applied sciences

MDPI

Article

InRes-ACNet: Gesture Recognition Model of Multi-Scale Attention Mechanisms Based on Surface Electromyography Signals

Xiaoyuan Luo [1], Wenjing Huang [1,*], Ziyi Wang [1], Yihua Li [2] and Xiaogang Duan [3]

[1] School of Materials Science and Engineering, Central South University of Forestry, Changsha 410004, China; 20211200173@csuft.edu.cn (X.L.); wangziyi2212@163.com (Z.W.)
[2] School of Logistics & Traffic, Central South University of Forestry, Changsha 410004, China; yhli@csuft.edu.cn
[3] Central South Intelligence Collaborative Research Center, Changsha 410004, China; 20231100273@csuft.edu.cn
* Correspondence: t20142191@csuft.edu.cn

Abstract: Surface electromyography (sEMG) signals are the sum of action potentials emitted by many motor units; they contain the information of muscle contraction patterns and intensity, so they can be used as a simple and reliable source for grasping mode recognition. This paper introduces the InRes-ACNet (inception–attention–ACmix-ResNet50) model, a novel deep-learning approach based on ResNet50, incorporating multi-scale modules and self-attention mechanisms. The proposed model aims to improve gesture recognition performance by enhancing its ability to extract channel feature information within sparse sEMG signals. The InRes-ACNet model is evaluated on the NinaPro DB1 and NinaPro DB5 datasets; the recognition accuracy for these datasets can reach 87.94% and 87.04%, respectively, and recognition accuracy can reach 88.37% in the grasping mode prediction of an electromyography manipulator. The results show that the fusion of multi-scale modules and self-attention mechanisms endows a strong ability for the task of gesture recognition based on sparse sEMG signals.

Keywords: multi-scale attention mechanisms; deep learning model; sEMG signals; gesture recognition; electromyography manipulator

Citation: Luo, X.; Huang, W.; Wang, Z.; Li, Y.; Duan, X. InRes-ACNet: Gesture Recognition Model of Multi-Scale Attention Mechanisms Based on Surface Electromyography Signals. *Appl. Sci.* **2024**, *14*, 3237. https://doi.org/10.3390/app14083237

Academic Editor: Yutaka Ishibashi

Received: 13 January 2024
Revised: 8 February 2024
Accepted: 14 February 2024
Published: 11 April 2024

1. Introduction

sEMG signals constitute bioelectric signals that record sequential muscle contraction processes within skeletal muscles, encompassing crucial information regarding muscle contraction modes and intensities; they can serve as a straightforward, reliable information source for the purpose of gesture recognition. In recent years, sEMG signals have found widespread applications in diverse fields, notably in human–computer interaction [1,2] and prosthetic limb control [3–5]. Gesture recognition based on sEMG signals presents itself as a multi-classification problem within the domain of pattern recognition; currently, two primary methodologies are employed to address this challenge: traditional machine-learning methods and deep-learning methods. Traditional machine-learning approaches typically involve the extraction of features from sEMG signals in the time domain, frequency domain, or time–frequency domain. Following dimensionality reduction in these features [6,7], conventional classification methods, such as support vector machines (SVMs) [8–11], random forest (RF) [12,13], linear discriminant analysis (LDA) [14–16], and others, are subsequently applied for effective gesture recognition.

Deep-learning methods have a strong ability to learn features from data and images; they have been confirmed to perform well in pattern recognition, so in recent years, various deep-learning structures and methods have been gradually applied for gesture recognition based on sEMG signals. For instance, in Ref. [17], a convolutional neural network (CNN)

was employed for electromyography gesture recognition, yielding superior accuracy compared to the traditional SVM. Atzori et al. [18] utilized the LeNet model for recognition, achieving a recognition accuracy equivalent to traditional classification methods in 53 gesture recognition tasks. Geng et al. [19] introduced an eight-layer CNN for gesture recognition. Soroushmojdehi et al. [20] proposed a topic transfer learning method for gesture recognition on the NinaPro DB2 dataset, elevating the recognition accuracy from 81.43% to 82.87%. Zhai et al. [21] utilized a CNN for gesture recognition on the NinaPro DB2 dataset, achieving a correct recognition rate of 78.7%. Cheng et al. [22] proposed a deep CNN model for gesture recognition, with the highest recognition accuracy reaching 82.54% on the NinaPro DB1 dataset. Wei et al. [23] employed a multi-stream CNN fusion network for gesture recognition, achieving a recognition accuracy of 85%.

Attention mechanisms in deep learning have proven instrumental in enhancing system focus on key information within signals, thereby improving classification accuracy and decoding. They have been used in gesture recognition problems based on sEMG signals. For instance, Hao et al. [24] integrated attention mechanism modules into a neural network's input layer for electromyography gesture recognition based on the CapgMyo and CSLHDEMG datasets, resulting in accuracy improvements of 4.44% and 2.71%, respectively. Wang et al. [25] enhanced the LSTM-CNN network by introducing the attention mechanism CBAM, leading to a notable 5.3% increase in recognition accuracy. Fan et al. [26] proposed the CSAC-Net network model, leveraging attention mechanisms to focus on crucial information in the channel space, achieving a gesture recognition accuracy of 82.50%. Rahimian et al. [27] employed the attention mechanism and temporal convolution in the TC-HGR architecture, achieving a gesture recognition accuracy of 81.65%. Hu et al. [28] proposed a hybrid CNN-RNN network structure based on the attention mechanism, achieving an average gesture recognition accuracy of 84.80% based on the NinaPro DB1 dataset.

Additionally, multi-scale modules in deep learning use convolution kernels and pooling operations of various sizes concurrently to facilitate feature extraction across different scales. This multi-scale design allows for the capture of information from diverse-sized areas in the image, thereby enhancing the model's perceptual capabilities. This approach has found application in electromyography gesture recognition as well. For example, Han et al. [29] introduced a novel CNN incorporating multi-scale kernels and feature fusion (MKFF-CNN), and it was applied to gesture recognition based on sEMG signals, resulting in a significant 6.54% increase in recognition accuracy compared to a single-scale convolutional subnetwork. Shen et al. [30] proposed an sEMG signal gesture classification model based on a multi-scale module, achieving a recognition accuracy of 79.43% on NinaPro DB5. Jiang et al. [31] introduced an RIE model based on inception multi-scale fusion convolution and the ECA mechanism, which achieved an average accuracy of 88.27% on NinaPro DB1, surpassing the traditional CNN by 7.89%.

There is multi-dimensional information in sEMG signals, such as the time domain, frequency domain, or time–frequency domain, so deep-learning methods for gesture recognition based on sEMG signals include two predominant approaches. The first involves feature extraction from the original sEMG signals, the subsequent conversion of features into maps, and finally the inputting of these maps into the deep-learning model. In this approach, manual feature extraction is required. The second involves the direct conversion of the original sEMG signals into an electromyography image, enabling the deep-learning model to autonomously learn sophisticated features in the gesture recognition process. This approach can avoid the step of manually extracting features and simplify the recognition process.

Hence, this study integrates both the attention mechanism and a multi-scale structure into the ResNet50 model in order to enhance the model's proficiency in capturing various receptive field image features and critical area information within sEMG signals. Simultaneously, features in sEMG signals can be autonomously learned, with no manually feature extracting required.

This study integrates the attention mechanism with a multi-scale structure into a deep-learning model, proposing the InRes-ACNet model based on ResNet50. Initially, to tackle challenges associated with the potential loss of channel information during feature extraction from sEMG signals, we introduce a multi-scale inception–attention module. This module aims to enhance the model's ability to extract features related to channel information. Furthermore, we incorporate the ACmix module into the ResNet50 model, providing a synergistic blend of the self-attention mechanism and convolution operations. This integration seeks to enhance the model's feature extraction capabilities. The simultaneous use of both the inception–attention and ACmix modules is intended to improve the model's proficiency in gesture recognition based on sEMG images. Additionally, the model's inputs are the original grayscale images of the sEMG signal. This approach eliminates the need for manual feature extraction during sEMG signal preprocessing, thereby streamlining the gesture recognition process.

