IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 5, 2019, accepted July 16, 2019, date of publication July 19, 2019, date of current version August 13, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2930005

sEMG-Based Gesture Recognition With Embedded Virtual Hand Poses and Adversarial Learning

YU HU[1], YONGKANG WONG[2], (Member, IEEE), QINGFENG DAI[1], MOHAN KANKANHALLI[2], (Fellow, IEEE), WEIDONG GENG[1], AND XIANGDONG LI[1]

[1]State Key Laboratory of CAD & CG, College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China
[2]School of Computing, National University of Singapore, Singapore 117417

Corresponding authors: Weidong Geng (gengwd@zju.edu.cn) and Xiangdong Li (axli@zju.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB1001302, in part by the National Natural Science Foundation of China under Grant 61379067, and in part by the National Research Foundation, Prime Minister's Office, Singapore, through its International Research Centre as a part of the Singapore Funding Initiative.

ABSTRACT To improve the accuracy of surface electromyography (sEMG)-based gesture recognition, we present a novel hybrid approach that combines real sEMG signals with corresponding virtual hand poses. The virtual hand poses are generated by means of a proposed cross-modal association model constructed based on the adversarial learning to capture the intrinsic relationship between the sEMG signals and the hand poses. We report comprehensive evaluations of the proposed approach for both frame- and window-based sEMG gesture recognitions on seven-sparse-multichannel and four-high-density-benchmark databases. The experimental results show that the proposed approach achieves significant improvements in sEMG-based gesture recognition compared to existing works. For frame-based sEMG gesture recognition, the recognition accuracy of the proposed framework is increased by an average of +5.2% on the sparse multichannel sEMG databases and by an average of +6.7% on the high-density sEMG databases compared to the existing methods. For window-based sEMG gesture recognition, the state-of-the-art recognition accuracies on three of the high-density sEMG databases are already higher than 99%, i.e., almost saturated; nevertheless, we achieve a +0.2% improvement. For the remaining eight sEMG databases, the average improvement with the proposed framework for the window-based approach is +2.5%.

INDEX TERMS Hand gesture recognition, surface electromyography (sEMG), myoelectric control, generative adversarial learning, virtual hand pose.

I. INTRODUCTION

The surface electromyography (sEMG) signal is a kind of biological signal collected by putting myoelectric electrodes on the skin. The sEMG-based gesture recognition plays an significant role in muscle-computer interface (MCI) since it provides a way to understand the user's intention. The sEMG-based gesture recognition has been employed in three major areas [1]: assistive technology [2], [3], rehabilitative technology [4], [5] and input technology [6].

From the perspective of the input data streams, sEMG-based gesture recognition can be further categorized into unimodal methods (based only on sEMG signals) and multimodal methods (involving data streams of no less two modalities, e.g., sEMG + inertial measurement unit (IMU)). In general, multimodal MCIs can achieve higher gesture recognition accuracies than unimodal MCIs can [7], [8]. However, the multimodal MCI requires users to wear additional sensors, which degrades the user experience during interaction. Unimodal systems have the advantage of higher usability since they require the user to wear only sEMG electrodes; however, their gesture recognition accuracy is relatively low. If a hybrid system combining the advantages of unimodal and multimodal systems could be built, it would be possible to improve sEMG-based gesture recognition accuracy while ensuring usability.

There is an intrinsic physiological relationship between sEMG signals and the concurrently performed gestures since each hand movement is driven by specific muscle groups [9]. Most existing sEMG benchmark databases contain synchronously captured hand poses. If we could establish a cross-modal association model between sEMG signals and hand

The associate editor coordinating the review of this manuscript and approving it for publication was Maurizio Tucci.

poses, this model could be used to generate corresponding virtual hand poses based on real sEMG signals. On this basis, a multimodal classifier could be constructed for sEMG-based gesture recognition. Considering the quality of the user experience, we wish to collect only unimodal data during system operation to achieve higher recognition accuracy without degrading the usability of the user interface.

The recognition accuracy increases as the observation delay increases [10]. To satisfy real-time usage constrains for MCI, the response time should be shorter than 300ms [11]. Therefore, we need to trade off between the observation delay (window length) and recognition accuracy. Depending on the requirements of application scenarios for the observational latency, the existing works can be divided into frame-based and window-based methods. The frame-based method [12], [13] only recognizes one frame sEMG signal which has the shortest observation delay, and is often used in areas such as prosthetic control that requires low response time. For a higher recognition accuracy, the window-based method [7], [14]–[16] classifies a segment of the sEMG signal in the window and is often applied in muscle-computer interaction.

The main contributions of this work are threefold:

- We propose a novel generative adversarial network (GAN)-based approach to construct a cross-modal association model through adversarial learning. The resulting model can better capture the intrinsic relationship between sEMG signals and hand poses and can be effectively used to generate corresponding hand poses based on input sEMG signals.
- To achieve higher accuracy without compromising the user experience, we design a two-step pipeline classification solution for gesture recognition in an MCI that is different from traditional gesture recognition based solely on classification. In step one, virtual hand poses are generated using the aforementioned cross-modal association model relating sEMG signals and hand poses. In step two, each sEMG signal is paired with its corresponding virtual hand pose and fed into the classifier for gesture recognition. From the perspective of the input and output, this solution functions in a ''unimodal'' manner because the virtual hand poses are ''unseen'' from the external point of view.
- We report comprehensive evaluations of both frame-based and window-based sEMG gesture recognition conducted on 7 sparse multichannel sEMG databases and 4 high-density sEMG databases. The results show that the proposed framework achieves better performance than state-of-the-art traditional unimodal methods. For frame-based sEMG gesture recognition, the improvements are $+5.2\%$ and $+6.7\%$ on the sparse multichannel and high-density sEMG databases, respectively. For window-based sEMG gesture recognition, the state-of-the-art recognition accuracies on 3 of the high-density sEMG databases are already higher than 99%, i.e., almost saturated; nevertheless, we achieve a $+0.2\%$ improvement. For the remaining 8 sEMG databases, the average improvement with the proposed framework for the window-based approach is $+2.5\%$.

II. RELATED WORK

The applications of gesture recognition are multifaceted, from sign language to medical rehabilitation to virtual reality [27]. Great progress has been achieved on vision-based gesture recognition, mainly exploiting hand RGB and depth images acquired by RGB cameras [28], [29], depth cameras [30]–[32] and binocular cameras [28], [33] as input to track hand movements or recognize gestures. To this end, sEMG signals provide a novel means for humans to communicate with computers; such signals are collected by recording the electrical activity produced by the skeletal muscles [34] by means of noninvasive sEMG electrodes placed on the skin. Recently, Phinyomark and Scheme [35] published an excellent survey on sEMG pattern recognition and divided the existing approached into two categories: feature engineering and feature learning. Simão *et al*. [36] also presented an extensive review of sEMG pattern recognition for human-computer interfaces, summarizing the current feature extraction techniques and novel classification methods.

We summarize the related sEMG-based gesture recognition works in recent years from various perspectives in Table 1, such as the number of modalities, electrode arrangement, observational latency, application scenarios, classification methodology, *etc*. The unimodal methods are based merely on sEMG signal during runtime while the multimodal methods recognize gestures through data of not less than two modalities. The sparse multi-channel methods mainly focus on the sEMG signal which is collected by placing a small number of electrodes on specific muscles, while the high-density methods employ the array of sEMG electrodes for signal acquisition. The frame-based methods classify gesture based on one frame of sEMG signal, but the window-based methods apply a segment of sEMG signal for recognition. The intrasubject and intersubject evaluations simulate different sEMG gesture recognition application scenarios. Specifically, the training and test sets of intrasubject evaluation are from the same subject, while those of intersubject evaluation are from different subjects. The intrasubject evaluation simulates the usage of sEMG-based gesture recognition in practical application scenarios, such as assistive technology and rehabilitative technology. In this study, we mainly focus on improving the performance of gesture recognition in intra-subject evaluation with unimodal input. The existing works can be categorized into unimodal and multimodal methods in terms of the number of modalities during runtime.

Unimodal methods provide better user experience, and many researchers have focused on developing novel sEMG-based models to achieve higher classification accuracy. Depending on the classification methodology used, unimodal methods can be broadly divided into conventional machine learning based method and deep learning based method [35]. The conventional machine learning based method consists of signal preprocessing, feature extraction, feature

TABLE 1. Summary of research perspectives of existing works in the literature.

