; revised 14 October 2023; accepted 17 November 2023. Date of publication 20 November 2023; date of current version 30 November 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2021YFF0501601; in part by the National Natural Science Foundation of China under Grant 32271431, Grant 82272070, and Grant 62301523; and in part by the Joint Fund for Medical Artificial Intelligence under Grant MAI2023C004. *(Corresponding authors: Le Wu; Xun Chen.)*

Shengcai Duan, Le Wu, and Aiping Liu are with the School of Information Science and Technology, University of Science and Technology of China, Hefei, Anhui 230027, China (e-mail: lewu@ustc.edu.cn).

Xun Chen is with the Department of Neurosurgery, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, and the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230001, China (e-mail: xunchen@ustc.edu.cn).

Digital Object Identifier 10.1109/TNSRE.2023.3335101

may be further enhanced when the interactive relationships are deeply explored [19], [22]. In addition, the recognition performance of multimodal HGR is still suffering from poor robustness, especially the problem of missing modality. The problem of incomplete multimodal caused by sensor failures or data corruption is likely to occur in real-world scenarios [20], [23], especially during long periods of exercise. Nevertheless, most existing multimodal methods for sEMG-ACC-based HGR assume that the two modalities are always available during testing [6], [16], [24]. Consequently, the problem of incomplete modality in testing samples tends to dramatically degrade the performance. Therefore, it is necessary for sEMG-ACC-based HGR to improve its robustness of missing modality, while maintaining the satisfactory performance of complete multimodal HGR.

In this study, aiming to simultaneously pursue the high effectiveness and quite robustness of sEMG-ACC-based HGR, a novel alignment-enhanced interactive fusion model is proposed, termed AiFusion. It innovatively designs a cascaded transformer-based progressive hierarchical fusion strategy to achieve effective fusion. Specifically, AiFusion mainly contains two unimodal branches and a multimodal fusion branch. The transformer-based fusion branch captures intricate cross-modal interactions by progressively integrating the hierarchical features from multiple branches. In this way, the model can not only adaptively learn the inter-modal knowledge with the help of the multi-head attention mechanism of transformers, but also explore the hierarchical relationships across multimodal features through the progressive fusion strategy. After that, to further improve fusion quality and modality-specific representations, the modality-interactive knowledge is aligned with that of the unimodality using cross-modal supervised contrastive learning and online distillation in the embedding and probability spaces. Based on the scalability of the three branches, the AiFusion is extended to perform the incomplete multimodal without additional training burden, where the recognition outcomes are set to be determined by available modalities. The extended experiments are completed on five public multimodal datasets of HGR. These datasets include multimodal gesture data for healthy individuals and trans-radial amputees, with up to 50 types of gestures.

In summary, the major contributions of this study are concluded as follows:

- An interactive fusion model based on a progressive hierarchical fusion strategy is proposed for sEMG-ACC-based HGR. It utilizes cascaded transformers to explore inter-modal knowledge by gradually fusing cross-modal hierarchical features, facilitating effective modality fusion.
- Cross-modal supervised contrastive learning and online distillation are utilized to align the interactive-knowledge-boosted integrated features and unimodal features from embedding and probability spaces, thus further enhancing the modality fusion and unimodal representations.
- AiFusion is further enabled to perform incomplete multimodal HGR that is challenging and frequently occurs in real-world scenarios. To the best of our knowledge, this is the first work that investigates the complete and incomplete multimodal HGR in a unified model.
- The extended experiments are completed on five public multimodal datasets. The proposed AiFusion outperforms most state-of-the-art benchmarks in complete multimodal HGR and also surpasses unimodal baselines in the challenging area of incomplete multimodal HGR.

The rest of the paper is organized as follows. Section II demonstrates the related work about fusion strategies and incomplete multimodal learning. Then, the problem formulation and the proposed AiFusion model are presented in section III. The experiments and results are included in section IV. A further discussion is exhibited in section V. Finally, section VI concludes our work.

II. Related Work

In this section, the representative fusion strategies for multimodal HGR in existing works are first introduced. Then, the incomplete multimodal learning is presented.

A. Fusion Strategy in Multimodal HGR

The classical myoelectric pattern recognition methods, including linear discriminant analysis (LDA) [15] and support vector machine (SVM) [25] are also utilized to perform sEMG-ACC-based HGR and obtained a better multimodal recognition baseline than unimodal recognition. Gijsberts et al. [26] adopted Kernel Regularized Least Squares (KRLS) algorithm to fuse sEMG and ACC and achieved the accuracy of 82.49% for 40-type hand gestures recognition. All these traditional methods utilized the concatenation of handcrafted features derived from sEMG and ACC as the input of the algorithm, which is a typical kind of early fusion manners. Recently, numerous studies have endeavored to take advantage of deep learning to complete sEMG-ACC-based HGR [6], [7], [8], which present the more powerful ability of knowledge learning and feature presentation. The classical multi-view learning with convolution neural networks (MVCNN) [24] was adopted to complete sEMG-IMU-based HGR. The MVCNN regarded each feature set of a modal signal as a view of hand gestures. The classification results of all views characterized by parallel convolution networks (CNN) were handled with a decision-level fusion approach, which is a classical manner of late fusion. Therefore, the MVCNN made full use of unimodal information, yet the interactive knowledge of multimodal signals was undervalued. Zhang et al. [16] proposed the SeeSign network for multimodal sign language recognition. The SeeSign utilized different attention mechanisms to integrate the deep features of sEMG and IMU extracted by CNN and long short-term memory (LSTM). Although the correlation between sEMG and IMU is considered with the attention mechanism in SeeSign, the hierarchical relation in deep features was absent. The hybrid multimodal fusion (HyFusion) [27] simultaneously acquired the intra-modal and inter-modal knowledge with parallel branches, yielding the state-of-the-art results on multiple public multimodal datasets of HGR. The multiple fusion strategies were equipped in HyFusion, but the inter-modal knowledge among unimodal and integrated signals still can be further explored.

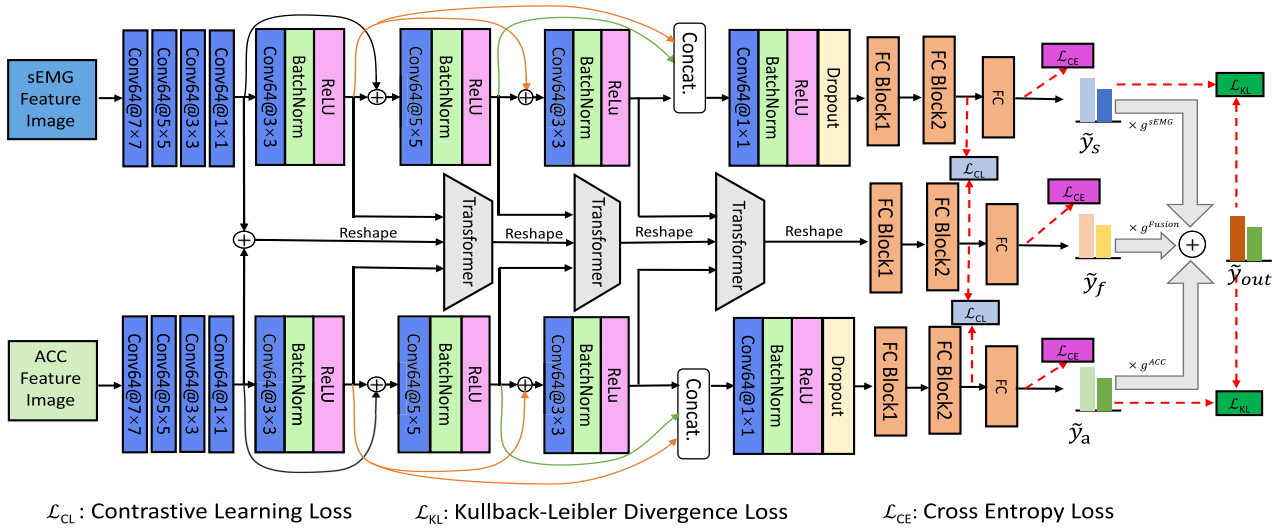


Fig. 1. The proposed Alignment-enhanced interactive Fusion model (AiFusion) for complete and incomplete multimodal HGR. Transformer is leveraged in the multimodal fusion branch and progressively fuses hierarchical multimodal features. Conv, BatchNorm, ReLU and Dropout indicate the layer of convolution, batch normalization, rectified linear unit and dropout layer, respectively. FC Block and FC denote the full connection block and classifier layer, respectively. The number after the Conv layer name is the number of filters. The number after @ denotes the convolution kernel shape. Concat refers to the concatenation of hierarchical unimodal features by channel.