The rest of this paper is organized as follows: Section 2 describes the preprocessing of original sEMG signals and the generation of sEMG grayscale image datasets. Section 3 introduces the proposed InRes-ACNet model. Section 4 verifies the effectiveness of the InRes-ACNet model through experiments and conducts grasping mode prediction for the electromyography manipulator. Section 5 summarizes the results of this paper's work.

2. Methods

The flow of grasping mode recognition for the manipulator based on sEMG signals is illustrated in Figure 1. Arm sEMG signals are collected by a wireless 8-channel sEMG armband, and the sEMG data are transmitted to the computer in real-time via Bluetooth. The computer, through software, communicates with the Bluetooth serial port and obtains the current sEMG data according to the serial communication protocol. Subsequently, a deep-learning model for predicting the manipulator's grasping mode is trained using sEMG data. The trained model is then integrated with the manipulator environment to achieve control of the manipulator's grasping mode, which has multiple degrees of freedom.

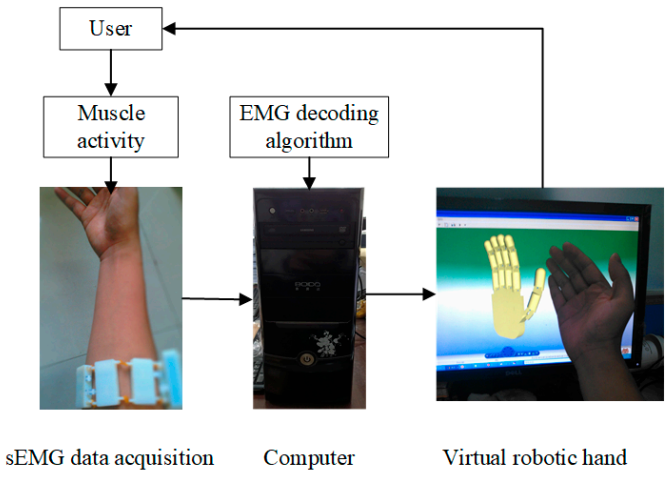

Figure 1. Basic flow chart of grasping mode recognition of manipulator.

2.1. sEMG Signals Acquisition of Wireless 8-Channel Armband

sEMG signals are collected using an 8-channel wireless sEMG armband developed by researchers at Shanghai Jiao Tong University. This armband, capable of recognizing 8 hand gestures, is illustrated in Figure 2. The armband consists of 8 dry electrodes, a DC power supply, and a microcontroller. Circuit-wise, the 8 dry electrodes are interconnected via flexible flat cables (FPC), which individually transmit the collected signals to the microcontroller for further processing. Structurally, the electrodes are paired and connected by elastic cords, allowing the armband to accommodate arms of various thicknesses,

as shown in Figure 2a,b. Each dry electrode incorporates a differential amplification circuit for filtering and amplifying sEMG signals. The signal processing unit, centered around the microcontroller, refilters, amplifies, and converts analog signals to digital before transmitting them to a computer or other devices via Bluetooth. The receiving device reads the 8-channel EMG signals in real-time following the serial communication protocol. The sEMG armband has a sampling frequency of 1000 Hz, an amplification factor of 500, and a baud rate of 57,600 bits/s.

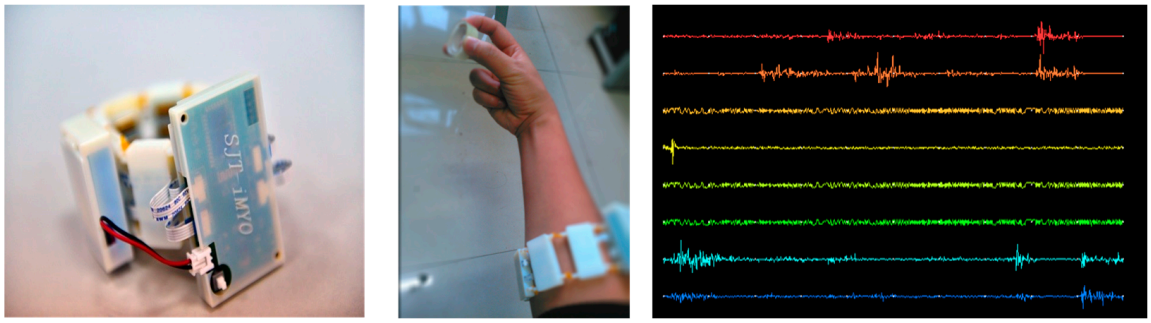

Figure 2. sEMG signals acquisition: (**a**) Wearable 8-channel wireless sEMG armband. (**b**) Grasp mode with sEMG armband. (**c**) sEMG signals acquisition.

Subjects participating in the data acquisition using an 8-channel wireless sEMG armband performed eight hand gestures in each cycle, with each gesture lasting 5 s. Upon hearing an announcement through headphones, subjects changed their grasp gestures. After each data acquisition cycle, subjects rested for 1 min before proceeding to the next cycle. From each subject, sEMG data totaling 100 s were collected, accumulating to 1000 s of sEMG data. Figure 2c shows the first round of sEMG signals collected from a subject. The eight hand gestures are natural grasp, spherical grasp, cylindrical grasp, tripod pinch, tip pinch, lateral pinch, hook grasp, and thumb extension grasp, as depicted in Figure 3.

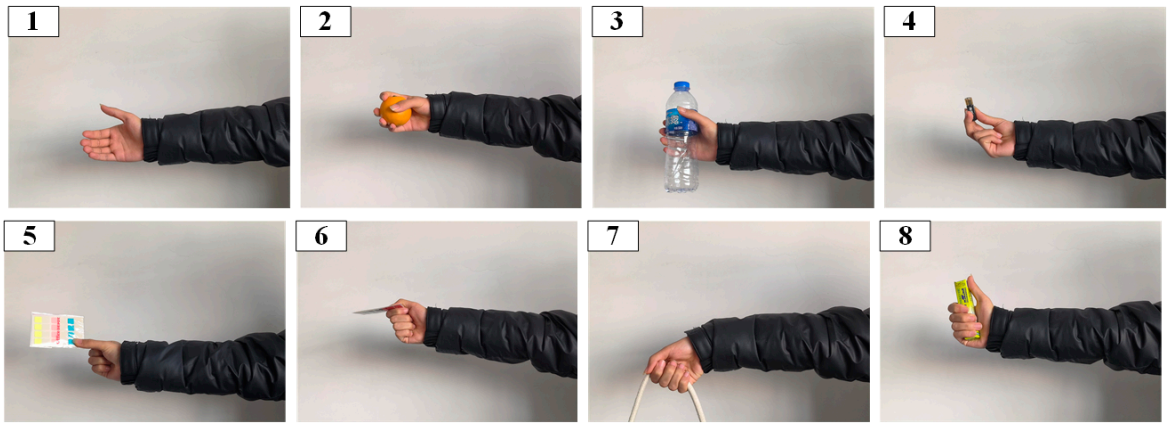

Figure 3. Eight hand gesture actions: 1, natural grasp; 2, spherical grasp; 3, cylindrical grasp; 4, tripod pinch; 5, tip pinch; 6, lateral pinch; 7, hook grasp; 8, thumb extension grasp.

2.2. sEMG Processing

The original sEMG signals include electrical activities from muscle contractions and relaxations, potentially carrying noise from power-line interference, motion artifacts, baseline drifts, and recording equipment [32]. Such noise significantly impacts the accuracy of hand gesture recognition. Butterworth filters, which can uniformly transfer all frequency components while maintaining the time responses of signals without a sharp transition between the passband and stopband, effectively preserve the original signal form. Therefore, we

used a full-wave rectifier and a Butterworth filter for noise filtering of the original signals. Initially, a full-wave rectifier processed signals collected by the 8-channel wireless sEMG armband. Subsequently, a 1 Hz Butterworth filter was employed for low-pass filtering to remove high-frequency noise. As shown in Figure 4, the sEMG signal curves become smoother after noise filtering, demonstrating the Butterworth filter's effective denoising. This ensures the quality of subsequent training for the hand gesture recognition model.