	Number of modalities		Electrode arrangement		Observational latency		Application scenarios		Classification methodology	
	unimodal	multimodal	sparse multi-channel	high-density	frame-based	window-based	intrasubject	intersubject	conventional machine learning based method	deep learning based method
Atzori *et al.* [17]	✓		✓			✓	✓		✓	
Geng *et al.* [12]	✓		✓	✓	✓		✓			✓
Atzori *et al.* [14]	✓		✓			✓	✓		✓	✓
Kyranou *et al.* [18]		✓	✓			✓	✓		✓	
Kim *et al.* [19]		✓	✓			✓	✓		✓	
Pizzolato *et al.* [20]	✓		✓			✓	✓		✓	
Palermo *et al.* [21]	✓		✓			✓	✓		✓	
Krasoulis *et al.* [22]	✓		✓			✓	✓		✓	
Zhai *et al.* [15]	✓		✓			✓	✓			✓
Du *et al.* [23]	✓		✓	✓	✓	✓		✓		✓
Du *et al.* [13]	✓		✓	✓	✓	✓	✓			✓
Shin *et al.* [24]		✓	✓			✓	✓	✓		✓
Hu *et al.* [16]	✓		✓	✓		✓	✓			✓
Jiang *et al.* [8]		✓	✓			✓	✓		✓	
Tao *et al.* [25]		✓	✓			✓		✓		✓
Kundu *et al.* [26]		✓	✓			✓	✓		✓	
Wei *et al.* [7]	✓	✓	✓			✓	✓	✓		✓
Proposed	✓		✓	✓	✓	✓	✓			✓

normalization and gesture classification. The core of conventional machine learning based method is to present novel handcrafted feature sets with high distinguishability [11], [37]–[39]. Hudgins *et al.* [11] proposed the most commonly used feature set, which contains zero crossings (ZC), slope sign changes (SSC), mean absolute values (MAV), and waveform lengths (WL). Khushaba *et al.* [38] designed a novel feature set containing 7 time-domain descriptors for the extraction of spatiotemporal information. After discovering the optimal feature set, the traditional machine learning classifiers has been employed for sEMG-based gesture recognition, such as Support Vector Machine (SVM) [20], [40], [41], Linear Discriminative Analysis (LDA) [42], [43], k-Nearest Neighbor (kNN) [22], [44], Random Forests [17], [20], *etc*. In recent years, feature learning methods based on deep learning have been applied for sEMG-based gesture recognition. Researchers have converted sEMG signals into various types of sEMG images and have presented multiple deep learning architectures for gesture recognition [12], [14], [15]. To address the problem of insufficient labeling of sEMG data, a semi-supervised learning architecture was recently presented [13]. Subsequently, researchers have considered intermuscle relationships to propose multistream CNN architectures [7], [45]. Since sEMG signal is sequential data by nature, a hybrid CNN-RNN architecture has been proposed to achieve a higher recognition accuracy by modeling both spatial and temporal information [16].

The multimodal approach is an effective way to improve recognition accuracy because it provides information on the same gesture from multiple perspectives. Existing works have combined sEMG signals with data of other modalities (e.g., IMU data) to enhance the recognition accuracy and robustness of gesture recognition systems. The conventional methods consist of signal preprocessing, feature extraction from the sEMG and IMU signals, feature fusion, and gesture classification. Kyranou *et al.* [18] trained a linear discriminant analysis (LDA) classifier to recognize six hand grip patterns and showed that the consideration of additional sensory modalities improved the robustness of prosthetic control. Jiang *et al.* [8] designed a wristband that fused sEMG and IMU signals for sEMG-based gesture recognition. Kim *et al.* [19] presented a novel wearable human-computer interface combining sEMG and IMU sensors. Deep-learning-based methods treat multimodal data as input to a multistream network and apply various fusion strategies for different sensors [7], [26]. Tao *et al.* [25] captured IMU and sEMG signals from a Myo armband and fed them into a CNN architecture to recognize human activities. The Myo armband is a commercial product that collects sEMG and IMU data from the user. Shin *et al.* [24] proposed an automatic Korean sign language recognition system based on sEMG and IMU sensors and group-dependent neural network models. Du *et al.* [13] proposed a novel semi-supervised CNN architecture that used dataglove data in the training phase to capture more discriminative features of sEMG signals.

GAN-based cross-modal generation provides association information between multimedia data, such as text, image and audio, which has become an active research topic [46]. The cross-modal generation can recover the missing modality from the existing one and researchers first explored the relationship between text and image [47]–[49]. Visual and audio are two symbiotic modalities behind the video, including common and complementary information [50], and it is an interesting research to minner the relation underlying the two modalities. Chen *et al.* [51] employed GAN to solve the cross-modal audio-image mutual generation for the first time and composed two datasets with paired images and sounds. Hao *et al.* [50] presented a cross-modal cycle GAN. As human can imagine a scene based on sound, Wan *et al.* [52] introduces a novel task to make machine thinking like humans and successfully used conditional GAN to generate images from sounds.

In general, multimodal systems achieve higher recognition accuracy than the unimodal systems do. However, the need

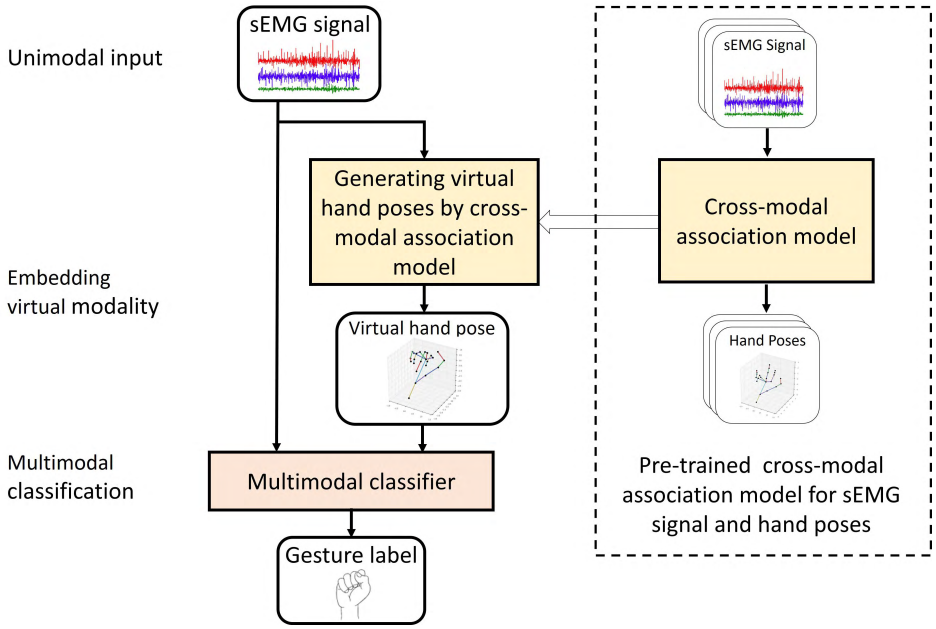


FIGURE 1. Diagram of the proposed hybrid architecture for sEMG-based gesture recognition.

to wear additional sensors during operation degrades the user experience. In this work, we build upon a novel two-step pipeline classification solution to improve the recognition accuracy achieved in intrasubject evaluation by exploiting multimodal data during training but using only sEMG signals for recognition at run time to ensure the quality of the user experience.

III. PROPOSED METHOD

A diagram of our hybrid framework for sEMG-based gesture recognition is shown in Figure 1. The overall workflow of the proposed hybrid framework is separated into two phases. During the offline preparation, we collect sEMG signals and the corresponding hand poses to construct a cross-modal association model to capture the intrinsic relationship between the signals and poses. During run-time testing, the only input collected consists of sEMG signals, and we establish a multimodal gesture classifier by incorporating the virtual hand poses generated by the pretrained cross-modal association model constructed in the offline preparation phase.

A. CROSS-MODAL ASSOCIATION MODEL WITH ADVERSARIAL LEARNING

1) PROBLEM STATEMENT

Given a segment of an sEMG signal of length L (denoted by $[\boldsymbol{FS_1}, \boldsymbol{FS_2}, ..., \boldsymbol{FS_L} | \boldsymbol{FS_t} \in \boldsymbol{R^{C1}}]$, where $\boldsymbol{FS_t}$ is the frame of the sEMG signal collected at time t and $C1$ is the number of sEMG signal channel), we need to generate the corresponding virtual hand pose for each frame (denoted by $[\boldsymbol{FM'_1}, \boldsymbol{FM'_2}, ..., \boldsymbol{FM'_L} | \boldsymbol{FM'_t} \in \boldsymbol{R^{C2}}]$, where $\boldsymbol{FM'_t}$ is the virtual hand pose corresponding to frame $\boldsymbol{FS_t}$ of the sEMG signal and $C2$ is the hand pose channel number). Therefore, we build a function to describe the cross-modal association, where the input to and output of this function are the sEMG signal and the hand pose, respectively.