B. Incomplete Multimodal Learning

Incomplete multimodal learning aims to address the missing modality problem in multimodal tasks, which is a classical problem and has attracted increasing attention recently [20], [28], [29]. The incomplete multimodal learning methods generally fall into two categories: generation-based methods and non-generation-based methods. Ma et al. [28] adopted a generation subnetwork to investigate multimodal learning with severely missing modality (SMIL) for image-text-based classification. Although the SMIL model alleviates the problem of incomplete modalities to some extent, the additional generative module also adds the burden of memory and computation for the multimodal fusion model. Recently, Ma et al. [29] published another work based on a novel transformer-based model with multi-task optimization, rather than generative methods to deal with the incomplete modalities problem. This model [29] can only guarantee that the performance was not worse than the unimodal one, but can not achieve satisfactory performance in complete multimodal recognition.

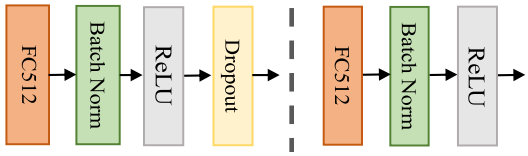
However, the challenging and frequent problem of incomplete modalities is rarely addressed in existing multimodal HGR. The incomplete multimodal HGR will be preliminarily explored in this study.

III. Methodology

In this section, we first present the sEMG-ACC-based HGR problem definition from the aspect of the presence of modality. Then, the proposed alignment-enhanced interactive fusion (AiFusion) model is exhibited.

A. Problem Formulation

This study aims to propose an effective and robust fusion model to accomplish sEMG-ACC-based HGR in complete and incomplete multimodal scenarios. We investigate multimodal HGR with two modalities, i.e., sEMG and ACC. Formally, $\mathcal{D}^f = \{\boldsymbol{x}_i^1, \boldsymbol{x}_i^2, y_i\}$ denotes the complete multimodal dataset



(a) FC Block1 (b) FC Block2

Fig. 2. The components of FC Block1 and Block2. The number after FC layer name denotes the number of neurons in the hidden layer.

with full modalities, where $\boldsymbol{x}_i^1$ and $\boldsymbol{x}_i^2$ represent the two modalities of i-th sample and y_i denotes the corresponding class label; $\mathcal{D}^p = \{\mathcal{D}^1, \mathcal{D}^2\}$ denotes the incomplete multimodal datasets with partial modalities, where $\mathcal{D}^1 = \{\boldsymbol{x}_i^1, y_i\}$ and $\mathcal{D}^2 = \{\boldsymbol{x}_i^2, y_i\}$ denote the datasets of missing $\boldsymbol{x}^2$ modality and missing $\boldsymbol{x}^1$ modality respectively. Specifically, our target is to obtain a projection function $\mathcal{F}$ which is trained with the complete dataset and can effectively and robustly classify complete or incomplete multimodal hand gestures in the testing phase.

B. Proposed Approach

In this work, a novel alignment-enhanced interactive fusion model (AiFusion) is presented. Firstly, a cascaded transformer-based progressive hierarchical fusion strategy is proposed. Then, the alignments using cross-modal supervised contrastive learning and online distillation are elaborated. At last, the optimization strategy and post-processing are presented.

1) Progressive Hierarchical Fusion Strategy: To effectively integrate the modalities and extract the interactive knowledge, the progressive hierarchical fusion strategy using transformers is designed. As illustrated in Fig. 1, the AiFusion model mainly contains two unimodal convolution branches (sEMG-CNN and ACC-CNN) and a multimodal fusion transformer branch (fusion-Transformer).

In AiFusion, the inputs are the handcrafted feature images of sEMG and ACC, which would be presented in IV-B. These inputs are first expanded to 64 channels by multiscale

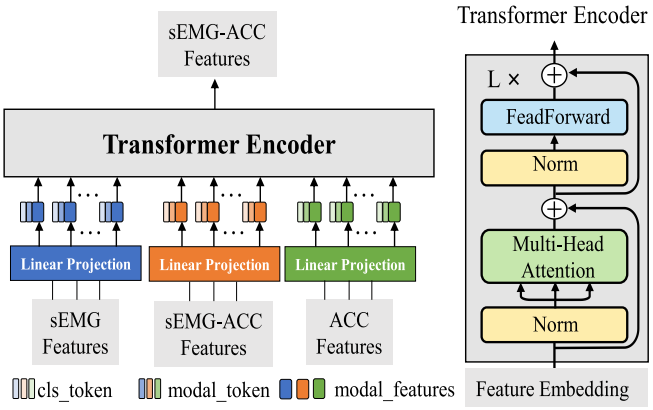


Fig. 3. The overview of multimodal Transformer. Norm denotes layer normalization. Feedforward contains the sequential layers: full connection, Gaussian Error Linear Unit, dropout, full connection and dropout.

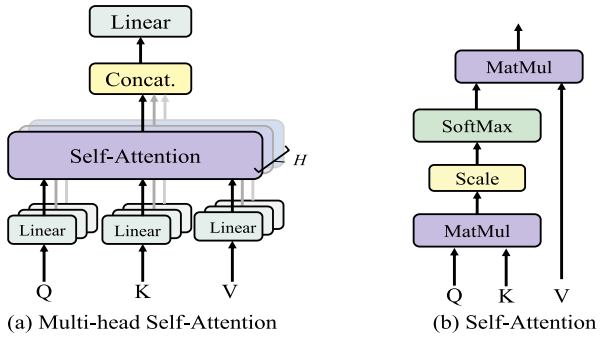


Fig. 4. The structure of multi-head self-attention.

convolution. In unimodal branches, the sEMG-CNN and ACC-CNN utilize convolution with residual skip to extract hierarchical features of sEMG and ACC, respectively. Then, the three-level hierarchical unimodal features are concatenated in channel dimension and merged by local convolutions. The local convolution, i.e., the convolution operations with 1×1 filter, is widely adopted in the model for HGR [24]. The local convolution contributes to extracting the cross-channel knowledge, which consists of different hierarchical unimodal features. The local convolution also reduces the dimension of the feature. Finally, the unimodal hybrid hierarchical features are put into the corresponding full connection blocks and classifiers. The full connection blocks are shown in Fig. 2. As a result, two unimodal branches independently represent and classify the unimodal features images and provide the corresponding classification scores $\tilde{\boldsymbol{y}}_s$ and $\tilde{\boldsymbol{y}}_a$, respectively.

In the fusion branch, the transformer is used to progressively integrate the hierarchical features across modalities, consisting of the sEMG modal, ACC modal and hybrid modal (sEMG-ACC). The fusion-Transformer contains a three-level transformer, FC Blocks and a classifier, shown in Fig. 1. The transformer of overview is presented in Fig. 3. Specifically, the hierarchical features of unimodal (sEMG features $\boldsymbol{x}^s$ and ACC features $\boldsymbol{x}^a$) and multimodal (sEMG-ACC features $\boldsymbol{x}^f$) are first input into the linear projection layer ($\boldsymbol{E}$), respectively. It is worth noting that we abandon the operation of image segmentation like in the vision transformer (ViT) [31], which significantly reduces the number of parameters and guarantees the complete semantic information of hand gestures. Then, not only the learnable parameters of class tokens (clc_token, $\boldsymbol{x}_{cls}$), but also the learnable parameters of corresponding modality tokens (modal_token, $\boldsymbol{x}_S, \boldsymbol{x}_A, \boldsymbol{x}_F,$) are experimentally attached to improve the representation ability of embedding features. In addition, positional sinusoidal embeddings ($\boldsymbol{E}^{pos}$) are also added to the feature embedding to retain relative positional information of various modalities, and then these feature embeddings are concatenated in channel dimension. The above steps provide the input ($\boldsymbol{Z}_0$) to the transformer encoder expressed in Eq. 1.