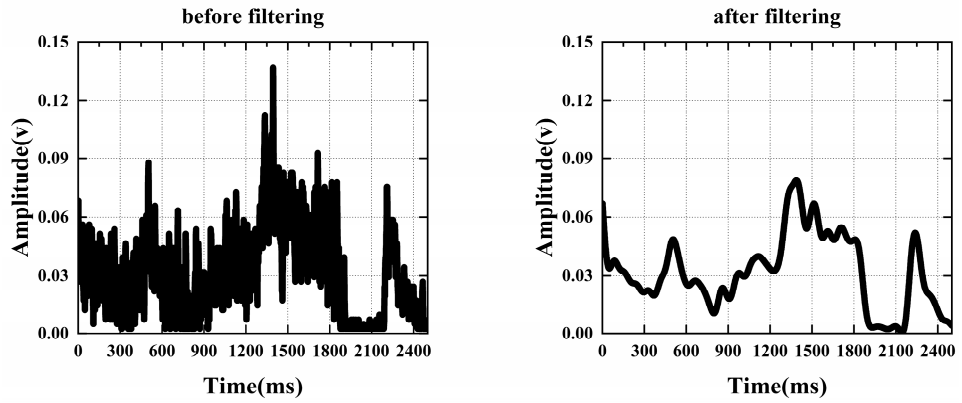

Figure 4. Comparison of sEMG signals before and after Butterworth filtering.

2.3. Construction of Grayscale Image Dataset of sEMG Signals

In the field of sEMG gesture recognition, when a deep-learning model is utilized for image recognition tasks, it is often necessary to convert one-dimensional sEMG signals into two-dimensional sEMG grayscale images for input into the deep model. In this paper, we form the sEMG matrix using a sliding window, and this matrix is transformed into an sEMG image via a mathematical mapping equation. The size of the generated sEMG image is detailed in Equation (1).

$$Image \in S^{C \times W \times H} \tag{1}$$

In this paper, we utilize original sEMG signals where the number of picture channels (C) equals the number of features, with $C = 1$ for original sEMG. Arturo et al. [12] demonstrated that a 150–250 ms window for sliding sampling of sEMG signals was optimal. Considering the computational power of computers and the requirements for real-time processing, the picture height (H), which is the total number of sliding windows in a 200 ms time period, is set to 20 for original sEMG; the picture width (W), representing the number of sampled electrode channels, is 10. We used only the original sEMG signals for grayscale image transformation without feature extraction. This approach simplifies the training process and reduces the number of parameters involved.

Each group of sEMG signals is filtered, and the filtered data are then mapped to the grayscale interval [0, 1], resulting in the creation of an sEMG grayscale image. Figure 5 presents a schematic diagram of the signal-to-image conversion process. The equation for maximum–minimum normalization is shown in Equations (2) and (3).

$$X_{[0,1]} = F(X_{(i,j)}) \tag{2}$$

$$F(X_{(i,j)}) = (X_{(i,j)} - X_{\min})/(X_{\max} - X_{\min}) \tag{3}$$

where $X_{[0,1]}$ is the sEMG matrix after mapping conversion, with element values ranging from 0 to 1; $F(\ )$ is the mapping function that converts one-dimensional signals to two-dimensional images; $X_{(i,j)}$ represents the value at row i and column j in the sEMG matrix, formed by a sliding window, where $0 < i \leq H$ and $0 < j \leq W$; $X_{\max}$ is the largest value within the sEMG matrix formed by the sliding window. Similarly, $X_{\min}$ is the smallest value in the matrix.

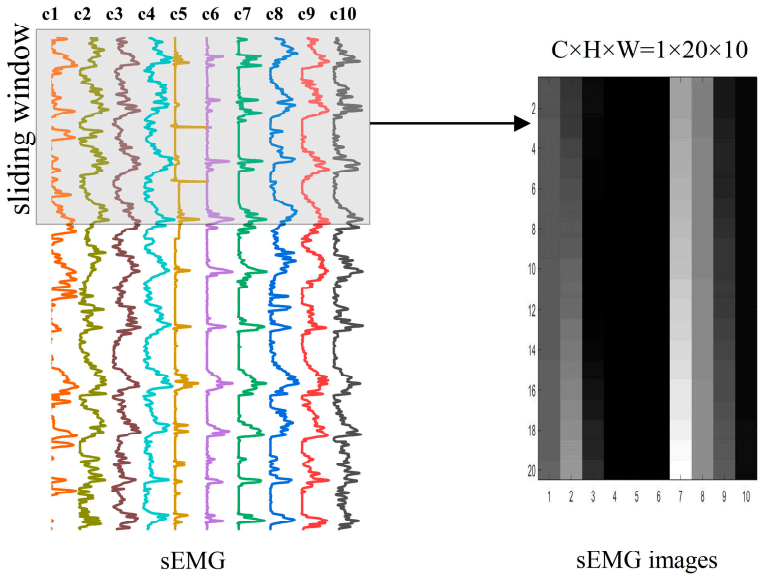

Figure 5. Imaging process of sEMG signals.

3. Construction of Recognition Model

3.1. InRes-ACNet Model

An InRes-ACNet model, based on the ResNet50 model [33], is presented in Figure 6. Initially, we design an inception–attention module for multi-scale feature extraction, enhancing the model's ability to capture abstract, complex, and representative features. This module, integrated into the InRes-ACNet model, employs a multi-scale attention mechanism, focusing on important features at various scales to improve detail capture in grayscale images. By incorporating the inception–attention module before ResNet50, the model benefits from early acquisition of multi-scale gray features, providing a strong foundation for subsequent feature learning.

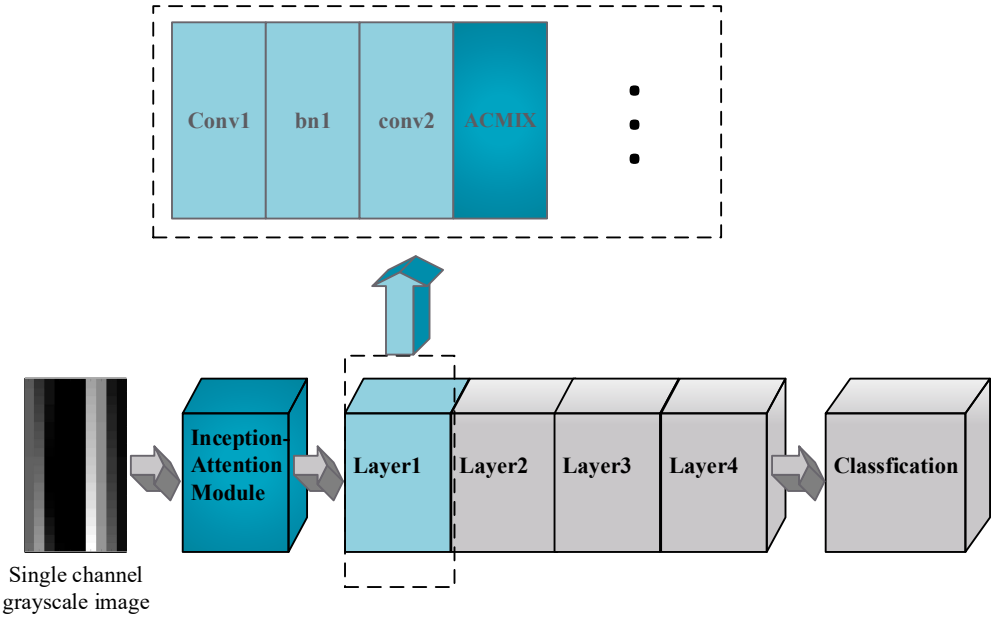

Figure 6. InRes-AcNet model.

Furthermore, to enhance the attention and representation capabilities of the ResNet50 model, an ACmix module is incorporated. This module merges convolution and self-attention mechanisms, thereby improving the model's ability to learn feature relevance

while retaining high efficiency. With the ACmix module integrated into ResNet50, the network more effectively captures long-range dependencies between features, thus enhancing the model's expressiveness and generalization capabilities.

3.2. Inception–Attention Module

The inception module [34], a classical feature extraction module, was first introduced by GoogLeNet. It captures feature information at various scales using different-sized convolution kernels and parallel pooling. These features are then concatenated in the channel dimension. This approach allows the model to simultaneously consider local details and global information at different scales, thereby enhancing its ability to understand images.

To extract sEMG features, this paper introduces the inception–attention module, depicted in Figure 7. This module combines the inception architecture with the SE (squeeze-and-excitation) module, a lightweight attention mechanism. The SE module adaptively learns the relationships between channels, allowing for the dynamic adjustment of each channel's importance. To fit the input requirements of the inception–attention module, we resize each pre-processed grayscale image to 224 × 224 pixels.