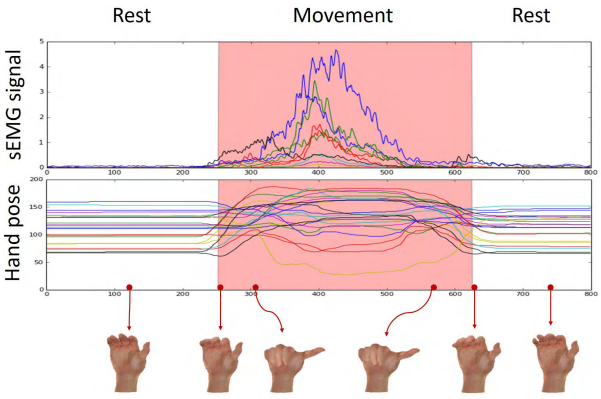


FIGURE 2. The sEMG signals and corresponding hand poses for the thumbs-up gesture.

To visualize the correlation between sEMG signals and hand poses, we plot the sEMG signal waveforms, dataglove data and the corresponding 3D virtual hand poses for the thumbs-up gesture in Figure 2. The gesture begins with the neutral hand pose and moves continuously into the static thumbs-up gesture; the changes in hand pose from one frame to the next are controlled by the sEMG signal. Based on the hypothesis of the continuity of hand movement, we can transform the problem of solving the above function into the following problem. Suppose that the neutral hand pose is denoted by $FM_0 \in R^{C2}$. Given single-frame sEMG signals

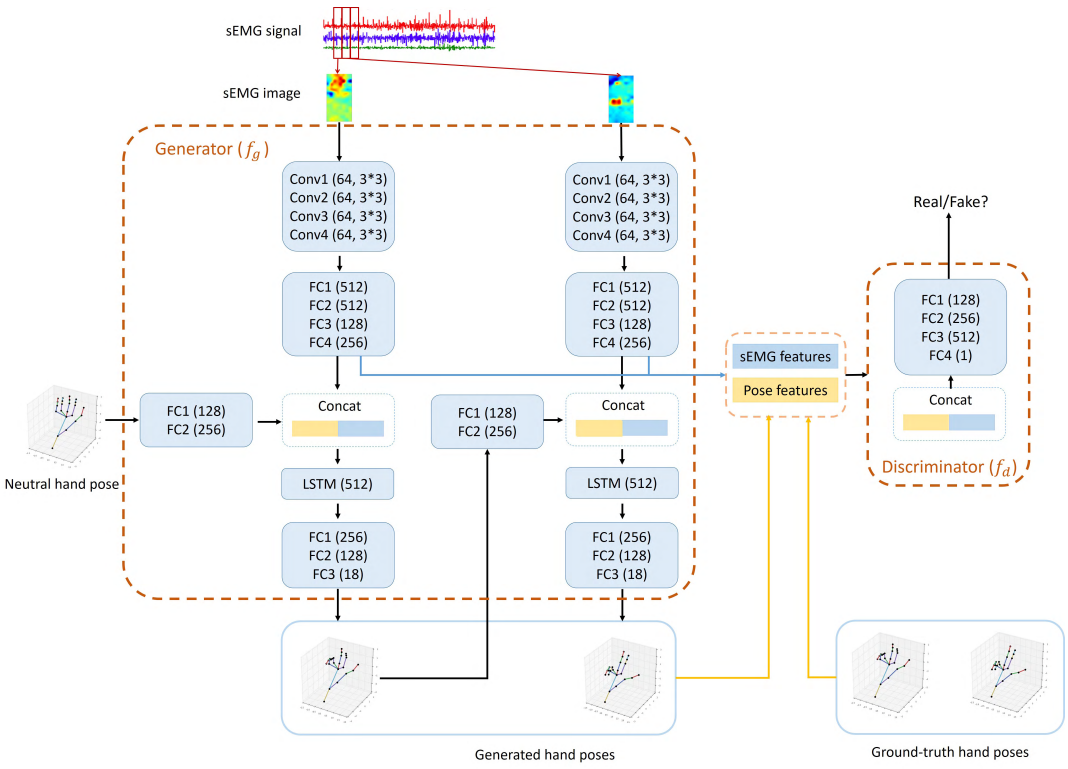


FIGURE 3. Architecture of the cross-modal association model with GAN.

$\boldsymbol{FS_t} \in R^{C1}$, we can calculate the corresponding virtual hand poses $\boldsymbol{FM'_t} \in \boldsymbol{R^{C2}}$ using the following solution procedure.

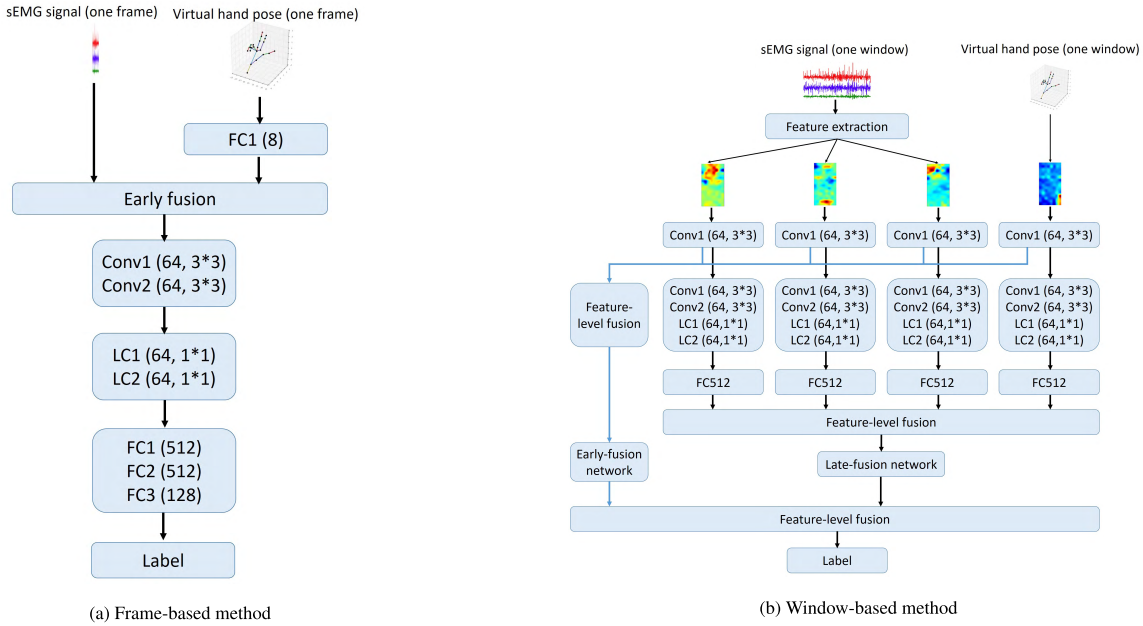
1) For t = 1, $\boldsymbol{FM'_1} = f_g(\boldsymbol{FS_1}, \boldsymbol{FM_0})$, where FM'_1 is the virtual hand pose at time 1.
2) For t = 2, ..., m, first repeat step 1, and then calculate $\boldsymbol{FM'_t} = f_g(\boldsymbol{FS_t}, \boldsymbol{FM'_{t-1}})$.

Therefore, our target is to build a deep learning network to solve for the cross-modal association model f_g.

2) GAN-BASED ASSOCIATION MODEL

The usual means of solving for the cross-modal association model f_g is to transform the problem into a prediction or regression problem [53]. The sEMG signals and hand poses are of different modalities, and each frame of an sEMG signal and its corresponding hand pose have an intrinsic correlation. One of the key requirements in cross-modal generation is that the output data generated in one modality should be matched with the input data, which are of another modality. Therefore, we need to employ suitable expertise to decide whether each hand pose does indeed match the corresponding input sEMG signal. However, unlike classical cross-modal generation problems (e.g., the generation of images from text or vice versa), it is difficult for human experts to judge whether an sEMG signal is properly matched with a hand pose. Motivated by the application of the "discriminator" in a GAN for image generation to judge whether an image is real [54], we similarly employ a GAN to help judge whether a generated hand pose is correctly matched with the corresponding sEMG signal; thus, the GAN implicitly plays the role of "expert judgment" during training.

We further introduce adversarial learning into the construction of the function f_g. The backbone network is a conditional GAN (CGAN) [55] that takes the sEMG signal as the condition input in the discriminator in addition to the virtual hand pose. Under the assumption that all paired samples can be labeled as belonging to one of two categories (matched or unmatched), the trained "discriminator" in the GAN will provide a "yes/no" answer for each paired sample, thus mimicking expert judgment on whether the paired sample is matched or not. The pose feature and sEMG feature are combined to form the paired sample. The input of discriminator is hand pose data with its corresponding sEMG signal, and the output label is matched or unmatched. We supply the positive case (matched, ground-truth hand pose with its sEMG signal) and the negative case (unmatched, generated hand pose with its sEMG signal) for training. The GAN-based association model consists of a generator module (f_g) and a discriminator module (f_d); the detailed architecture is shown in Figure 3. The architecture of the generator module f_g is based on a hybrid CNN-RNN. The sEMG stream is composed of 4 convolutional layers and 4 fully connected (FC) layers. Each convolutional layer consists of 64 filters with dimensions of 3×3, and the four FC layers have 512, 512, 128 and



(a) Frame-based method

(b) Window-based method

FIGURE 4. Architecture of the gesture classifier f_c.