Thereafter, the transformer encoder is utilized to fuse cross-modal hierarchical features. Similar to ViT, a layer normalization, multi-head self-attention units (MSA), a layer normalization and a feedforward module are used in sequence in a one-layer encoder. The number of layers for a transformer encoder is set as a hyperparameter L. The MSA is assigned for the attention mechanism [32] while the feedforward acts as a multilayer perception. The MSA is calculated as in Eq. 2. The input sequence $\boldsymbol{Z}_0$ is projected $\boldsymbol{W}_{qkv}$ to identical-dimension Keys (K), Queries (Q) and Values (V). In MSA block, shown in Fig. 4, there are H number of identical heads with distinct learnable parameters operating parallelly. Therefore, there are H units of self-attention (SA), resulting in an attention matrix indicating the similarity between each token. The outputs of the scaled dot-product attention are concatenated and transferred to the linear layers $\boldsymbol{W}_{msa}$ expressed in Eq. 2.

$$\begin{aligned}\boldsymbol{Z}_0 &= [\boldsymbol{x}_{cls}; \boldsymbol{x}_S; \boldsymbol{x}^s\boldsymbol{E}; \boldsymbol{x}_{cls}; \boldsymbol{x}_A; \boldsymbol{x}^a\boldsymbol{E}; \boldsymbol{x}_{cls}; \boldsymbol{x}_F; \boldsymbol{x}^f\boldsymbol{E}] \\ &\quad + \boldsymbol{E}^{pos} \end{aligned} \tag{1}$$

$$\begin{aligned}[Q, K, V] &= \boldsymbol{Z}_0\boldsymbol{W}_{qkv} \\ A &= softmax(\frac{QK^T}{\sqrt{d_h}}) \\ SA(Z) &= AV \\ MSA(Z) &= [SA_1(Z); SA_2(Z); \ldots; SA_H(Z)]\boldsymbol{W}_{msa} \end{aligned} \tag{2}$$

where d_h is the scale coefficient. Finally, the outputs of MSA are input into the feedforward part. The sEMG-ACC features merged by the three-level transformer are passed to the FC Blocks and a classifier. Consequently, the fusion-transformer branch utilizing progressive hierarchical features provides the corresponding classification results $\boldsymbol{y}_f$. The decision fusion manner adopts the summation and the final classification score equals to $\boldsymbol{y}_{out} = \boldsymbol{y}_f + \boldsymbol{y}_a + \boldsymbol{y}_s$.

2) Improvement With Alignment: Multimodal alignment is one of the key challenges for multimodal classification tasks [20]. In AiFusion, to further improve the fusion quality, the alignments between modality-interactive knowledge and unimodal modality are adopted with cross-modal supervised contrastive learning and online distillation in embedding space and probability space, respectively.

In embedding space, cross-modal supervised contrastive learning is utilized to pull together the clusters of cross-modal samples belonging to the same class, while simultaneously pushing apart clusters of cross-modal features from different classes. The utilized cross-modal supervised contrastive

learning loss $\mathcal{L}_{CL}$ [35] can be expressed as:

$$\mathcal{L}_{CL} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(z_i \cdot z_p / \tau\right)}{\sum_{a \in A(i)} \exp\left(z_i \cdot z_a / \tau\right)} \tag{3}$$

Here, within a batch of N samples, $i \in I \equiv \{1, \ldots, 2N\}$ is the index of an arbitrary sample. The cross-modal samples contain N unimodal-feature (sEMG Feature or ACC Feature) samples from unimodal branches and N corresponding hybrid-feature (sEMG-ACC Feature) samples from fusion-transformer branch. $A(i) \equiv I \backslash \{i\}$. $P(i) \equiv \{p \in A(i) : \boldsymbol{y}(p) = \boldsymbol{y}(i)\}$ is the set of indices of all positives in the batch distinct from i, and $|P(i)|$ is its cardinality. $\boldsymbol{y}(p)$ and $\boldsymbol{y}(i)$ is the one-hot label of p-th and i-th sample respectively. $\tau \in \mathcal{R}^+$ is a scalar temperature parameter.

The cross-modal supervised contrastive learning loss $\mathcal{L}_{CL}$ makes full use of label information. The $\mathcal{L}_{CL}$ regards the samples with the same label as the positive sample and other samples as negative samples. Therefore, the $\mathcal{L}_{CL}$ not only can preserve the ability of original contrastive learning [33], [34], but also gain the generation ability to an arbitrary number of positives [35]. Specifically, as seen in Fig. 1, the $\mathcal{L}_{CL}$ contrasts the view of unimodal features (sEMG or ACC) and multimodal features (sEMG-ACC) contributing to aligning the samples with the same label.

In probability space, online distillation based on Kullback-Leibler Divergence Loss is used to align unimodal and multimodal feature distribution. The Kullback-Leibler Divergence Loss $\mathcal{L}_{KL}$ evaluates the distribution difference of unimodal classification scores ($\tilde{\boldsymbol{y}}_s$ or $\tilde{\boldsymbol{y}}_a$) and final fusion classification score $\tilde{\boldsymbol{y}}_{out}$:

$$\begin{aligned} \mathcal{L}_{KL} &= \sum_{i=1}^{N} \mathcal{L}\left(\tilde{\boldsymbol{y}}_m(i) \,||\, \tilde{\boldsymbol{y}}_{out}(i)\right) \\ &= \sum_{i=1}^{N} \sum_{j=1}^{C} \tilde{\boldsymbol{y}}_m(i,j) \log \frac{\tilde{\boldsymbol{y}}_m(i,j)}{\tilde{\boldsymbol{y}}_{out}(i,j)} \end{aligned} \tag{4}$$

In Eq. 4, N is the number of samples in a batch; C is the classes of hand gestures; $\tilde{\boldsymbol{y}}_m \in \{\tilde{\boldsymbol{y}}_a, \tilde{\boldsymbol{y}}_s\}$; $\tilde{\boldsymbol{y}}_m(i,j)$ and $\tilde{\boldsymbol{y}}_{out}(i,j)$ represents the j-th value in the classfication score of i-th sample for $\tilde{\boldsymbol{y}}_m$ and $\tilde{\boldsymbol{y}}_{out}$, respectively. In AiFusion, the richer information of multimodal distribution is distilled and transported to unimodal distribution through the Kullback-Leibler Divergence Loss $\mathcal{L}_{KL}$. Thereby, the representation capability of unimodal, especially the weak unimodal (i.e., sEMG), is boosted.

3) Multitask Optimization and Post-Processing: The three parallel branches in AiFusion consisting of sEMG-CNN, ACC-CNN and fusion-Transformer can be seen as three tasks. Three branches provide the results of hand gesture classification from the corresponding modality perspective. Therefore, the proposed AiFusion can be trained with multitask optimization strategy. The cross-entropy loss $\mathcal{L}_{CE}$ is applied to the optimization of each branch:

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} \boldsymbol{y}(i,j) \log \tilde{\boldsymbol{y}}(i,j) \tag{5}$$