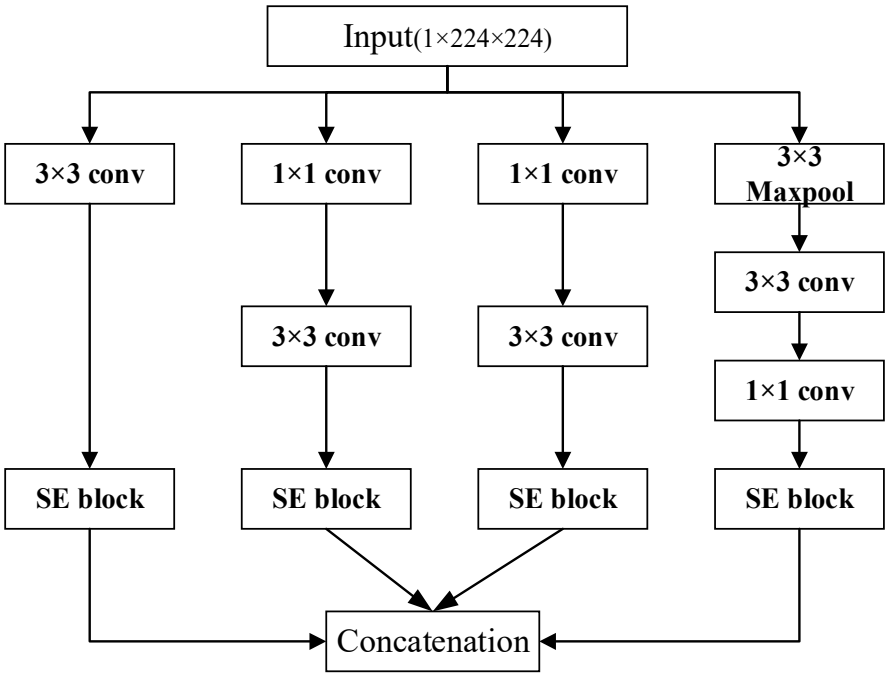

Figure 7. Inception–attention module.

The development of the inception–attention module involves the following process: Four branches are established using convolution kernels of different sizes (3 × 3 and 1 × 1), enabling the extraction of information across various spatial scales. Specifically, the first branch employs 3 × 3 convolution kernels to extract features and capture local details. The two middle branches first use 1 × 1 convolution kernels to increase the channel count, followed by 3 × 3 convolution kernels for feature extraction. The fourth branch uses a 3 × 3 max-pooling operation to highlight important information in the images, followed by feature extraction with 3 × 3 convolution kernels, and then alters the channel count using a 1 × 1 convolutional layer. To enhance the representation capability of the inception–attention module further, an SE block is integrated into each branch. The outputs of the four branches are then concatenated in the channel dimension, forming the final output of the inception–attention module.

The four branches of the inception–attention module capture diverse features of sEMG signals across multiple scales, enhancing adaptation to the complexity of sEMG signal

grayscale maps. Integrating the SE block into the inception–attention module allows for the reweighting of features across channels, enabling the module to adaptively learn channel relationships and optimize feature representation. Features extracted by each branch are then fused in the channel dimension, creating varied feature representations and enriching the information provided to subsequent tasks. With these improvements, the inception–attention layer is better suited to the characteristics of sEMG signal grayscale maps, thus enhancing the performance and effectiveness of the feature extraction layer in sEMG-related tasks.

3.3. ACmix Module

The ACmix module [35], proposed by researchers at Tsinghua University, introduces a new deep-learning architecture specifically designed to enhance the ability of convolutional neural networks to process complex features. This module combines convolution and self-attention operations, as illustrated in Figure 8. Initially, feature maps are processed through three 1×1 kernel convolutional layers, producing $3 \times N$ intermediate feature maps. Subsequently, based on the initial feature conversion, two operations are executed: a convolution operation via a lightweight fully-connected layer to transform feature maps and generate k^2 new feature representations; and the introduction of a self-attention mechanism to compute attention weights for the query, key, and value, followed by their fusion. Finally, the outcomes of the convolution and self-attention processes are integrated with varying weights at the ACmix module's conclusion using the following formula:

$$F_{out} = \alpha F_{att} + \beta F_{conv} \tag{4}$$

where α and β are learnable parameters; F_{att} is the weight obtained by self-attention operation; F_{conv} is the feature obtained by convolution operation; and F_{out} is the fusion output.

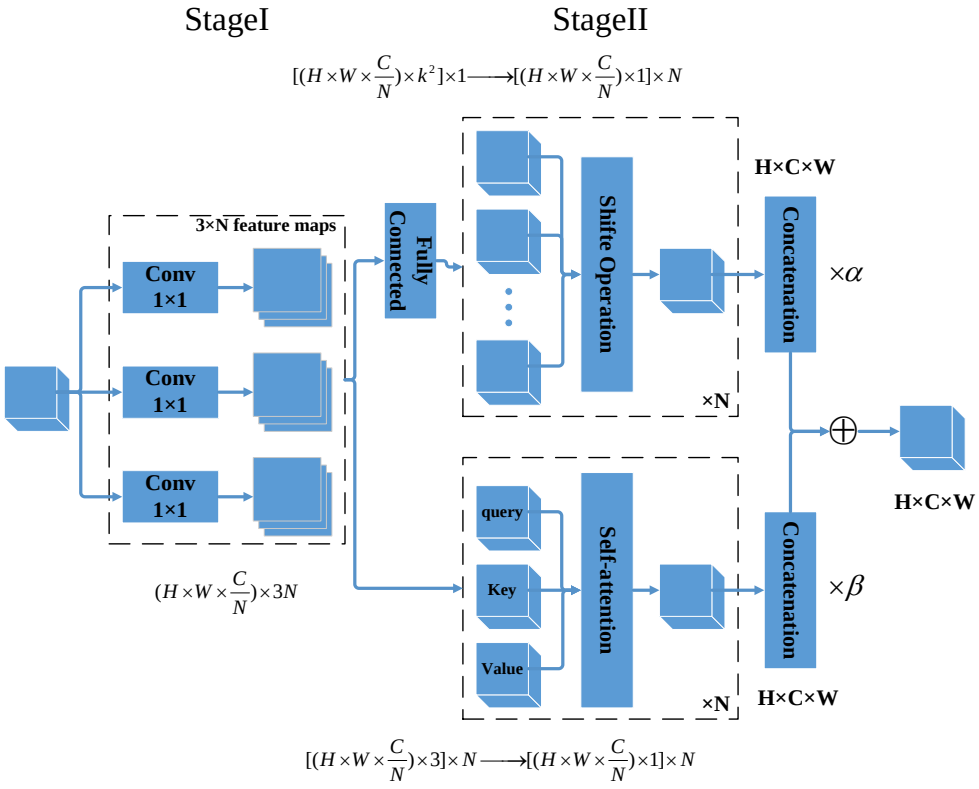

Figure 8. ACmix module.

In this research, ACmix is integrated into the ResNet50 model to enhance the traditional convolutional network's ability to process high-level features. The ACmix model introduces a self-attention mechanism at the network's early stages, enabling the capture of more comprehensive information from sEMG signals while maintaining sensitivity to local features. The integration of the ACmix module into ResNet50 occurs as follows: The ACmix module is inserted immediately after the first convolutional layer (conv2d) within the first residual block (layer1) of the ResNet50 network, as shown in Figure 6. Initially, the original feature maps are processed by ResNet50′s initial convolutional layer, then undergo deep feature extraction and transformation within the ACmix module. The output from the Acmix module serves as the input for the subsequent residual block, facilitating close collaboration between ACmix and ResNet50′s other modules. Through this strategy, a more robust and adaptable model is developed, achieving an enhanced performance across various complex grayscale image vision tasks.

4. Model Training and Experimental Results

Two kinds of experiments were conducted: the first experiment was based on the publicly available dataset NinaPro DB1 and NinaPro DB5; the second experiment was based on sEMG signals collected by 8-channel wireless sEMG armband.