256 hidden units. The hand pose stream consists of 2 FC layers with 128 and 256 hidden units. To prevent overfitting, we apply dropout to the FC layers and the last convolutional layer. Then, we concatenate the output from the last FC layers of both the sEMG and hand pose streams to form the input of a long short-term memory (LSTM) layer with 512 hidden units, followed by 3 FC layers. The 3 FC layers have 256, 128 and 18 hidden units, where 18 is the dimensionality of the hand pose data. We design a neural network to serve as the discriminator f_d, taking both the sEMG feature vector and the hand poses as input, where the sEMG feature vector is specified as the condition. The high-level sEMG feature vector and hand poses are concatenated and fed into the discriminator, which is composed of 4 FC layers with 128, 256, 512 and 1 hidden units. Then, the discriminator outputs the probability that each input hand pose is ''real'' (a ground-truth hand pose) or ''fake'' (a generated virtual hand pose).

3) LOSS FUNCTIONS

The loss functions of the GAN-based cross-modal association model are as follows:

$$Loss_g = -\frac{\alpha}{N}\sum_{i=1}^{N} log((\boldsymbol{FM}_t')^{\boldsymbol{d}}) + \frac{\beta}{N}\sum_{i=1}^{N}((FM_t' - FM_t)^2) \tag{1}$$

$$Loss_d = -\frac{1}{N}\sum_{i=1}^{T} log((\boldsymbol{FM}_t)^{\boldsymbol{d}}) + \frac{1}{N}\sum_{i=1}^{N} log(1 - (\boldsymbol{FM}_t')^{\boldsymbol{d}}) \tag{2}$$

where $Loss_g$ is the loss function of the generator, $Loss_d$ is the loss function of the discriminator, $\boldsymbol{FM}_t'$ is the virtual hand pose at time t, $\boldsymbol{FM}_t$ is the real hand pose at time t, $(\boldsymbol{FM}_t')^{\boldsymbol{d}}$ is the output of f_d with $\boldsymbol{FM}_t'$ as the input, $(\boldsymbol{FM}_t)^{\boldsymbol{d}}$ is the output of f_d with $\boldsymbol{FM}_t$ as the input, N is the number of samples, and α and β are two weight parameters.

B. GESTURE CLASSIFICATION WITH THE HYBRID FRAMEWORK

We embed the virtual hand poses into the sEMG-based gesture classification framework and construct a multimodal gesture classifier $label_t = f_c(\boldsymbol{FS}_t, \boldsymbol{FM}_t')$ with two modalities. The first modality is the real sEMG signal $\boldsymbol{FS}_t$, and the second modality is the virtual hand pose. The cross-modal association model is trained in advance using real sEMG signals and hand poses and subsequently remains unchanged during the classification process.

In general, we establish the function f_c through multimodal learning, for which the multimodal fusion strategy is key. The available strategies for this purpose mainly include model-agnostic approaches (including early, late and hybrid fusion) and model-based approaches (including kernel-based models, graphical models and neural networks) [57]. The proposed hybrid framework can be utilized in both frame-based and window-based applications and the network structure diagrams are shown in Figure 4.

There are two main differences between the frame-based and window-based methods: the input images generation and classification network architecture. Frame-based method respectively converts one frame raw sEMG signal and its corresponding virtual hand pose into images. Window-based method extracts traditional feature sets from a segment of sEMG signal to reform sEMG images and converts the segment of virtual hand poses associated with the segment of sEMG signal into an image. The specific image

TABLE 2. Descriptions of the benchmark sEMG databases used in this paper.

Name	Number of gestures	Subjects	Number of channels	Number of trials	Trials for training	Trials for testing
NinaDB1 [17]	53	27	10	10	1,3,4,6,7,8,9	2,5,10
NinaDB2 [17]	50	40	12	6	1,3,4,6	2,5
NinaDB3 [17]	50	11	12	6	1,3,4,6	2,5
NinaDB4 [20]	53	10	12	6	1,3,4,6	2,5
NinaDB5 [20]	53	10	16	6	1,3,4,6	2,5
NinaDB6 [21]	7	10	14	120	odd trials	even trials
NinaDB7 [22]	41	22	12	6	1,3,4,6	2,5
CapgDBa [12]	8	18	128	10	LOOCV	LOOCV
CapgDBb [12]	8	20	128	10	LOOCV	LOOCV
CapgDBc [12]	12	10	128	10	LOOCV	LOOCV
csl-hdemg [56]	27	25	192	10	LOOCV	LOOCV

representation method described above can be found in Hu *et al*. [16]. The network architecture of frame-based method consists of sEMG stream and virtual hand pose stream. The early fusion technique is applied in frame-based sEMG gesture recognition, in which the input sEMG stream and the virtual hand pose stream are fused via simple concatenation. The concatenated data are utilized as the input to GengNet [12] for sEMG-based gesture recognition, which has 2 convolutional layers, 2 locally connected (LC) layers and 3 FC layers. In the window-based method, the input sEMG images from three traditional feature sets and the virtual hand pose image are treated as four views and the WeiNet [7] with an additional FC layer at the beginning of the virtual hand pose view is applied for multi-view learning to fuse the multimodal data for sEMG-based gesture recognition. The number of hidden units in this FC layer is equal to the number of channels of the sEMG signal.

During run-time recognition, our classification solution is a two-step pipeline; it is not based solely on a classifier. In step one, virtual hand poses are generated from the input sEMG signals via the learned cross-modal association model. In step two, each input sEMG signal is paired with its corresponding virtual hand pose and fed into the learned multimodal classifier for gesture recognition. From the perspective of the input and output, this solution functions in a ''unimodal'' manner because the virtual hand poses are ''unseen'' from the external point of view.

IV. EXPERIMENTAL RESULTS

In this section, we first introduce the experimental setup and then compare the proposed method with the state-of-the-art methods on seven benchmark databases. Subsequently, we evaluate and discuss the effects of different parts of the proposed architecture.

A. EXPERIMENTAL SETUP

1) DATA PREPARATION

We conducted evaluations on 7 subdatabases of the sparse multichannel sEMG benchmark database NinaPro [17] [20]–[22] (denoted by NinaDB1-NinaDB7), 3 subdatabases of the high-density sEMG benchmark database CapgMyo [12] and the csl-hdemg database [56].

The NinaPro database is the largest, best known sparse multichannel sEMG benchmark database [35]. It contains 7 subdatabases; detailed information on each subdatabase can be seen in Table 2. It collects sEMG signals from 117 able-bodied subjects and 13 amputees performing a subset of 61 predefined hand movements and represents more than 48,000 trials and 326,000 s of muscle contractions in total [35]. Data in other modes were also synchronously acquired, including IMU data and hand kinematic data. The CapgMyo and csl-hdemg databases are two widely used high-density sEMG benchmark databases [35]. The CapgMyo database contains 128 channels constituting a $16*8$ array. The csl-hdemg database contains 192 channels, but, following previous work [56], we used only 168 of these channels, forming a $24 * 7$ array.

To facilitate the performance comparison, we adopted the same database split for intrasubject gesture recognition used in previous works [12], [17], [20]–[22], [56]. In previous studies on the NinaPro databases [14], [15], the training set consisted of approximately 2/3 of the gesture trials for each subject, and the remaining trials constituted the test set. However, for NinaDB6, which contains 120 trials for each gesture, the odd trials constituted the training set, and the remaining trials formed the test set. We followed the leave-one-out cross-validation (LOOCV) evaluation procedure described in previous works [12], [13], [56] for the CapgMyo and csl-hdemg databases.

We evaluated the proposed method on a total of 11 benchmark databases, as described above, but only the NinaDB1, NinaDB2 and NinaDB5 databases contain dataglove data that provide the hand poses associated with each gesture. For the other databases, which contain only sEMG signals, we artificially specified a static hand pose (finger joint angles) for each gesture based on the gesture images provided in each database and generated the corresponding dynamic process of hand pose variation. The detailed generation process is as follows: (1) Each repetition was divided into three parts. The second part corresponds to the static gesture, and the first and third parts are the dynamic gesture processes.

(2) A hand possesses 18 joint points. We specified that the joint angles of the neutral hand pose are all equal to 0. Referring to the neutral hand pose, we rotated each joint as necessary to achieve each gesture, and the corresponding rotation angles were defined as the final joint angles for that gesture. (3) For the static gesture, the hand pose in each frame was defined by the specified finger joint angles. (4) For the dynamic gesture processes, the hand poses were obtained through spherical interpolation between the neutral hand pose and the specified gesture.