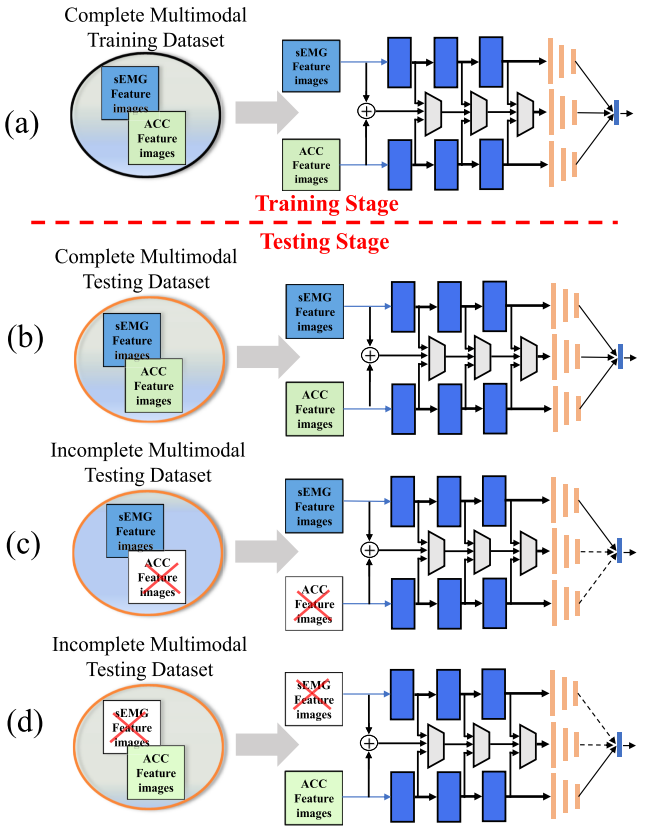


Fig. 5. Scheme of training and testing stages for AiFusion. (a)AiFusion is trained with complete multimodal training datasets. (b), (c) and (d) The trained AiFusion model is tested with complete and incomplete multimodal testing datasets, respectively. The white feature image with a red cross represents that with missing modality. The classification scores corresponding to the branch of the dotted line will be discarded.

In Eq. 5, N is the size of samples in a batch and C represents the number of categories of hand gestures. $\boldsymbol{y}(i,j)$ is a binary label: it equals 1 when the i-th sample in the batch truly belongs to class j, otherwise it equals 0. $\tilde{\boldsymbol{y}}(i,j)$ is the predicted probability that the i-th sample belongs to class j. It is worth noting that the proposed AiFusion model is trained with complete multimodal training datasets. As a result, as shown in Fig. 1, the summation of all loss functions for the AiFusion can be expressed as:

$$\mathcal{L}_{sum} = \sum_{j=1}^{3} \mathcal{L}_{CE_j} + \sum_{i=1}^{2} \left(\mathcal{L}_{CL_i} + \mathcal{L}_{KL_i}\right) \tag{6}$$

The parallel architecture of the AiFusion and multi-task optimization training method provide three benefits for multimodal HGR. First, the ensemble results of the three branches can provide better multimodal fusion recognition performance. The second advantage is to improve unimodal representation capability. At last, it provides great scalability of the model, which facilitates the handling of missing modalities. Furthermore, a simple and effective post-processing mechanism is utilized to perform the incomplete multimodal HGR. Specifically, as shown in Fig. 5, the post-processing mechanism can

TABLE I
SPECIFICATIONS OF THE SELECTED DATASETS EVALUATED IN THIS RESEARCH

Name	Total number of gestures	Number of gestures to be classified	Intact subjects	Amputed subjects	Number of sEMG channels	Number of ACC channels	Number of trials	Trials for training	Trials for testing	Sampling rate
Ninapro DB2 [36]	50	50	40	0	12	36	6	1,3,4,6	2,5	2000Hz
Ninapro DB3 [36]	50	50	0	11	12	36	6	1,3,4,6	2,5	2000Hz
Ninapro DB5 [37]	41	41	10	0	16	3	6	1,3,4,6	2,5	2000Hz
Ninapro DB6 [38]	8	8	10	0	14	42	120	1,3,...,119	2,4,...,120	2000Hz
Ninapro DB7 [15]	41	41	20	2	12	36	6	1,3,4,6	2,5	2000Hz

be expressed as:

$$\{\boldsymbol{g}^{sEMG}, \boldsymbol{g}^{Fusion}, \boldsymbol{g}^{ACC}\} = \begin{cases} \{1,0,0\}, & \boldsymbol{x}\,(i) = \{\boldsymbol{x_i}^1, \boldsymbol{0}\} \\ \{0,0,1\}, & \boldsymbol{x}\,(i) = \{\boldsymbol{0}, \boldsymbol{x_i}^2\} \\ \{1,1,1\}, & \boldsymbol{x}\,(i) = \{\boldsymbol{x_i}^1, \boldsymbol{x_i}^2\} \end{cases}. \tag{7}$$

As shown in Eq. 7, $\boldsymbol{g}^{sEMG}$, $\boldsymbol{g}^{Fusion}$ and $\boldsymbol{g}^{ACC}$ represent the multiplication factors for classification score of the corresponding sample from sEMG-CNN, fusion-Transformer and ACC-CNN branch, respectively. $\boldsymbol{x_i} = \{\boldsymbol{x_i}^1, \boldsymbol{x_i}^2\}$ indicates the complete multimodal sample. $\boldsymbol{x_i} = \{\boldsymbol{x_i}^1, \boldsymbol{0}\}$ and $\boldsymbol{x_i} = \{\boldsymbol{0}, \boldsymbol{x_i}^2\}$ represents the incomplete multimodal sample, missing the information of ACC and sEMG, respectively. It is worth noting that the absent modality information is handled by zero-padding [30]. Accordingly, when the input is the complete multimodal sample, the final classification score is set as the summation of three scores from three branches. When only one modality is existing, the final score equals the classification score of the corresponding branch with the present modality.

IV. EXPERIMENTS AND RESULTS

In this section, the selected datasets and experimental setup are first demonstrated. Then, the designed experiments of complete multimodal HGR are exhibited and its results are compared with existing advanced methods. Then, to evaluate the capability to handle missing modality, the experiments of incomplete multimodal HGR are conducted and the results are compared with the unimodal baseline method.

A. Datasets

To evaluate the proposed AiFusion model, five public datasets, namely Ninapro DB2, DB3, DB5, DB6, and DB7, are adopted to perform the complete and incomplete multimodal HGR experiments. The Ninapro database are publicly available [15], [36], [37], [38] and are widely utilized to help the research of prosthetic hand systems and validation study of decoding algorithms. The sEMG and ACC signals are synchronously collected with sparse channels. The specifications of these datasets are presented in Table. I.

In Ninapro DB2 [36], there are 50-class hand gestures of sEMG and ACC signals. The data were all collected from 40 healthy subjects. Six trials (repeated six times) was composed of a gesture; the duration time of a trial was 5 seconds and the rest time between adjacent trials was set to 3 seconds. The sEMG and ACC signals were synchronously collected with 12 Delsys Trigno Wireless electrodes.

In Ninapro DB3 [36], the number of gestures and the paradigm of collecting data was the same with Ninapro DB2. Notably, the 11 subjects are amputees with transradial amputation. As same as previous studies [24], [27], the data from three amputees with fewer gestures and two amputees with missing electrodes are excluded.

In Ninapro DB5 [37], 10 intact subjects performed 41-class hand gestures. The signals are collected with two Thalmic Myo Armbands at 200Hz. Notably, there are 16-channel sEMG signals and only one-channel ACC signals. To provide sufficient samples for deep learning training, the data are upsampled to 2000Hz in this study.

In Ninapro DB6 [38], the synchronized sEMG and ACC signals were collected for 8-class hand gestures from 10 intact subjects. However, the experimental paradigm is different. The subjects were asked to repeat 7 grasps 12 times, twice a day for 5 days.

In Ninapro DB7 [15], there are 41-type hand gestures of sEMG and IMU (including accelerometers (ACC), magnetometers (MAG) and gyroscopes (GYR)) signals [27]. There are 20 intact subjects and two amputees with transradial amputation. In our work, the data of amputees was excluded because of missing electrodes.

B. Experimental Setup

The experimental setup of this study follows the classical myoelectric control paradigm [39], including signal preprocessing, feature extraction and classification. Therefore, data preprocessing and feature extraction are first demonstrated. Then, the preparation of the training and testing datasets are described. Finally, the training and testing paradigm and the evaluation metric are presented. All experiments are completed with an NVIDIA GeForce GTX 3080 GPU in Pytorch.