4.1. Experiment Based on Publicly Available Dataset

The NinaProDB1 dataset encompasses 52 hand gestures, classified into three categories for training: (A) 12 finger gestures; (B) 8 gestures of uniform opening and closure with equal length, along with 9 wrist gestures; (C) 23 basic grasp gestures. A total of 10 healthy volunteers (7 males and 3 females, aged 22 to 30 years old, without any medical history and of similar physique) participated in the collection of sEMG signals from forearm muscles using an ELONXI electromyography system equipped with 18 dry electrodes, including 2 for grounding. The sampling frequency was set to 100 Hz. Each participant performed the 52 hand gestures 10 times, with each gesture lasting 10 s. A 3-s rest was allowed between gestures, and subjects could rest for 10 min between different sessions.

The NinaProDB5 datasets encompass 52 hand gestures, classified into three categories for training: (A) finger gestures; (B) gestures characterized by uniform opening and closure of equal length, along with wrist gestures; (C) basic grasp gestures. Ten healthy volunteers participated in the collection of sEMG signals, utilizing a 16-channel sampling system. The sampling frequency was set at 200 Hz. Each participant was required to perform the 52 hand gestures in the experiment, repeating each gesture 6 times, with each repetition lasting 5 s. Participants were allowed a 3-s rest period between different hand gestures.

In the experiment, the sEMG data from the second and sixth repetitions of each gesture constituted the test sets, while the remaining data formed the training and validation sets. The distribution among the training, validation, and test sets was approximately in a 6:2:2 ratio. We employed data augmentation techniques, such as random and center cropping, to enhance the diversity of the training data. Each recognition model was trained individually, and the corresponding trained model was then utilized for each individual. The batch size for training was set at 64, with the Adam optimizer (adaptive moment estimation method) selected. The learning rate was established at 0.001, and a dropout rate of 0.4 was applied to improve the model's generalization capability.

4.1.1. Ablation Experiment

We conducted comparative experiments using four distinct network architectures: the ResNet50 model, the ACResNet model (which integrates ACmix into ResNet50), the InResNet model (introducing inception–attention into ResNet50), and the InRes-ACNet

model. We employed the cross-entropy loss function for forward propagation, a standard approach in addressing multi-classification problems, as illustrated in Equations (5) and (6).

$$loss = \frac{1}{S}\sum_{i}^{S} loss_i \tag{5}$$

$$loss_i = -\sum_{c=1}^{M} y_{ic} \log(p_{ic}) \tag{6}$$

where $loss$ is the average loss rate; $loss_i$ is the loss rate of the ith sample; S is the total number of samples; M is the number of classes; The value of y_{ic} can be 0 or 1, it is 1 if the true class of the ith sample is to class c, 0 otherwise; p_{ic} is the predicted probability that the observed sample i belongs to class c.

The results of the ablation experiments are presented in Table 1. The ResNet50 model demonstrated an accuracy of 82.71% in the sEMG gesture recognition test, attributed to its robust feature extraction capabilities embedded in the deep convolutional structure. The ACResNet model achieved an increased accuracy rate of 85.98%, owing to the incorporation of the ACmix module, which facilitated more precise information integration, thereby enhancing the model's ability to comprehend and characterize image content.

Table 1. Ablation experiment of the InRes-ACNet model.

Method	Params	Best Test Acc
ResNet50	24.03 M	82.71%
ACResNet	24.04 M	85.98%
InResNet	24.25 M	86.31%
InRes-ACNet	24.26 M	87.82%

The InResNet model demonstrated a slight improvement in accuracy to 86.31%. This enhancement is attributed to the integration of the inception–attention module, which enables the model to capture multi-scale feature information. The InRes-ACNet model, combining the inception and ACmix modules, achieved a notable accuracy of 87.82% in sEMG gesture recognition, significantly surpassing other individually enhanced models. Compared with the ResNet50, the recognition accuracy of the ACResNet and InResNet models increased by 3.24% and 3.6%, respectively. The InRes-ACNet model's accuracy improved by 5.11%, with only a 0.23 M increase in the number of parameters. As indicated by the parameter values in Table 1, the InRes-ACNet model effectively extracts both local details and global information of sEMG signals, thanks to its improved module structure, while maintaining parameter efficiency, thus significantly enhancing model performance.

In experiments, it was demonstrated that both the inception–attention module and the ACmix module individually enhance the model's recognition performance. The InRes-ACNet model, which integrates these two modules simultaneously, exhibits a superior classification performance in sEMG gesture recognition tasks.

4.1.2. Comparative Analysis of Identification Performance of Different Model

This experiment utilized the NinaPro DB1 dataset, covering 17 different gestures. The InRes-ACNet model, as proposed in this study, was employed for sEMG gesture recognition. The experimental results were compared with mainstream recognition methods reported by other researchers, as illustrated in Table 2. The results indicated that traditional sEMG gesture recognition methods, such as SVM and the random forest algorithm, yielded relatively low gesture recognition rates of 69.45% and 75.36%, respectively. The adoption of deep-learning methods, specifically the convolutional neural network variant MyoCNN and VGGNet models, significantly enhanced the recognition performance, with rates of 78.25% and 81.12%, respectively. Furthermore, the accuracy of the MSCNet model

increased to 83.24% through the extraction of multi-scale features. The InRes-ACNet model proposed in this study, integrating a multi-scale module, the ResNet50 architecture, and the ACmix module, achieved a recognition rate of 86.46%, demonstrating considerable promise in gesture recognition. This outcome further substantiates the efficacy of the attention mechanism and multi-scale feature fusion strategy in enhancing the performance of sEMG gesture recognition tasks.

Table 2. Gesture recognition results of different models.

Method Author	Model	Recognition Acc
Pizzolato	SVM [7]	69.45%
Atzori	Random forests [12]	75.36%
Wei	MyoCNN [23]	78.25%
Atzori	VGGNet [12]	81.12%
Simonyan	MSCNet [36]	83.24%
This paper	InRes-ACNet	86.46%

4.1.3. Individual Variability Analysis

The primary objective of this research is to assess the model's performance when applied to various datasets, with a particular focus on evaluating its recognition performance across different gestures of different individuals. The experiment involved four datasets: NinaPro DB1 E1, NinaPro DB1 E2, NinaPro DB5 E1, and NinaPro DB5 E2.

Figure 9 illustrates the validation effect of using the InRes-ACNet model on the NinaPro DB1 E1 dataset, which encompasses 12 different gestures. It presents the validation loss values and recognition accuracy for five subjects across 50 validation cycles (epochs). In the initial stages of validation, the loss values for all subjects decreased rapidly, while the accuracy rates increased significantly, nearing their peak levels. This indicates that the model exhibits a fast convergence rate and robust learning capabilities. Throughout the validation process, despite minor fluctuations, the loss and accuracy curves generally remained stable, demonstrating the recognition stability of the InRes-ACNet model.

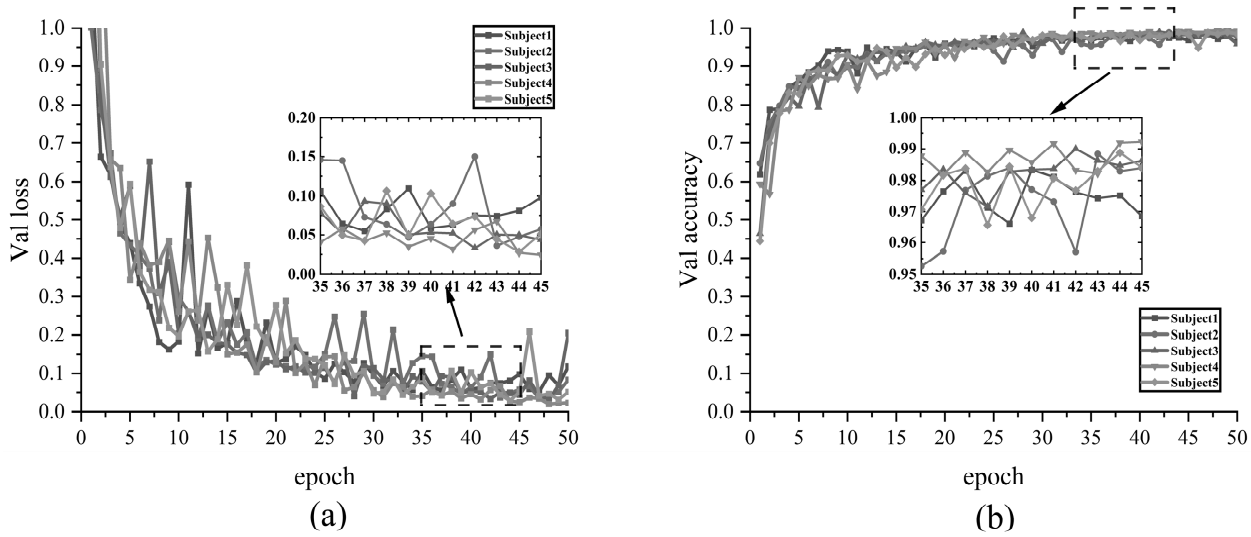

Figure 9. Validation loss and accuracy for NinaPro DB1 E1. (**a**) Validation loss for NinaPro DB1 E1. (**b**) Validation accuracy for NinaPro DB1 E1.