We assume that the subjects have similar neutral hand pose and the neutral hand pose collected during training can be used as the FM_0 during evaluation. Since only NinaDB1, NinaDB2 and NinaDB5 databases contain dataglove data, we calculate the average value of the neutral hand pose in training data as the FM_0 applied for testing. For the other 9 benchmark databases merely contain the sEMG signals, the hand poses data relative to neutral hand pose are generated by the process described in the previous paragraph, and the joint angles of the neutral hand pose are set to 0.

Existing works have mainly focused on window-based recognition. Accordingly, to ensure fair comparisons, we divided the sEMG signals into small segments to evaluate the proposed framework. The sliding window strategy was utilized for this division of the sEMG signals. We set the window length to 200 ms in accordance with previous work [13].

2) EVALUATION METRICS AND PROTOCOLS

There are two classical evaluation approaches for sEMG-based gesture recognition, i.e., intrasubject evaluation and intersubject evaluation. Methods based on intrasubject evaluation involve mining the similarities between gestures performed by the same subject to enable more accurate recognition [12], [38], [45]. Methods based on intersubject evaluation aim to eliminate the differences between subjects via domain adaptation [23], [58] and transfer learning [59], [60]. In this work, we mainly focused on improving the intrasubject performance for sEMG-based gesture recognition.

For each subject i, the corresponding data were divided into a training subset (T_i) and a test subset (S_i), and the evaluation metrics and protocols presented in [13], [17] were applied. The network for each subject was trained based on that subject's own sEMG training data (T_i). Then, for a database with N subjects, each of the N learned models was evaluated on the test subset for the same subject (i). For each classification solution j, its accuracy was calculated as given below:

$$Accuracy = \frac{Number\ of\ correct\ samples}{Total\ number\ of\ test\ samples} \times 100\% \quad (3)$$

Then, we calculated the average accuracy over all subjects to obtain the final gesture recognition accuracy for each database. In our experiments, we obtained 203 classification solutions in total over the 11 benchmark databases.

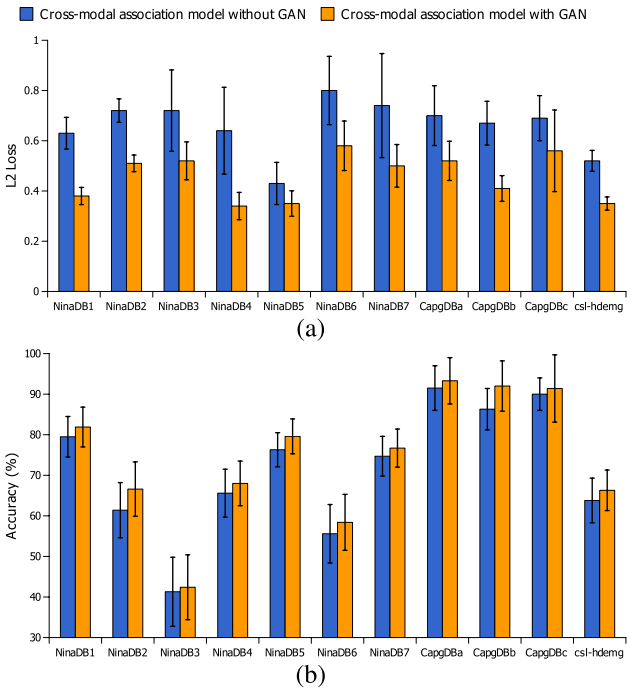


FIGURE 5. Cross-modal association models constructed with and without adversarial learning. Each bar represents the average value, while the error bars represent the standard deviations. (a) L2 loss between the ground-truth and virtually generated hand poses. (b) Recognition accuracy of sEMG-based gesture recognition.

3) NETWORK TRAINING

The proposed framework was implemented with TensorFlow and trained via stochastic gradient descent (SGD). We performed intrasubject evaluation on each of the 11 benchmark databases, meaning that the sEMG signals used for training and evaluation were obtained from the same subject. The training subset T_i was used for network training, and the test subset S_i was used to evaluate the performance of the proposed method. To achieve faster convergence, we employed a pretraining strategy in which the network for each subject was first trained on T, the union of the training subsets for all subjects, i.e., $T = \{Union\ of\ all\ T_i\ from\ subjects\ i, i = 1, ..., N\}$. The network for each subject was initialized with the pretrained model and was then refined based on that subject's own sEMG training data (T_i). In each experiment, 28 training epochs were performed to fine-tune the model parameters, and dropout and batch normalization were applied to make the training processes easier and faster. The learning rate was initially set to 0.1 and was subsequently divided by 10 after the 16th and 24th epochs.

B. EFFECT OF ADVERSARIAL LEARNING ON THE CROSS-MODAL ASSOCIATION MODEL

The achitecture of cross-modal association model with adversarial learning is shown in Figure 3, which has both the generator (f_g) and discriminator (f_d). The architecture of compared cross-modal association model without adversarial learning merely contains the generator f_g and there is no discriminator

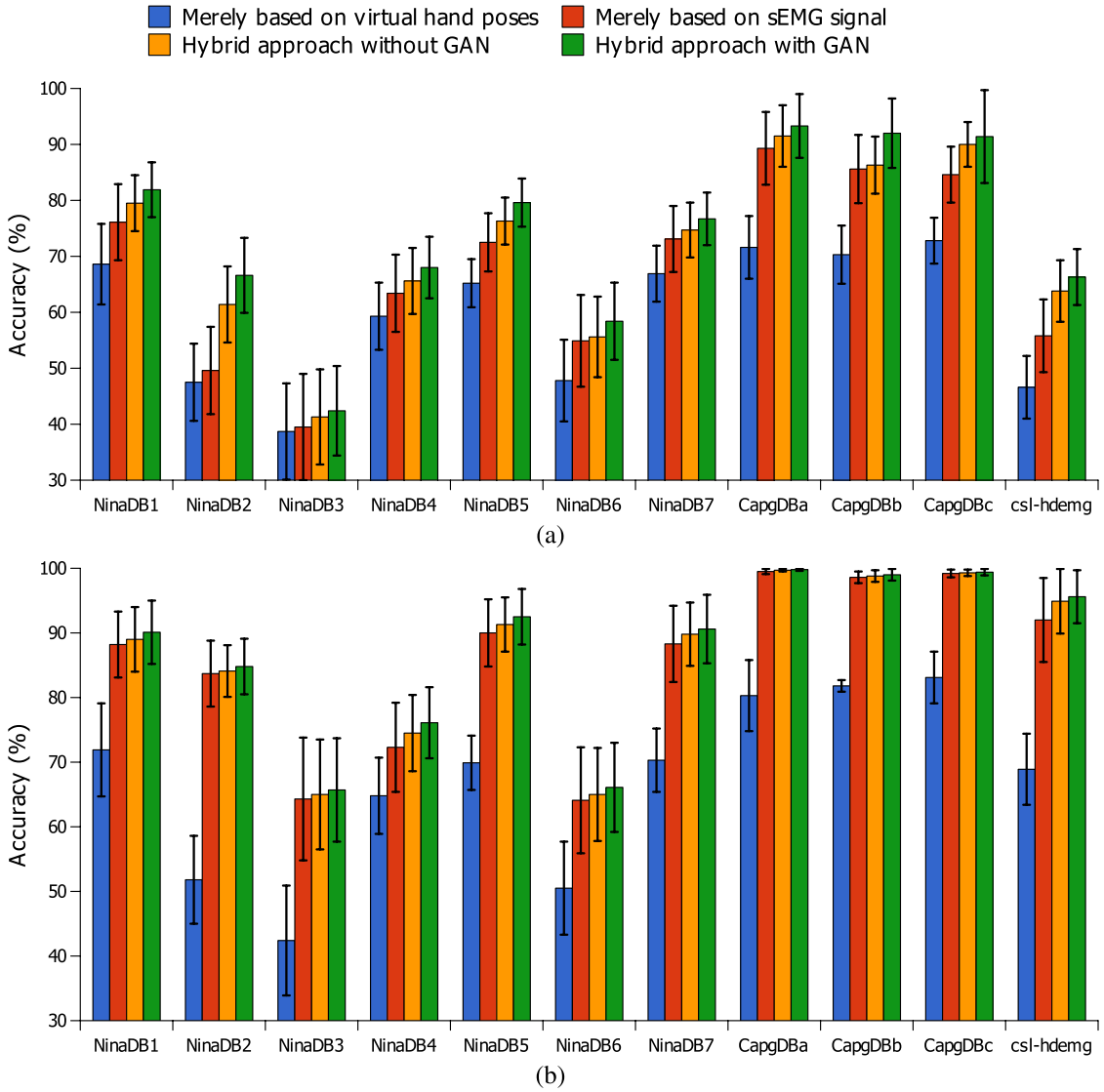


FIGURE 6. Ablation studies on different configurations for sEMG-based gesture recognition on the 11 benchmark databases. The height of each column represents the average accuracy, while the error bars represent the standard deviations. (a) Frame-based sEMG gesture recognition. (b) Window-based sEMG gesture recognition.