1) Data Preprocessing and Feature Extraction: First, the serial signals of sEMG and ACC of each subject are preprocessed respectively. To get raw data samples, the operation of window segmentation and normalization are sequentially carried out on sEMG and ACC signals, which is the same as previous studies of gesture recognition [27]. The length of each window is fixed to 200 ms and the step time is set as 10 ms [24], [27]. The myoelectric control application [41] requires that the length of a sliding window should be within 300 ms. Therefore, a window length of 200 ms is reasonable for realizing myoelectric control. The implemented normalization of raw data samples can be expressed as:

$$\tilde{x}_m(k, ch) = \frac{x_m(k, ch) - \bar{x}_m(ch)}{\sigma\,(x_m(ch))} \tag{8}$$

where $m = \{sEMG, ACC\}$; k is the index for all the samples; ch represents the number of channels. $\tilde{x}_m(k, ch)$

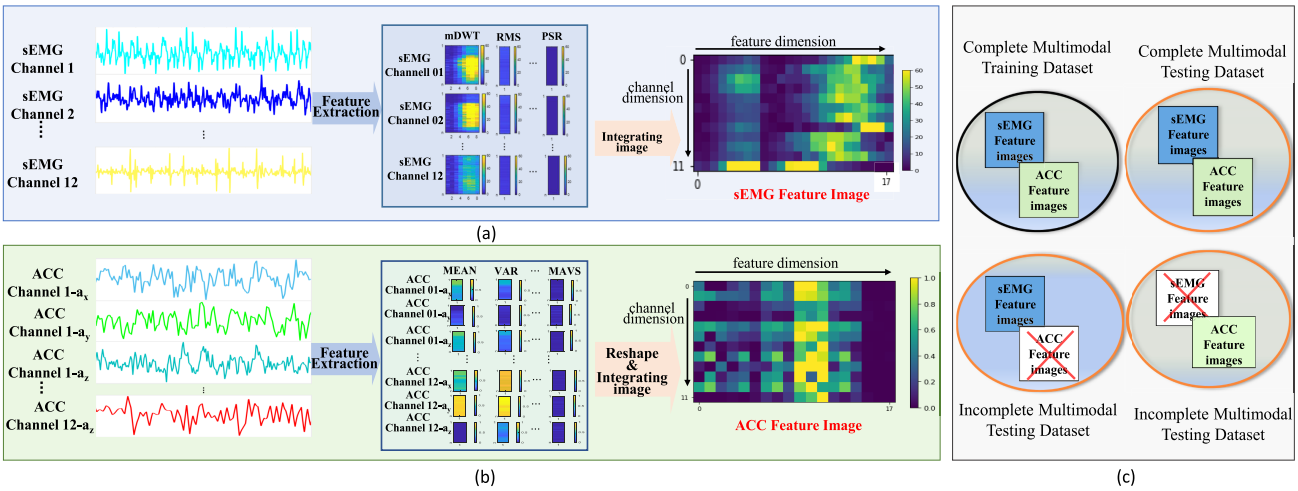


Fig. 6. The process of preparing complete and incomplete multimodal datasets. (a)The process of obtaining sEMG feature images from sEMG signals. (b)The process of obtaining ACC feature images from ACC signals. (c)The complete and incomplete multimodal datasets. The white feature images with red cross represent missing modality and they are handled with zero padding in this study.

TABLE II
SPECIFICATIONS OF HANDCRAFTED FEATURES FOR sEMG AND ACC SIGNALS

sEMG Features	Dimensions × Channels	ACC Features	Dimensions × Channels
mDWT	8×12	MEAN	1×36
SSC	1×12	VAR	1×36
WL	1×12	RMS	1×36
MAV	1×12	WL	1×36
WAMP	1×12	MAV	1×36
ARC	4×12	MAVS	1×36
MNF	1×12		
PSR	1×12		
Total	18×12	Total	6×36

is the normalized raw data sample. $\bar{x}_m(ch)$ and $\sigma(x_m(ch))$ indicate the mean value and standard deviation value for each channel of all training data.

The utilized handcrafted features for sEMG in our study are classical in the time domain and frequency domain, including a part of features from [36], [39], and [40]. Specific handcrafted features contain Marginal of Discrete Wavelet Transform (mDWT), Slope Sign Change (SSC), Waveform Length (WL), Mean Absolute Value (MAV), Willison Amplitude (WAMP), Autoregressive Coefficients (ARC), Mean Frequency (MNF) and Power Spectrum Ratio (PSR). The features for ACC are commonly used features in the time domain, including mean (MEAN), variance (VAR), RMS, WL, MAV, MAVS [27]. The similar features for serial ACC signal were also used in the previous study [22], [27]. The specifications of handcrafted features in this research are listed in Table. II. Furthermore, to make full use of the information of sEMG and ACC in the time and space domain, these handcrafted features are arranged into feature images. The process for composing handcrafted features into feature images is shown in Fig. 6 (a) and (b). It should be noted that since there is only one ACC sensor in the DB5, shown in Table. I, the acceleration information will be reused by other channels in order to align the space information of sEMG and ACC in DB5.

2) Complete and Incomplete Multimodal Dataset: To evaluate the capability of the proposed AiFusion to perform HGR in complete and incomplete multimodal conditions, the complete and incomplete multimodal dataset is prepared. As stated in III-A, the samples in the complete multimodal datasets have two modalities (sEMG and ACC). The samples in the incomplete multimodal datasets have only one modality (sEMG or ACC). These training and test datasets are presented in Fig. 6 (c). In this work, the AiFusion model is trained with the complete multimodal training datasets because it can always be achieved during the training phase. And the trained AiFusion model is independently tested with the complete and incomplete multimodal testing datasets. The complete multimodal testing samples are utilized to simulate normal testing scenarios. The incomplete multimodal testing samples are used to simulate the unnormal signal acquisition scenarios where one modality is missing. The part of missing modality is dealt with zero-padding to meet the input format of the model and simplify the detection of samples with missing modality.

3) Evaluation Metric: The setup of training data and testing data is shown in Table. I. The setting is the same as previous studies [24], [27], [36]. The scheme of training and testing stages for AiFusion is demonstrated in Fig. 5. Although the AiFusion is trained with the complete multimodal training datasets, the trained AiFusion model is independently tested with the complete and incomplete multimodal testing datasets. All experiments are completed in an intra-subject paradigm. The recognition accuracy for a subject is calculated as the evaluation in Eq. 9 and the average accuracy is the mean accuracy of overall subjects [24], [27].

$$\text{accuracy} = \frac{\text{number of correct classifying samples}}{\text{total number of samples}} \times 100\% \tag{9}$$

C. Complete Multimodal HGR

In this part, the experiment for complete multimodal HGR with the proposed AiFusion is demonstrated. Furthermore, the performance comparison between the AiFusion model and previous representative methods for sEMG-ACC-based HGR is presented. The complete multimodal HGR experiments

TABLE III
THE AVERAGE ACCURACIES OF COMPLETE MULTIMODAL HGR ACHIEVED BY THE AIFUSION MODEL AND EXISTING STUDIES ON NINAPRO DATASETS. THE RESULTS IN **BOLD** TEXT REPRESENT THE BEST PERFORMANCE

Method	Dataset	Modalities	Number of gestures to be classified	Window length	Accuracy
KRLS [26]	Ninapro DB2	sEMG-ACC	40	400ms	82.49%
Improved KRLS [42]	Ninapro DB2	sEMG-ACC	40	400ms	92.10%†
MV-CNN [24]	Ninapro DB2	sEMG-ACC	50	200ms	94.40%
HyFusion [27]	Ninapro DB2	sEMG-ACC	50	200ms	94.73%
AiFusion	Ninapro DB2	sEMG-ACC	50	200ms	**95.28%****
Improved KRLS [42]	Ninapro DB3	sEMG-ACC	40	400ms	88.90%†
SVM [25]	Ninapro DB3	sEMG-ACC	40	250ms	88.72%
MV-CNN [24]	Ninapro DB3	sEMG-ACC	50	200ms	87.06%
HyFusion [27]	Ninapro DB3	sEMG-ACC	50	200ms	89.60%
AiFusion	Ninapro DB3	sEMG-ACC	50	200ms	**91.11%***
MV-CNN [24]	Ninapro DB5	sEMG-ACC	41	200ms	**91.31%**
HyFusion [27]	Ninapro DB5	sEMG-ACC	41	200ms	86.37%
AiFusion	Ninapro DB5	sEMG-ACC	41	200ms	87.04%**
MV-CNN [24]	Ninapro DB6	sEMG-ACC	7	200ms	77.10%
HyFusion [27]	Ninapro DB6	sEMG-ACC	8	200ms	77.86%
AiFusion	Ninapro DB6	sEMG-ACC	8	200ms	**80.62%****
LDA [15]	Ninapro DB7	sEMG-ACC	40	256ms	82.70%
MV-CNN [24]	Ninapro DB7	sEMG-IMU+	41	200ms	94.54%
HyFusion [27]	Ninapro DB7	sEMG-ACC	41	200ms	96.44%
AiFusion	Ninapro DB7	sEMG-ACC	41	200ms	**96.76%****