Figure 10 illustrates the validation effect of employing the InRes-ACNet model on the NinaPro DB1 E2 dataset, encompassing 17 different gestures. It presents the validation loss values and recognition accuracy for five subjects across 50 validation cycles (epochs).

At the initial stage of validation, the loss value decreased rapidly, while the accuracy rate ascended to 95%, indicating the model's rapid learning capability and convergence speed. Despite minor fluctuations in loss values and accuracy between the 40th and 50th epochs, the model demonstrated stability and high accuracy throughout the entire gesture recognition process.

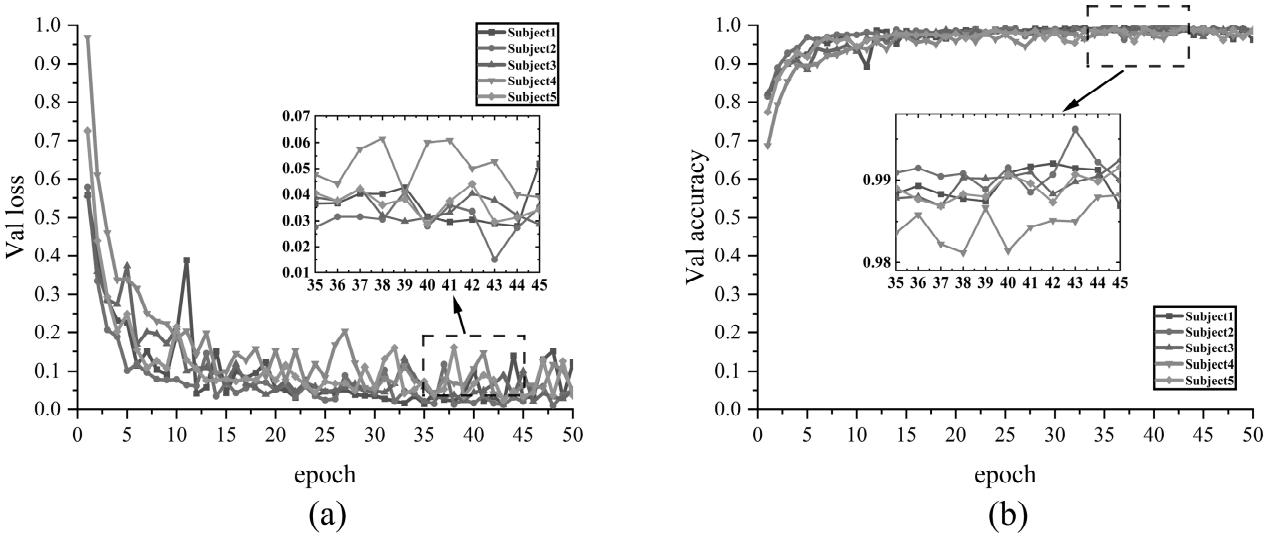

Figure 10. Validation loss and accuracy for NinaPro DB1 E2. (**a**) Validation loss for NinaPro DB1 E2. (**b**) Validation accuracy for NinaPro DB1 E2.

The variability experiment results for the datasets NinaPro DB1 E1, NinaPro DB1 E2, NinaPro DB5 E1, and NinaPro DB5 E2 are presented in Table 3, with the boxplot of recognition accuracy for these four datasets depicted in Figure 11. This boxplot illustrates the distribution of recognition accuracy for each dataset, including statistical measures such as the quartile, median, and mean. The NinaPro DB1 E1 and E2 datasets cover 12 and 17 gestures, respectively. As indicated in Table 3, the average recognition accuracy for these datasets was 87.94% and 86.00%, respectively, with individual variances of 8.04% and 6.25%. The boxplot reveals that the median recognition accuracy closely aligns with the mean, showing a relatively balanced distribution without extreme outliers or significant skewness. For the NinaPro DB5 E1 and E2 datasets, which cover 12 and 17 gestures respectively, the average recognition accuracy was noted as 87.04% and 85.39%, respectively, with individual differences of 8.45% and 8.16%. The boxplot indicates that, despite the quartiles' wide range, the median remains close to the mean, suggesting that the model maintains a consistent performance even in more complex gesture recognition tasks. Overall, the results demonstrate the model's consistent performance across different datasets in gesture recognition tasks.

Table 3. Recognition performance of different datasets.

Dataset	Average Recognition Rate	Recognition Rate Distribution	Number of Movements
NinaPro DB1 E1	87.94%	8.04%	12
NinaPro DB1 E2	86.00%	6.25%	17
NinaPro DB5 E1	87.04%	8.45%	12
NinaPro DB5 E2	85.39%	8.16%	17

Overall, the InRes-ACNet model achieves a high accuracy rate in various gesture recognition tasks, demonstrating its effectiveness in capturing and utilizing key features

in sEMG images. While recognition accuracy rates varied across datasets, the model's performance underlines its effectiveness and reliability in sEMG gesture recognition. This further validates the InRes-ACNet model's generalization capability across diverse datasets.

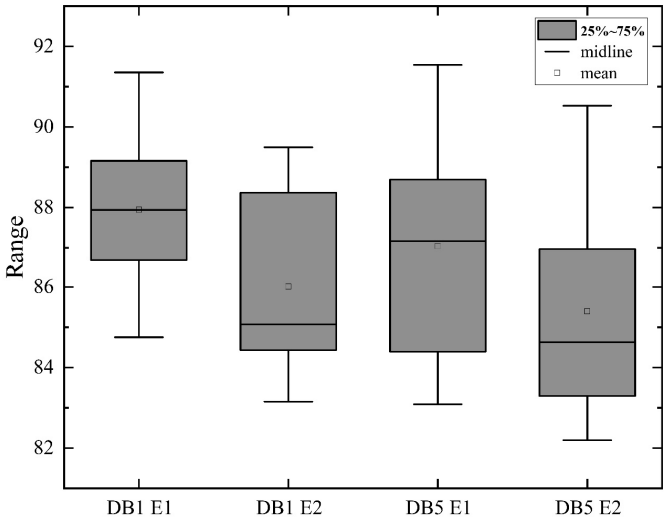

Figure 11. Boxplot of recognition accuracy for different datasets.

4.2. Experiment Based on sEMG Signals from 8-Channel Wireless Armband

This paper presents the integration of the inception–attention module and the ACmix module to develop the InRes-ACNet model as a gesture recognition framework. Initially, a multi-scale inception–attention module is constructed and integrated into the ResNet50 model. Subsequently, the self-attention mechanism, the ACmix module, is incorporated into ResNet50, resulting in the formulation of the InRes-ACNet gesture prediction model. For online prediction, the InRes-ACNet model is employed for gesture recognition tasks.

Experimenters donned an sEMG armband and executed eight types of gestures on a manipulator grasping mode control platform (Figure 12). For each gesture, 480 muscle signal segments were extracted and transformed into 480 sEMG grayscale images, yielding a dataset size of 3840 for each experimenter. To assess the model's robustness, ten diverse experimenters, including seven males and three females, were selected. Each experimenter replicated the specified actions to capture muscle electrical signals. The dataset from each experimenter was utilized to train the model, allocating 60% of the data for training, 20% for validation, and 20% for testing. For the test database, twenty groups of data corresponding to different gestures were randomly selected, totaling 160 data groups, and multiple rounds of testing were conducted on the outcomes. A subset of the test results is displayed in Figure 13.