(f_d) to judge the correlation between sEMG signals and hand poses. To make an objective comparison, the generator (f_g) of both cross-modal association model without and with GAN has the same network architecture.

To demonstrate the advantages of exploiting adversarial learning when constructing the cross-modal association model, we carried out two types of experiments. First of all, we compared the L2 loss between the ground-truth hand poses and virtual hand poses; the results are presented in Figure 5 (a), where the hand poses are represented by the corresponding finger joint angles (in radians). Then, we embedded virtual hand poses into the proposed sEMG-based gesture recognition framework and compared the recognition accuracies achieved with the hybrid approach with and without the GAN, as shown in Figure 5 (b). The experimental results show the following: (1) The virtual hand poses generated by the cross-modal association model constructed with adversarial learning have fewer errors than those generated by the model constructed without the GAN on both sparse multichannel and high-density sEMG databases. (2) The hybrid approach with GAN outperforms the hybrid approach without GAN for both frame-based and window-based sEMG gesture recognition. Therefore, we conclude that adding GAN during training can help the cross-modal association model better capture the intrinsic relationship between sEMG signals and hand poses.

C. ABLATION STUDIES ON THE PROPOSED METHOD

We conducted an ablation study of the proposed framework in which we compared the accuracy of four approaches to sEMG-based gesture recognition, as shown in Figure 6: (1) merely based on virtual hand poses, in which only the

TABLE 3. Recognition accuracies (%) on the 11 benchmark sEMG databases. The reported performance was achieved with a window length of 200 ms. The frame-based recognition accuracies are shown in parentheses.

	NinaDB1	NinaDB2	NinaDB3	NinaDB4	NinaDB5	NinaDB6	NinaDB7	CapgDBa	CapgDBb	CapgDBc	csl-hdemg
Atzori [17]	75.3(-)	-	-	-	-	-	-	-	-	-	-
Pizzolato [20]	-	-	-	69.1(-)	69.0(-)	-	-	-	-	-	-
Palermo [21]	-	-	-	-	-	52.4(-)	-	-	-	-	-
Krasoulis [22]	-	-	-	-	-	-	82.7(-)	-	-	-	-
AtzoriNet [14]	66.6(-)	-	-	-	-	-	-	-	-	-	-
ZhaiNet [15]	-	78.7(-)	-	-	-	-	-	-	-	-	-
GengNet [12]	77.8(76.1)	50.2(49.6)	41.0(39.5)	64.8(63.4)	74.0(72.5)	56.4(54.9)	74.6(73.1)	99.5(89.3)	98.6(85.6)	99.2(84.6)	92.0(55.8)
DuNet [13]	79.4(78.1)	52.6(52.2)	41.3(39.8)	64.8(63.9)	77.9(74.9)	56.8(55.2)	74.2(73.4)	99.6(89.5)	98.7(85.9)	99.2(85.0)	78.3(56.0)
HuNet [16]	87.0(-)	82.2(-)	46.7(-)	68.6(-)	81.8(-)	58.0(-)	80.7(-)	99.7(-)	98.7(-)	99.2(-)	94.5(-)
WeiNet [7]	88.2(-)	83.7(-)	64.3(-)	51.6(-)	90.0(-)	64.1(-)	88.3(-)	-	-	-	-
Proposed method	**90.1(81.9)**	**84.8(66.6)**	**65.7(42.4)**	**76.1(68.0)**	**92.5(79.6)**	**66.1(58.4)**	**90.6(76.7)**	**99.8(93.3)**	**99.0(92.0)**	**99.4(91.4)**	**95.6(66.3)**

generated virtual hand poses were applied for gesture recognition; (2) merely based on sEMG signals, in which only the sEMG signals were used for gesture recognition; (3) hybrid approach without GAN, in which the sEMG signals were combined with the virtual hand poses generated by the model learned without the GAN; and (4) hybrid approach with GAN (proposed).

The networks of the approach (1), approach (2) and approach (3) are parts of the network of the approach (4). For fair comparisons, the parameter settings of the four networks are consistent. The virtual hand pose applied for gesture recognition in approach (1) is generated by the GAN-based association model. Since the proposed framework (hybrid approach with GAN) fuses the sEMG signal stream and the virtual hand pose stream in the classification stage, the network used in approach (1) is the same as that for the virtual hand pose stream, and the network used in approach (2) is the same as that for the sEMG signal stream. The detailed network architecture can be found in Section III.B.

We compare the gesture recognition accuracy of the above four approaches in Figure 6. The upper histogram shows the accuracies of frame-based sEMG gesture recognition and the lower histogram indicates the accuracies of window-based sEMG gesture recognition. The experimental results show that whether or not the GAN is introduced, the proposed hybrid approach yields a better classification accuracy than the traditional unimodal method based solely on the sEMG signals for both frame-based and window-based sEMG gesture recognition. This finding indicates that the embedding of the virtual hand poses generated by the cross-modal association model can effectively improve the recognition accuracy. Overall, the hybrid approach with the GAN outperforms the other three approaches on all 11 sEMG benchmark databases for both frame-based and window-based sEMG gesture recognition.

D. COMPARISONS WITH STATE-OF-THE-ART APPROACHES

We compared the proposed framework with the state-of-the-art approaches [7], [12]–[17], [20]–[22] on the 7 sparse multichannel and 4 high-density sEMG benchmark databases; the results are presented in Table 3. Five state-of-the-art unimodal approaches based only on sEMG signals were considered in the comparison: AtzoriNet [14], ZhaiNet [15], GengNet [12], DuNet [13], HuNet [16] and WeiNet [7]. We also considered state-of-the-art traditional machine learning methods, such as random forest classifier (Atzori [17], Palermo [21]), SVM classifier (Pizzolato [20]) and LDA classifier (Krasoulis [22]). Previously, the best three existing works [7], [13], [16] had not been tested on all 11 benchmark databases considered in their works. To better compare with the best three existing works, we evaluate them on all 11 benchmark databases and show the results in Table 3. For DuNet [13], evaluations had been conducted only on the NinaDB1, CapgDBa, CapgDBb, CapgDBc and csl-hdemg databases; we fill in the frame-based and majority voting results on the other 6 databases in Table 3. For HuNet [16], evaluations had been conducted only on NinaDB1, NinaDB2, CapgDBa and csl-hdemg, and we fill in the window-based results on the other 7 databases in Table 3. WeiNet [7] was mainly designed to improve the recognition accuracy based on sparse multichannel sEMG signals but had been evaluated on only 6 of the NinaPro subdatabases. Therefore, in Table 3, we fill in the window-based results on NinaDB4 for WeiNet.

The proposed framework achieves a classification accuracy of 81.9% for 52 gestures based on single-frame sEMG signals and generated hand poses; this accuracy result is 5.8% higher than that of GengNet [12] and 3.8% higher than that of DuNet [13]. Our framework also achieves 4.1% higher accuracy than GengNet with a 200 ms window. For NinaDB2, the accuracy of our framework for 40 gestures is 66.6%, which is 17.0% higher than that of GengNet and 14.4% higher than that of DuNet. The proposed framework also achieves improved accuracy for frame-based gesture recognition on NinaDB5, with a result of 79.6%, which is higher than that of DuNet [13]. For NinaDB5, the accuracy of frame-based gesture recognition is 10.6% higher than that of window-based gesture recognition with a support vector machine classifier [20]. The state-of-the-art recognition results for window-based sEMG gesture recognition (200 ms) on the high-density databases are already very high and the recognition accuracies on 3 of the high-density sEMG databases are already higher than 99%, i.e., almost saturated; nevertheless, our framework achieves a +0.2% improvement. For the

remaining 8 sEMG databases, the average improvement with our framework for the window-based approach is +2.5%. For the frame-based approach, our average improvement over all 11 databases compared with the existing works is +5.7%. The overall improvements achieved with our approach are statistically significant and the detailed improvements can be seen in Table 3. These experimental results show that the proposed framework outperforms other state-of-the-art sEMG gesture recognition methods on both sparse multichannel and high-density sEMG databases.