† indicates the maximum accuracy reported in the literature.
* Sign * denotes $p<0.05$; sign ** denotes $p<0.01$.
+ IMU consists of ACC, MAG, and GYR. The specifications can be found in Section. IV-A.

are completed on Ninapro DB2, DB3, DB5, DB6, and DB7. The AiFusion is trained with an Adam optimizer, which is especially useful for the model to rapidly achieve convergence state [43]. The learning rate during model training adopts a step-down strategy. The learning rate is 0.001 for the first three epochs during training. Then, the learning rate is divided by 10 after the 6 epochs and 3 epochs, respectively. There are 13 epochs in all, which is determined by empirical experiments. The dropout rate is set to 0.65. The batch size is 512. The experimental setting and evaluation metrics utilized for the selected datasets are detailed in the Section. IV-B. The overview of training and testing stages for AiFusion on complete multimodal HGR is presented in Fig. 5 (a) and (b), respectively.

The current state-of-art work HyFusion [27] and classical work MVCNN [24] are taken as representative works of deep learning to be compared. Because the Ninapro DB5 and DB6 were not evaluated on the HyFusion model in [27], these two datasets are evaluated on HyFusion model in this research, where the experimental setup is the same as the setup in [27]. In addition, the reported results of some traditional methods applied to these selected datasets are also listed. The compared traditional methods include kernel regularized least squares (KRLS) [26], the improved KRLS [42], support vector machine (SVM) [25] and linear discriminant analysis (LDA) [15]. The average accuracy comparison of the proposed AiFusion and existing studies for complete multimodal HGR are listed in Table. III. In the complete modalities HGR experiments, the proposed AiFusion model obtains the average classification accuracy of 95.28%, 91.11%, 87.04%, 80.62% and 96.76% on Ninapro DB2, DB3, DB5, DB6 and DB7, respectively.

TABLE IV
THE AVERAGE ACCURACIES OF INCOMPLETE MULTIMODAL HGR OF THE AIFUSION AND CNN-BASELINE ON NINAPRO DATASETS. THE RESULTS IN ***Bold*** TEXT REPRESENT THE BETTER PERFORMANCE

Dataset	Modality	**AiFusion**	CNN-baseline
Ninapro DB2	sEMG	**79.73%****	79.40%
	ACC	**93.48%****	93.20%
Ninapro DB3	sEMG	**61.45%***	60.96%
	ACC	**88.06%**	87.43%
Ninapro DB5	sEMG	**74.01%****	73.18%
	ACC	**59.87%****	59.21%
Ninapro DB6	sEMG	**69.18%****	68.63%
	ACC	**71.27%****	70.13%
Ninapro DB7	sEMG	**85.23%***	84.95%
	ACC	**94.66%****	94.46%

Sign * denotes $p<0.05$; sign ** denotes $p<0.01$.

In order to demonstrate the superior performance of the proposed AiFusion for the complete HGR, the statistical tests are completed between the accuracies of HyFusion and AiFusion on the three selected datasets with SPSS R26.0.0.0 software. The results of paired t-test demonstrate that the AiFusion obtains significantly better performance compared to the existing studies for complete multimodal HGR.

D. Incomplete Multimodal HGR

The experiments of incomplete multimodal HGR are tested on incomplete multimodal testing datasets, while the model is still trained with complete multimodal training datasets.

TABLE V
THE AVERAGE ACCURACIES OF COMPLETE AND INCOMPLETE MULTIMODAL HGR ACHIEVED BY THE AIFUSION, AIFUSION-CNN AND AIFSUION-NOPRO ON NINAPRO DATASETS. THE RESULTS IN ***Bold*** TEXT REPRESENT THE BETTER PERFORMANCE

Dataset	Modality	**AiFusion**	AiFusion -CNN	AiFusion -noPro
Ninapro DB2	sEMG-ACC	**95.28%**	95.08%**	93.88%**
	sEMG	79.73%	79.63%	**79.80%**
	ACC	**93.48%**	93.35%*	93.20%**
Ninapro DB3	sEMG-ACC	**91.11%**	90.88%*	87.61%**
	sEMG	**61.45%**	60.99%	61.39%
	ACC	**88.06%**	87.80%	87.55%*
Ninapro DB5	sEMG-ACC	**87.04%**	86.80%*	85.81%**
	sEMG	**74.01%**	73.96%	74.00%
	ACC	59.87%	59.81%	**59.94%**
Ninapro DB6	sEMG-ACC	**80.62%**	79.67%*	79.57%**
	sEMG	**69.18%**	69.07%	69.11%
	ACC	71.27%	**71.51%**	70.71%*
Ninapro DB7	sEMG-ACC	**96.76%**	96.57%**	95.84%**
	sEMG	**85.23%**	84.62%**	85.20%
	ACC	**94.66%**	94.43%**	94.63%

Sign * denotes $p<0.05$; sign ** denotes $p<0.01$.

It should be noted that the testing datasets in the incomplete multimodal HGR contain only one modality and the information of absent modality is replaced by zero. In this way, extreme conditions can be simulated where one of the two modalities is completely absent. The overview of the testing stages of AiFusion for incomplete multimodal HGR is demonstrated in Fig. 6 (c) and (d). Furthermore, to present the capability of AiFusion to tackle the incomplete multimodal HGR, a unimodal comparative method, named CNN-baseline, is also implemented to perform hand gesture recognition with unimodal input. To make a fair comparison, the architecture of CNN-baseline is the same as CNN branch of AiFusion. The CNN-baseline is trained and tested with the datasets containing the corresponding existing modality signal. The hyper-parameters of training and testing setup for CNN-baseline are the same as AiFusion. The average accuracy of the proposed AiFusion and CNN-baseline for incomplete multimodal HGR are listed in Table. IV. The average results of AiFusion for incomplete multimodal HGR on all selected datasets are higher than the average results of the unimodal CNN-baseline method on unimodal HGR. The paired t-test between AiFusion and CNN-baseline is performed and the results prove the significant superiority of most datasets on incomplete multimodal HGR. We also find that the performance of Acc is better than that of sEMG on most datasets except for DB5. This may be because DB5 has only one 3D acceleration channel, leading to a severe decline in feature characterization. This also causes the performance of the multimodal on DB5 to fall short of MVCNN [24], which is good at unimodal characterization using multi-view technology.

V. DISCUSSION

To analyze the reasons for performance improvement from the AiFusion model, the comparison experiments of the fusion module and ablation experiments of alignment components are conducted on all selected datasets.

A. Effectiveness of Fusion Module

To analyze the effectiveness of the transformer-based progressive hierarchical fusion module in the AiFusion model, we conduct comparison experiments of different multimodal fusion modules, i.e., transformer and CNN, with or without progressive fusion. A comparison method performing multimodal fusion with CNN, termed AiFusion-CNN, is implemented. To make a fair comparison between the AiFusion and AiFusion-CNN models, the architecture of AiFusion-CNN is the same as AiFusion except that the transformer fusion module is replaced by CNN fusion module. Specifically, the CNN fusion module in AiFusion-CNN contains three levels of CNN block. Each CNN block consists of 64 convolution kernels, the Batch Normalization layer and ReLU layer. The kernel size of convolution kernels is set to 3×3, 5×5 and 3×3, respectively. In addition, another experiment where the unimodal hierarchical features do not participate in progressive cross-modal fusion (AiFusion-noPro), is completed. The input of the transformer fusion branch in AiFsuion-noPro only contains multi-scale modal information of sEMG and ACC. The experiments of the complete and incomplete multimodal HGR are conducted on all selected datasets in this part. The other experimental settings of training and testing for AiFusion-CNN and AiFsuion-noPro remain consistent with the AiFusion model. The average accuracy of the different fusion modules for complete and incomplete multimodal HGR are listed in Table. V.