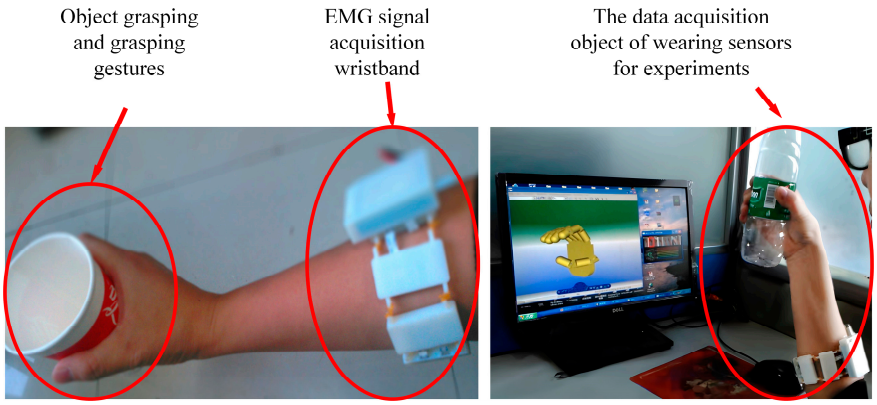

Figure 12. Experimental scene diagram.

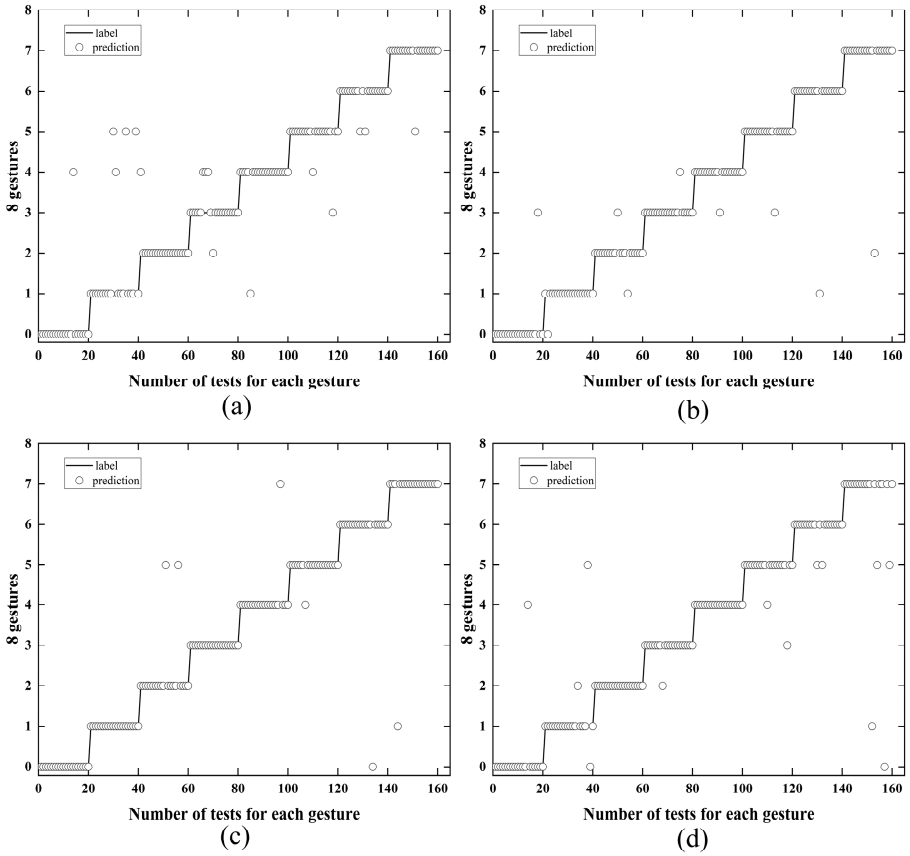

Figure 13. Comparison of test results for four experiments (to demonstrate the robustness of the model, we randomly selected some data and performed gesture recognition with different experimenters, obtaining comparative results. (**a**) is the gesture predictions for experimenter 1; (**b**) is the gesture predictions for experimenter 2; (**c**) is the gesture predictions for experimenter 3; (**d**) is the gesture predictions for experimenter 4).

The test results from four experimenters, as depicted in Figure 13, demonstrate that the InRes-ACNet model delivers commendable recognition performance in gesture recognition tasks. Given that the test data are randomly selected, this underscores the robustness of the InRes-ACNet. By integrating both the attention mechanism and a multi-scale structure, the InRes-ACNet showcases a high recognition accuracy and stability in human gesture recognition challenges.

The average recognition accuracy across ten experimenters is 88.37%. To elaborate on the results obtained from our model, we have selected four illustrative examples. These examples depict the gesture tests conducted by four experimenters. Figure 13a presents the gesture predictions for Experimenter 1, with an average recognition accuracy of 84.63%. Notably, there are a considerable number of deviations from the target category, such as the circle, indicating that the predicted gestures significantly diverge from the actual gestures. This is especially true for Gesture 2 (spherical grasping), which deviates markedly from its actual form. The figure illustrates that each gesture is prone to misidentification, such as Gesture 5 (two-finger pinch) and Gesture 6 (side pinch). Figure 13b displays the gesture predictions for Experimenter 2, achieving an average recognition accuracy of 90.63%. Most of the predicted category tags align closely with the actual category line, indicating more precise gesture recognition. Figure 13c reveals the gesture predictions for Experimenter 3, with an average recognition accuracy of 92.96%. The minimal deviation from the actual category (i.e., circles) suggests highly accurate gesture recognition results. Figure 13d showcases the gesture predictions for Experimenter 4, with an average recognition accuracy of 87.32%. Despite the recognition errors, particularly for Gesture 8 (thumb extension grip), the findings indicate variability in recognition accuracy across different experiments.

This variability may stem from several factors, such as inconsistent muscle strength exerted by experimenters during gesture performance or discrepancies in movement standards. Despite these variations, the gesture recognition model maintains a high accuracy and stability, underscoring the InRes-ACNet model's robust capability in gesture recognition tasks, particularly in natural grasp, cylindrical grasp, and tip pinch.

For comparative purposes, the ResNet50 model was also employed in gesture prediction tasks. Utilizing the same ten experimenters as in the previous InRes-ACNet model tests, the gesture prediction results are summarized in Table 4. The InResNet50 model achieves an average recognition accuracy of 85.63%, while the InRes-ACNet model attains an average recognition accuracy of 88.37%, highlighting the superior performance of the InRes-ACNet model in gesture prediction tasks.

Table 4. Comparison of recognition performance for the models ResNet50 and InRes-ACNet.

Model	Average Recognition Rate	Epoch	Number of Movements
ResNet50	85.63%	50	8
InRes-ACNet	88.37%	50	8

5. Conclusions

In this study, by incorporating an attention mechanism and multi-scale module into the ResNet50 model, we propose the InRes-ACNet model for gesture recognition based on surface electromyography (sEMG) signals. Initially, the study constructs the inception–attention module based on the multi-scale inception module, which is then integrated into the ResNet50 model to enhance its multi-scale feature extraction capabilities. Subsequently, the self-attention mechanism, the ACmix module, is incorporated into ResNet50, enabling the model to maintain a lower parameter count while improving its feature extraction performance. Ultimately, employing the InRes-ACNet model on the NinaPro DB1 and NinaPro DB5 datasets for gesture recognition yielded accuracy rates of 87.94% and 87.04%, respectively. Additionally, the InRes-ACNet model was applied to the prediction of grasping modes in an electromyography manipulator, achieving an average recognition accuracy of 88.37%. These results confirm the effectiveness of the InRes-ACNet model for gesture recognition tasks based on sEMG signals.

Our research enhances gesture recognition performance by incorporating multi-scale modules and attention mechanisms, alongside utilizing the grayscale images of the original sEMG signals. Nevertheless, the InRes-ACNet model, integrating the inception–attention module, ACmix module, and ResNet50, entails a considerable number of parameters. This complexity results in extensive computations and slower training speeds. Variations in the dataset sizes, attributable to the differing collection times and frequencies for each gesture, also affect the duration of each training step. Given the non-stationary and random characteristics of sEMG signals, coupled with significant variations among individuals, the model's generalization capacity in real-time recognition is constrained. Direct application of model training tailored to specific individuals to others may lead to a suboptimal recognition performance.