V. CONCLUSION AND DISCUSSION

A. CONLUSION

Multimodal systems can achieve higher accuracy that unimodal systems for sEMG-based gesture recognition, but the additional required sensors reduce usability. Therefore, we propose a novel two-step pipeline classification solution for sEMG-based gesture recognition, which we have evaluated on 7 sparse multichannel and 4 high-density sEMG benchmark databases. First, we present a cross-modal association model with adversarial learning to capture the intrinsic relationship between sEMG signals and hand poses. Experimental results indicate that compared with a cross-modal association model constructed without adversarial learning, the proposed model enables improved gesture recognition accuracy based on both sparse multichannel and high-density sEMG signals, although the improvements achieved on sparse multichannel sEMG databases are higher than those achieved on high-density sEMG databases. Then, we propose our two-step pipeline classification solution for sEMG-based gesture recognition. In step one, we generate virtual hand poses using the pretrained cross-modal association model. In step two, we pair each sEMG signal with the corresponding virtual hand pose and feed the paired samples into a multimodal classifier for gesture recognition. More specifically, during the learning phase, multimodal data samples are used to train the classifier; then, during the runtime phase, our solution functions like a unimodal pipeline, recognizing gestures based solely on sEMG signals, while achieving higher accuracy.

B. DISCUSSION

Evaluations conducted on 11 sEMG benchmark databases show that the proposed two-step solution achieves significant improvements in recognition rate without requiring the user to wear additional sensors. The main reason is that we incorporate data from another mode during the training phase to establish a common basis for gesture recognition. Finally, we compare the proposed method with existing traditional machine learning and novel deep learning approaches on 11 sEMG benchmark databases. The experimental results show that the proposed method achieves significant improvements in both frame-based and window-based sEMG gesture recognition compared with state-of-the-art methods.

The ability to significantly improve the recognition accuracy through the embedding of virtual hand poses can be partially attributed to the fact that we implicitly make use of the intrinsic relationship between sEMG signals and hand poses, which provide information from different perspectives for use in sEMG-based gesture recognition. Because hand movements are driven by sEMG signals [9], we can construct a cross-modal association model, which we then can use to generate virtual hand poses based on the intrinsic physiological relationship between sEMG signals and hand poses. Then, we embed these virtual hand poses into a multimodal gesture classifier to serve as a foundation for the hybrid recognition approach. In this way, the proposed hybrid framework achieves superior recognition accuracy compared with traditional unimodal classification without the need for additional sensors. Moreover, there are many sEMG benchmark databases available that contain both sEMG signals and hand pose data, thus supporting the feasibility of the proposed framework.

Our future work will focus on incorporating new data modalities, such as IMU data, to further improve the robustness and accuracy of the proposed framework. The upgraded system will allow both users with fully intact limbs and amputees to interact with computers more efficiently and freely. We will also attempt to mine the hidden relationships among data of different modalities using the latest deep learning approaches, such as unsupervised domain adaptation and transfer learning.

REFERENCES

[1] M. Hakonen, H. Piitulainen, and A. Visala, "Current state of digital signal processing in myoelectric interfaces and related applications," *Biomed. Signal Process. Control*, vol. 18, pp. 334–359, Apr. 2015.
[2] B. Karlik, M. O. Tokhi, and M. Alci, "A fuzzy clustering neural network architecture for multifunction upper-limb prosthesis," *IEEE Trans. Biomed. Eng.*, vol. 50, no. 11, pp. 1255–1261, Nov. 2003.
[3] I. Moon, M. Lee, J. Ryu, and M. Mun, "Intelligent robotic wheelchair with EMG-, gesture-, and voice-based interfaces," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2003, pp. 3453–3458.
[4] J. Rosen, M. B. Fuchs, and M. Arcan, "Performances of hill-type and neural network muscle models—Toward a Myosignal-based exoskeleton," *Comput. Biomed. Res.*, vol. 32, no. 5, pp. 415–439, Oct. 1999.
[5] L. van Dijk, C. K. van der Sluis, H. W. van Dijk, and R. M. Bongers, "Learning an EMG controlled game: Task-specific adaptations and transfer," *Plos One*, vol. 11, no. 8, pp. 1–14, Aug. 2016.
[6] Y. Asai, S. Tateyama, and T. Nomura, "Learning an intermittent control strategy for postural balancing using an EMG-based human-computer interface," *PLoS One*, vol. 8, no. 5, 2013, Art. no. e62956.
[7] 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.*, to be published.
[8] 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 Trans Ind. Informat.*, vol. 14, no. 8, pp. 3376–3385, Aug. 2018.
[9] Y. Y. Huang, K. H. Low, and H. B. Lim, "Objective and quantitative assessment methodology of hand functions for rehabilitation," in *Proc. IEEE ROBIO*, Feb. 2009, pp. 846–851.
[10] O. Banos, J.-M. Galvez, M. Damas, H. Pomares, and I. Rojas, "Window size impact in human activity recognition," *Sensors*, vol. 14, no. 4, pp. 6474–6499, Apr. 2014.
[11] 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.
[12] W. Geng, Y. Du, W. Jin, W. Wei, Y. Hu, and J. Li, "Gesture recognition by instantaneous surface EMG images," *Sci. Rep.*, vol. 6, p. 36571, Nov. 2016.

[13] Y. Du, Y. Wong, W. Jin, W. Wei, Y. Hu, M. S. Kankanhalli, and W. Geng, "Semi-supervised learning for surface EMG-based gesture recognition," in *Proc. IJCAI*, Aug. 2017, pp. 1624–1630.

[14] M. Atzori, M. Cognolato, and H. Müller, "Deep learning with convolutional neural networks applied to electromyography data: A resource for the classification of movements for prosthetic hands," *Frontiers Neurorobotics*, vol. 10, p. 9, Sep. 2016.

[15] X. Zhai, B. Jelfs, R. H. M. Chan, and C. Tin, "Self-recalibrating surface EMG pattern recognition for neuroprosthesis control based on convolutional neural network," *Frontiers Neurosci.*, vol. 11, p. 379, Jul. 2017.

[16] Y. Hu, Y. Wong, W. Wei, Y. Du, M. Kankanhalli, and W. Geng, "A novel attention-based hybrid CNN-RNN architecture for sEMG-based gesture recognition," *Plos One*, vol. 13, no. 10, Oct. 2018, Art. no. e0206049.

[17] M. Atzori, A. Gijsberts, C. Castellini, B. Caputo, A.-G. M. Hager, S. Elsig, G. Giatsidis, F. Bassetto, and H. Müller, "Electromyography data for non-invasive naturally-controlled robotic hand prostheses," *Sci. Data.*, vol. 1, Dec. 2014, Art. no. 140053.

[18] I. Kyranou, A. Krasoulis, M. S. Erden, K. Nazarpour, and S. Vijayakumar, "Real-time classification of multi-modal sensory data for prosthetic hand control," in *Proc. 6th IEEE Int. Conf. Biomed. Robot. Biomechatronics (BioRob)*, Jun. 2016, pp. 536–541.

[19] J. Kim, M. Kim, and K. Kim, "Development of a wearable HCI controller through sEMG & IMU sensor fusion," in *Proc. URAI*, Aug. 2016, pp. 83–87.

[20] S. Pizzolato, "Comparison of six electromyography acquisition setups on hand movement classification tasks," *Plos One*, vol. 12, no. 10, Oct. 2017, Art. no. e0186132.

[21] 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.*, Jul. 2017, pp. 1154–1159.

[22] 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, p. 71, Dec. 2017.

[23] 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, p. 458, 2017.

[24] S. Shin, Y. Baek, J. Lee, Y. Eun, and S. H. Son, "Korean sign language recognition using EMG and IMU sensors based on group-dependent NN models," in *Proc. IEEE Symp. Series Comput. Intell. (SSCI)*, Dec. 2017, pp. 1–7.

[25] W. Tao, Z.-H. Lai, M. C. Leu, and Z. Yin, "Worker activity recognition in smart manufacturing using IMU and sEMG signals with convolutional neural networks," *Procedia Manuf.*, vol. 26, pp. 1159–1166, Jul. 2018.

[26] A. S. Kundu, O. Mazumder, P. K. Lenka, and S. Bhaumik, "Hand gesture recognition based omnidirectional wheelchair control using IMU and EMG sensors," *J. Intell. Robotic Syst.*, vol. 91, nos. 3–4, pp. 529–541, Sep. 2018.

[27] S. Mitra and T. Acharya, "Gesture recognition: A survey," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 37, no. 3, pp. 311–324, May 2007.

[28] P. Wang, Q. Song, H. Han, and J. Cheng, "Sequentially supervised long short-term memory for gesture recognition," *Cogn. Comput.*, vol. 8, no. 5, pp. 982–991, Oct. 2016.

[29] G. Zhu, "Multimodal gesture recognition using 3-D convolution and convolutional LSTM," *IEEE Access*, vol. 5, pp. 4517–4524, 2017.

[30] C. Yang, D. K. Han, and H. Ko, "Continuous hand gesture recognition based on trajectory shape information," *Pattern Recognit. Lett.*, vol. 99, pp. 39–47, Nov. 2017.

[31] M. Ataş, "Hand tremor based biometric recognition using leap motion device," *IEEE Access*, vol. 5, pp. 23320–23326, 2017.