The results in the first row of each dataset in Table. V, i.e., sEMG-ACC, is obtained in complete multimodal HGR. The second and third rows of each dataset, i.e., sEMG and ACC, present the results of the incomplete multimodal HGR. The AiFusion obtains the best average accuracies compared to the AiFusion-CNN and AiFusion-noPro for complete multimodal HGR on all datasets. For the incomplete multimodal HGR, the AiFusion also achieves better results on most datasets. The paired t-test for is completed (AiFusion vs. AiFusion-CNN and AiFusion vs. AiFusion-noPro). The statistical results verify that the transformer fusion module plays a more vital role than CNN in extracting interactive knowledge among various modalities. And the progressive hierarchical cross-modal fusion strategy especially contributes to effective cross-modal fusion. The interactive knowledge indeed contributes to the performance of AiFusion on complete and incomplete multimodal HGR.

B. Ablation Experiments of Alignment Components

To explore the effectiveness of the alignments in the AiFusion, we implement the ablation experiments of alignment components, i.e., cross-modal supervised contrastive learning loss $\mathcal{L}_{CL}$ and Kullback-Leibler Divergence Loss $\mathcal{L}_{KL}$. These ablation experiments are conducted on all the selected datasets for complete and incomplete multimodal conditions, respectively. First, the AiFusion model is trained with only the cross entropy loss $\mathcal{L}_{CE}$ and independently tested with complete and incomplete datasets. The results of the AiFusion trained without alignments are regarded as a baseline in the ablation experiments. Then, the AiFusion model is respectively trained

TABLE VI

THE AVERAGE ACCURACIES FOR ABLATION STUDY OF ALIGNMENT COMPONENT IN AIFUSION. THE RESULTS IN **BOLD** TEXT REPRESENT THE BEST PERFORMANCE. ✓AND ✗DENOTE THAT THE CORRESPONDING COMPONENT IS USED AND NOT USED, RESPECTIVELY

Dataset	$\mathcal{L}_{CL}-$ Alignment of embedded space	$\mathcal{L}_{KL}-$ Alignment of probability space	AiFusion with sEMG-ACC	AiFusion with sEMG	AiFusion with ACC
Ninapro DB2	✗	✗	95.18%	79.47%	93.38%
	✓	✗	95.22%↑	79.45%	93.41%↑
	✗	✓	95.23%↑	79.71%↑	93.37%
	✓	✓	**95.28%**↑	**79.73%**↑	**93.48%**↑
Ninapro DB3	✗	✗	90.92%	61.01%	**88.10%**
	✓	✗	**91.14%**↑	61.26%↑	88.02%
	✗	✓	90.83%	61.21%↑	87.83%
	✓	✓	91.11%↑	**61.45%**↑	88.06%
Ninapro DB5	✗	✗	86.11%	73.52%	59.82%
	✓	✗	86.72%↑	73.80%↑	59.63%
	✗	✓	86.56%↑	73.77%↑	59.83%↑
	✓	✓	**87.04%**↑	**74.01%**↑	**59.87%**↑
Ninapro DB6	✗	✗	79.90%	68.62%	70.80%
	✓	✗	80.11%↑	68.53%	70.93%↑
	✗	✓	79.88%	69.14%↑	**71.34%**↑
	✓	✓	**80.63%**↑	**69.18%**↑	71.27%↑
Ninapro DB7	✗	✗	96.66%	85.11%	94.52%
	✓	✗	96.73%↑	85.11%	94.65%↑
	✗	✓	96.67%↑	85.21%↑	94.62%↑
	✓	✓	**96.76%**↑	**85.23%**↑	**94.66%**↑

↑ indicates the result is higher than the corresponding baseline.

with $\mathcal{L}_{CE}+\mathcal{L}_{CL}$ and $\mathcal{L}_{CE}+\mathcal{L}_{KL}$. These AiFusion models trained with alignments are independently tested with complete and incomplete multimodal testing datasets. The other experimental setting is the same as Section. IV-B. The average accuracies for the ablation study of alignment components on Ninapro datasets are presented in Table. VI.

As shown in Table. VI, almost all models with the help of alignment operation obtain higher average accuracy compared to the corresponding baseline, i.e., $\mathcal{L}_{CL}$ or $\mathcal{L}_{KL}$, except for the experiment of AiFusion with ACC for Ninapro DB3. The slight decrease in average accuracy in AiFusion with ACC may result in the individual difference and a few subjects' performance is not improved. The outcomes of the ablation study demonstrate that the combined $\mathcal{L}_{CL}$ and $\mathcal{L}_{KL}$ contribute to achieving the best performance for complete and incomplete multimodal HGR on Ninapro DB2, DB5, DB7. It is worth noting that the alignment operations provide more improvement on Ninapro DB2, DB3 and DB7 for incomplete multimodal HGR, especially for the sEMG-based HGR. It may be explained that knowledge distilled from multimodal fusion has a bigger positive guidance for the weak modality. The results of ablation experiments demonstrate that the alignment in AiFusion plays a positive role in complete and incomplete multimodal HGR.

The current research has its limitations. For example, the information of missing modality is dealt with zero padding in this study. In this way, it is convenient for classification model to detect incomplete multimodal conditions during the testing phase. In the future, it is necessary to provide a more advanced and intelligent detector of missing modalities, such as anomaly detection, in application scenarios. In addition, although the AiFusion obtains state-of-the-art performance on complete multimodal HGR, the performance for incomplete multimodal HGR is obviously lower than that of complete multimodal HGR. This may be due to the limited ability of the alignment operation in reducing the gap between weak and strong modalities. The approach of cross-modal generation can be leveraged to further solve the problem of missing modality.

VI. CONCLUSION

This study is the first work that realizes the complete and incomplete multimodal HGR in a unified fusion model, termed AiFusion. AiFusion not only sufficiently explores the multimodal interactive information with a transformer-based progressive hierarchical fusion strategy, but also aligns the various modalities from both embedding space and probability space, thus promoting the effectiveness and robustness of the sEMG-ACC-based HGR. Extensive experiments on five public datasets corroborate that AiFusion achieves state-of-the-art performance on complete multimodal HGR and also surpasses unimodal baselines in the challenging area of incomplete multimodal HGR on most datasets. This innovative AiFusion model provides a promising solution to construct more effective and robust multimodal HGR-based HMI systems.

REFERENCES

[1] T. Wang, Y. Zhao, and Q. Wang, "Hand gesture recognition with flexible capacitive wristband using triplet network in inter-day applications," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 2876–2885, 2022.

[2] B. Xue, L. Wu, A. Liu, X. Zhang, X. Chen, and X. Chen, "Reduce the user burden of multiuser myoelectric interface via few-shot domain adaptation," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 972–980, 2023.

[3] M. Garg, D. Ghosh, and P. M. Pradhan, "Multiscaled multi-head attention-based video transformer network for hand gesture recognition," *IEEE Signal Process. Lett.*, vol. 30, pp. 80–84, 2023.

[4] X. Zhang, X. Zhang, L. Wu, C. Li, X. Chen, and X. Chen, "Domain adaptation with self-guided adaptive sampling strategy: Feature alignment for cross-user myoelectric pattern recognition," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 1374–1383, 2022.

[5] L. Wu, X. Zhang, K. Wang, X. Chen, and X. Chen, "Improved high-density myoelectric pattern recognition control against electrode shift using data augmentation and dilated convolutional neural network," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 12, pp. 2637–2646, Dec. 2020.

[6] P. Kang, J. Li, S. Jiang, and P. B. Shull, "Reduce system redundancy and optimize sensor disposition for EMG–IMU multimodal fusion human–machine interfaces with XAI," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–9, 2023.