[32] G. Li, H. Wu, G. Jiang, S. Xu, and H. Liu, "Dynamic gesture recognition in the Internet of Things," *IEEE Access*, vol. 7, pp. 23713–23724, 2019.

[33] D. Jiang, Z. Zheng, G. Li, Y. Sun, J. Kong, G. Jiang, H. Xiong, B. Tao, S. Xu, and H. Yu, "Gesture recognition based on binocular vision," *Cluster Comput.*, vol. 1, pp. 1–11, Feb. 2018.

[34] M. B. I. Reaz, M. S. Hussain, and F. Mohd-Yasin, "Techniques of EMG signal analysis: Detection, processing, classification and applications," *Biol. Procedures Online*, vol. 8, no. 1, p. 11, 2006.

[35] A. Phinyomark and E. Scheme, "EMG pattern recognition in the era of big data and deep learning," *Big Data Cogn. Comput.*, vol. 2, no. 3, p. 21, 2018.

[36] M. Sim ao, N. Mendes, O. Gibaru, and P. Neto, "A review on electromyography decoding and pattern recognition for human-machine interaction," *IEEE Access*, vol. 7, pp. 39564–39582, 2019.

[37] A. Phinyomark, P. Phukpattaranont, and C. Limsakul, "Feature reduction and selection for EMG signal classification," *Expert Syst. Appl.*, vol. 39, pp. 7420–7431, Jun. 2012.

[38] R. N. Khushaba, A. H. Al-Timemy, A. Al-Ani, and A. Al-Jumaily, "A framework of temporal-spatial descriptors-based feature extraction for improved myoelectric pattern recognition," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 10, pp. 1821–1831, Oct. 2017.

[39] K. Kiatpanichagij and N. Afzulpurkar, "Use of supervised discretization with PCA in wavelet packet transformation-based surface electromyogram classification," *Biomed. Signal Process. Control*, vol. 4, no. 2, pp. 127–138, 2009.

[40] A. Doswald, F. Carrino, and F. Ringeval, "Advanced processing of sEMG signals for user independent gesture recognition," in *Proc. MEDICON*, 2014, pp. 758–761.

[41] N. Patricia, T. Tommasit, and B. Caputo, "Multi-source adaptive learning for fast control of prosthetics hand," in *Proc. Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 2769–2774.

[42] R. Menon, G. D. Caterina, H. Lakany, L. Petropoulakis, B. Conway, and J. Soraghan, "Study on interaction between temporal and spatial information in classification of EMG signals for myoelectric prostheses," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 10, pp. 1832–1842, Oct. 2017.

[43] N. Jarrassé, C. Nicol, A. Touillet, F. Richer, N. Martinet, J. Paysant, and J. B. de Graaf, "Classification of phantom finger, hand, wrist, and elbow voluntary gestures in transhumeral amputees with sEMG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 1, pp. 71–80, Jan. 2017.

[44] J. Kim, S. Mastnik, and E. André, "EMG-based hand gesture recognition for realtime biosignal interfacing," in *Proc. Int. Conf. Intell. User Interfaces*, Jan. 2008, pp. 30–39.

[45] W. Wei, Y. Wong, Y. Du, Y. Hu, M. Kankanhalli, and W. Geng, "A multi-stream convolutional neural network for sEMG-based gesture recognition in muscle-computer interface," *Pattern Recognit. Lett.*, vol. 119, pp. 131–138, Mar. 2017.

[46] X. Xu, L. He, H. Lu, L. Gao, and Y. Ji, "Deep adversarial metric learning for cross-modal retrieval," *World Wide Web*, vol. 22, no. 2, pp. 657–672, 2019.

[47] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," 2016, *arXiv:1605.05396*. [Online]. Available: https://arxiv.org/abs/1605.05396

[48] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jul. 2017, pp. 5907–5915.

[49] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "Attngan: Fine-grained text to image generation with attentional generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2018, pp. 1316–1324.

[50] W. Hao, Z. Zhang, and H. Guan, "Cmcgan: A uniform framework for cross-modal visual-audio mutual generation," in *Proc. 32nd AAAI Conf. Artif. Intell.*, Apr. 2018, pp. 6886–6893.

[51] L. Chen, S. Srivastava, Z. Duan, and C. Xu, "Deep cross-modal audio-visual generation," in *Proc. Thematic Workshops ACM Multimedia 2017*, Oct. 2017, pp. 349–357.

[52] C.-H. Wan, S.-P. Chuang, and H.-Y. Lee, "Towards audio to scene image synthesis using generative adversarial network," in *Proc. ICASSP*, Jul. 2019, pp. 496–500.

[53] A. Oussidi and A. Elhassouny, "Deep generative models: Survey," in *Proc. Int. Conf. Intell. Syst. Comput. Vis. (ISCV)*, Apr. 2018, pp. 1–8.

[54] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[55] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*. [Online]. Available: https://arxiv.org/abs/1411.1784

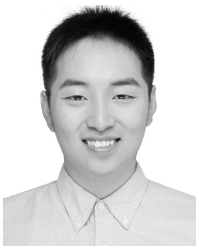
[56] C. Amma, T. Krings, J. Böer, and T. Schultz, "Advancing muscle-computer interfaces with high-density electromyography," in *Proc. CHI*, Apr. 2015, pp. 929–938.

[57] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[58] R. Chattopadhyay, N. C. Krishnan, and S. Panchanathan, "Topology preserving domain adaptation for addressing subject based variability in semg signal," in *Proc. AAAI Spring Symp. Ser.*, 2011, pp. 4–9.
[59] U. Cote-Allard, C. L. Fall, A. Campeau-Lecours, C. Gosselin, F. Laviolette, and B. Gosselin, "Transfer learning for sEMG hand gestures recognition using convolutional neural networks," in *Proc. IEEE Int. Conf. Syst., Man, (SMC)*, Oct. 2017, pp. 1663–1668.
[60] U. Côté-Allard, C. L. Fall, A. Drouin, A. Campeau-Lecours, C. Gosselin, K. Glette, F. Laviolette, and B. Gosselin, "Deep learning for electromyographic hand gesture signal classification using transfer learning," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 4, pp. 760–771, Apr. 2019.

YU HU received the B.S. degree from Xidian University, Xi'an, China, in 2012. She is currently pursuing the Ph.D. degree with the College of Computer Science and Technology, Zhejiang University, China. She is also with the Computer Animation and Perception Group, State Key Laboratory of CAD and CG, Zhejiang University. Her primary research interests include artificial intelligence, biological signal analysis, and human–computer interaction.

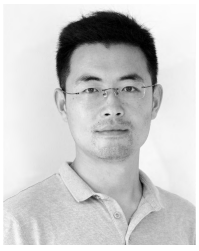
YONGKANG WONG received the B.Eng. degree from The University of Adelaide and the Ph.D. degree from The University of Queensland. He was a Graduate Researcher with the NICTA's Queensland laboratory, Brisbane, OLD, Australia, from 2008 to 2012. He is currently a Senior Research Fellow with the School of Computing, National University of Singapore. He is also the Assistant Director of the NUS Centre for Research in Privacy Technologies (N-CRiPT). His current research interests include the areas of image/video processing, machine learning, and social scene analysis.



QINGFENG DAI received the B.Sc. degree in applied mathematics from Zhejiang University, in 2016, where he is currently pursuing the Ph.D. degree in computer–human interaction and computer vision.

MOHAN KANKANHALLI received the B.Tech. degree from IIT Kharagpur and the M.S. and Ph.D. degrees from the Rensselaer Polytechnic Institute. He is currently the Provost's Chair Professor with the Department of Computer Science, National University of Singapore. He is also the Director of N-CRiPT and the Dean of the School of Computing, NUS. His current research interests include multimedia computing, multimedia security, image/video processing, and social media analysis. He is active in the Multimedia Research Community. He is on the editorial boards of several journals.

WEIDONG GENG received the B.Sc. degree from the Computer Science Department, Nanjing University, China, in 1989, the M.Sc. degree from the Computer Science Department, National University of Defense Technology, in 1992, and the Ph.D. degree from the Computer Science and Engineering Department, Zhejiang University, China, in 1995, where he is currently a Professor with the College of Computer Science. From 1995 to 2000, he was with Zhejiang University, where he took charge of a number of projects about CAD/CG and intelligent systems. In 2000, he joined the Fraunhofer Institute for Media Communication (formerly GMD.IMK), Germany, as a Research Scientist. In 2002, he was with the Multimedia Innovation Center, The Hong Kong Polytechnic University, Hong Kong. Since 2003, he has been with the State Key Laboratory of CAD & CG, Zhejiang University. His current research interests include computer-aided design, computer animation, perceptual user interface, interactive media, and digital entertainment.



XIANGDONG LI is currently an Associate Professor with the College of Computer Science and Technology, Zhejiang University. His research interests include intelligent user interfaces and cross-device computing, with a focus on leveraging the natural interaction between human and devices.