[7] D. Xiong, D. Zhang, X. Zhao, and Y. Zhao, "Deep learning for EMG-based human-machine interaction: A review," *IEEE/CAA J. Autom. Sinica*, vol. 8, no. 3, pp. 512–533, Mar. 2021.

[8] S. Jiang, P. Kang, X. Song, B. P. L. Lo, and P. B. Shull, "Emerging wearable interfaces and algorithms for hand gesture recognition: A survey," *IEEE Rev. Biomed. Eng.*, vol. 15, pp. 85–102, 2022.

[9] M. Wang et al., "Gesture recognition using a bioinspired learning architecture that integrates visual data with somatosensory data from stretchable sensors," *Nature Electron.*, vol. 3, no. 9, pp. 563–570, Jun. 2020.

[10] X. Zhou, C. Wang, L. Zhang, J. Liu, G. Liang, and X. Wu, "Continuous estimation of lower limb joint angles from multi-stream signals based on knowledge tracing," *IEEE Robot. Autom. Lett.*, vol. 8, no. 2, pp. 951–957, Feb. 2023.

[11] B. Yang, C. Shi, Z. Liu, Y. Hu, M. Cheng, and L. Jiang, "Fingertip proximity-based grasping pattern prediction of transradial myoelectric prosthesis," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 1483–1491, 2023.

[12] G. Vecchiato, M. Del Vecchio, J. Ambeck-Madsen, L. Ascari, and P. Avanzini, "EEG–EMG coupling as a hybrid method for steering detection in car driving settings," *Cognit. Neurodyn.*, vol. 16, no. 5, pp. 987–1002, Oct. 2022.

[13] X. Wang et al., "Alternative muscle synergy patterns of upper limb amputees," *Cognit. Neurodynamics*, vol. 2023, pp. 1–15, Apr. 2023.

[14] X. Zhang, X. Chen, Y. Li, V. Lantz, K. Wang, and J. Yang, "A framework for hand gesture recognition based on accelerometer and EMG sensors," *IEEE Trans. Syst., Man, Cybern., A, Syst. Hum.*, vol. 41, no. 6, pp. 1064–1076, Nov. 2011.

[15] A. Krasoulis, I. Kyranou, M. S. Erden, K. Nazarpour, and S. Vijayakumar, "Improved prosthetic hand control with concurrent use of myoelectric and inertial measurements," *J. NeuroEng. Rehabil.*, vol. 14, no. 1, pp. 1–14, Dec. 2017.

[16] J. Zhang, Q. Wang, Q. Wang, and Z. Zheng, "Multimodal fusion framework based on statistical attention and contrastive attention for sign language recognition," *IEEE Trans. Mobile Comput.*, early access, Jan. 10, 2023, doi: 10.1109/TMC.2023.3235935.

[17] Y. Yu, X. Chen, S. Cao, X. Zhang, and X. Chen, "Exploration of Chinese sign language recognition using wearable sensors based on deep belief net," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 5, pp. 1310–1320, May 2020.

[18] K. Shatilov, Y. D. Kwon, L. H. Lee, D. Chatzopoulos, and P. Hui, "MyoKey: Inertial motion sensing and gesture-based QWERTY keyboard for extended realities," *IEEE Trans. Mobile Comput.*, vol. 22, no. 8, pp. 4807–4821, Aug. 2023.

[19] S. Yang, X. Yang, R. Zhang, and K. Liu, "Hierarchical progressive network for multimodal medical image fusion in healthcare systems," *IEEE Trans. Computat. Social Syst.*, vol. 10, no. 4, pp. 1540–1558, Aug. 2023.

[20] A. Rahate, R. Walambe, S. Ramanna, and K. Kotecha, "Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions," *Inf. Fusion*, vol. 81, pp. 203–239, May 2022.

[21] X. Liang, Y. Qian, Q. Guo, H. Cheng, and J. Liang, "AF: An association-based fusion method for multi-modal classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9236–9254, Dec. 2022.

[22] T.-Y. Pan, W.-L. Tsai, C.-Y. Chang, C.-W. Yeh, and M.-C. Hu, "A hierarchical hand gesture recognition framework for sports referee training-based EMG and accelerometer sensors," *IEEE Trans. Cybern.*, vol. 52, no. 5, pp. 3172–3183, May 2022.

[23] M. Jing, J. Li, L. Zhu, K. Lu, Y. Yang, and Z. Huang, "Incomplete cross-modal retrieval with dual-aligned variational autoencoders," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 3283–3291.

[24] W. Wei, Q. Dai, Y. Wong, Y. Hu, M. Kankanhalli, and W. Geng, "Surface-electromyography-based gesture recognition by multi-view deep learning," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 10, pp. 2964–2973, Oct. 2019.

[25] J. Liu, W. Chen, M. Li, and X. Kang, "Continuous recognition of multifunctional finger and wrist movements in amputee subjects based on sEMG and accelerometry," *Open Biomed. Eng. J.*, vol. 10, no. 1, pp. 101–110, Nov. 2016.

[26] A. Gijsberts, M. Atzori, C. Castellini, H. Müller, and B. Caputo, "Movement error rate for evaluation of machine learning methods for sEMG-based hand movement classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 22, no. 4, pp. 735–744, Jul. 2014.

[27] S. Duan, L. Wu, B. Xue, A. Liu, R. Qian, and X. Chen, "A hybrid multimodal fusion framework for sEMG-ACC-Based hand gesture recognition," *IEEE Sensors J.*, vol. 23, no. 3, pp. 2773–2782, Feb. 2023.

[28] M. Ma, J. Ren, L. Zhao, S. Tulyakov, C. Wu, and X. Peng, "SMIL: Multimodal learning with severely missing modality," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 2302–2310.

[29] M. Ma, J. Ren, L. Zhao, D. Testuggine, and X. Peng, "Are multimodal transformers robust to missing modality?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18156–18165.

[30] F. Ma, X. Xu, S.-L. Huang, and L. Zhang, "Maximum likelihood estimation for multimodal learning with missing modality," 2021, *arXiv:2108.10513*.

[31] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[32] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–13.

[33] P. Poklukar, M. Vasco, H. Yin, F. S. Melo, A. Paiva, and D. Kragic, "Geometric multimodal contrastive representation learning," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 17782–17800.

[34] H. Bao, Y. Nagano, and K. Nozawa, "On the surrogate gap between contrastive and supervised losses," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 1585–1606.

[35] P. Khosla et al., "Supervised contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 18661–18673.

[36] M. Atzori et al., "Electromyography data for non-invasive naturally-controlled robotic hand prostheses," *Sci. Data*, vol. 1, no. 1, pp. 1–13, Dec. 2014.

[37] S. Pizzolato, L. Tagliapietra, M. Cognolato, M. Reggiani, H. Müller, and M. Atzori, "Comparison of six electromyography acquisition setups on hand movement classification tasks," *PLoS ONE*, vol. 12, no. 10, Oct. 2017, Art. no. e0186132.

[38] F. Palermo, M. Cognolato, A. Gijsberts, H. Müller, B. Caputo, and M. Atzori, "Repeatability of grasp recognition for robotic hand prosthesis control based on sEMG data," in *Proc. Int. Conf. Rehabil. Robot. (ICORR)*, Jul. 2017, pp. 1154–1159.

[39] 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.

[40] A. Phinyomark, P. Phukpattaranont, and C. Limsakul, "Feature reduction and selection for EMG signal classification," *Expert Syst. Appl.*, vol. 39, no. 8, pp. 7420–7431, Jun. 2012.

[41] R. N. Khushaba and K. Nazarpour, "Decoding HD-EMG signals for myoelectric control—How small can the analysis window size be?" *IEEE Robot. Autom. Lett.*, vol. 6, no. 4, pp. 8569–8574, Oct. 2021.

[42] M. Atzori, A. Gijsberts, H. Müller, and B. Caputo, "Classification of hand movements in amputated subjects by sEMG and accelerometers," in *Proc. 36th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2014, pp. 3545–3549.

[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.