Author Contributions: Conceptualization, X.L. and W.H.; methodology, X.L.; software, X.L.; validation, X.L.; formal analysis, X.L.; investigation, X.L.; resources, X.L. and W.H.; data curation, X.L. and W.H.; writing—original draft preparation, X.L. and W.H.; writing—review and editing, X.L., W.H. and Y.L.; visualization, X.L.; supervision, W.H. and Z.W.; project administration, X.L., W.H., and X.D.; funding acquisition, W.H. and Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Hunan Natural Science Foundation (No. 2022JJ31015), General Project of the National Social Science Fund (No. 22BGL173), Major Project of the Social Science Evaluation Committee of Hunan Province (No. XSP22ZDA006), Hunan Teaching Reform Research Project (No. HNJG-20230470), and the Graduate Science and Technology Innovation Fund of Central South University of Forestry and Technology (No. 2023CX02070).

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Sun, Y.; Xu, C.; Li, G.; Xu, W.; Kong, J.; Jiang, D.; Tao, B.; Chen, D. Intelligent Human Computer Interaction Based on Non Redundant EMG Signal. *Alex. Eng. J.* **2020**, *59*, 1149–1157. [CrossRef]
2. Li, K.; Zhang, J.; Wang, L.; Zhang, M.; Li, J.; Bao, S. A Review of the Key Technologies for sEMG-Based Human-Robot Interaction Systems. *Biomed. Signal Process. Control* **2020**, *62*, 102074. [CrossRef]
3. Palermo, F.; Cognolato, M.; Gijsberts, A.; Muller, H.; Caputo, B.; Atzori, M. Repeatability of Grasp Recognition for Robotic Hand Prosthesis Control Based on sEMG Data. In Proceedings of the 2017 International Conference on Rehabilitation Robotics (ICORR), London, UK, 17–20 July 2017; pp. 1154–1159.
4. Zhang, Y.; Duan, X.-G.; Zhong, G.; Deng, H. Initial Slip Detection and Its Application in Biomimetic Robotic Hands. *IEEE Sens. J.* **2016**, *16*, 7073–7080. [CrossRef]
5. Xu, X.; Deng, H.; Zhang, Y.; Chen, J. Continuous Grasping Force Estimation with Surface EMG Based on Huxley-Type Musculoskeletal Model. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2023**, *31*, 346–355. [CrossRef] [PubMed]
6. Arozi, M.; Caesarendra, W.; Ariyanto, M.; Munadi, M.; Setiawan, J.D.; Glowacz, A. Pattern Recognition of Single-Channel sEMG Signal Using PCA and ANN Method to Classify Nine Hand Movements. *Symmetry* **2020**, *12*, 541. [CrossRef]
7. Mendes Junior, J.J.A.; Freitas, M.L.B.; Siqueira, H.V.; Lazzaretti, A.E.; Pichorim, S.F.; Stevan, S.L. Feature Selection and Dimensionality Reduction: An Extensive Comparison in Hand Gesture Classification by sEMG in Eight Channels Armband Approach. *Biomed. Signal Process. Control* **2020**, *59*, 101920. [CrossRef]
8. Oskoei, M.A.; Hu, H. Support Vector Machine-Based Classification Scheme for Myoelectric Control Applied to Upper Limb. *IEEE Trans. Biomed. Eng.* **2008**, *55*, 1956–1965. [CrossRef]
9. Pizzolato, S.; Tagliapietra, L.; Cognolato, M.; Reggiani, M.; Müller, H.; Atzori, M. Comparison of Six Electromyography Acquisition Setups on Hand Movement Classification Tasks. *PLoS ONE* **2017**, *12*, e0186132. [CrossRef]
10. Wang, B.; Wang, C.; Wang, L.; Xie, N.; Wei, W. Recognition of semg hand actions based on cloud adaptive quantum chaos ions motion algorithm optimized svm. *J. Mech. Med. Biol.* **2019**, *19*, 1950047. [CrossRef]
11. Kuzborskij, I.; Gijsberts, A.; Caputo, B. On the Challenge of Classifying 52 Hand Movements from Surface Electromyography. In Proceedings of the 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, San Diego, CA, USA, 28 August–1 September 2012; pp. 4931–4937.
12. Atzori, M.; Gijsberts, A.; Castellini, C.; Caputo, B.; Hager, A.-G.M.; Elsig, S.; Giatsidis, G.; Bassetto, F.; Müller, H. Electromyography Data for Non-Invasive Naturally-Controlled Robotic Hand Prostheses. *Sci. Data* **2014**, *1*, 140053. [CrossRef]
13. Xue, Y.; Ji, X.; Zhou, D.; Li, J.; Ju, Z. SEMG-Based Human In-Hand Motion Recognition Using Nonlinear Time Series Analysis and Random Forest. *IEEE Access* **2019**, *7*, 176448–176457. [CrossRef]
14. Tkach, D.; Huang, H.; Kuiken, T.A. RSetseuardchy of Stability of Time-Domain Features for Electromyographic Pattern Recognition. *J. Neuroeng. Rehabil.* **2010**, *7*, 21. [CrossRef]
15. Krasoulis, A.; Kyranou, I.; Erden, M.S.; Nazarpour, K.; Vijayakumar, S. Improved Prosthetic Hand Control with Concurrent Use of Myoelectric and Inertial Measurements. *J. Neuroeng. Rehabil.* **2017**, *14*, 71. [CrossRef] [PubMed]
16. Namazi, H. Decoding of hand gestures by fractal analysis of electromyography (emg) signal. *Fractals* **2019**, *27*, 1950022. [CrossRef]
17. Park, K.-H.; Lee, S.-W. Movement Intention Decoding Based on Deep Learning for Multiuser Myoelectric Interfaces. In Proceedings of the 2016 4th International Winter Conference on Brain-Computer Interface (BCI), Gangwon, Republic of Korea, 20–22 February 2016; pp. 1–2.
18. Atzori, M.; Cognolato, M.; Müller, H. Deep Learning with Convolutional Neural Networks Applied to Electromyography Data: A Resource for the Classification of Movements for Prosthetic Hands. *Front. Neurorobot.* **2016**, *10*, 9. [CrossRef] [PubMed]
19. Geng, W.; Du, Y.; Jin, W.; Wei, W.; Hu, Y.; Li, J. Gesture Recognition by Instantaneous Surface EMG Images. *Sci. Rep.* **2016**, *6*, 36571. [CrossRef] [PubMed]
20. Soroushmojdehi, R.; Javadzadeh, S.; Pedrocchi, A.; Gandolla, M. Transfer Learning in Hand Movement Intention Detection Based on Surface Electromyography Signals. *Front. Neurosci.* **2022**, *16*, 977328. [CrossRef]
21. Zhai, X.; Jelfs, B.; Chan, R.H.M.; Tin, C. Self-Recalibrating Surface EMG Pattern Recognition for Neuroprosthesis Control Based on Convolutional Neural Network. *Front. Neurosci.* **2017**, *11*, 379. [CrossRef] [PubMed]
22. Cheng, Y.; Li, G.; Yu, M.; Jiang, D.; Yun, J.; Liu, Y.; Liu, Y.; Chen, D. Gesture Recognition Based on Surface Electromyography. *Concurr. Comput.* **2021**, *33*, e6051. [CrossRef]
23. Wei, W.; Wong, Y.; Du, Y.; Hu, Y.; Kankanhalli, M.; Geng, W. myocnn-A Multi-Stream Convolutional Neural Network for sEMG-Based Gesture Recognition in Muscle-Computer Interface. *Pattern Recognit. Lett.* **2019**, *119*, 131–138. [CrossRef]
24. Hao, S.; Wang, R.; Wang, Y.; Li, Y. A Spatial Attention Based Convolutional Neural Network for Gesture Recognition with HD-sEMG Signals. In Proceedings of the 2020 IEEE International Conference on E-health Networking, Application & Services (HEALTHCOM), Shenzhen, China, 1 March 2021; pp. 1–6.

25. Wang, L.; Fu, J.; Zheng, B.; Zhao, H. Research on sEMG–Based Gesture Recognition Using the Attention-Based LSTM-CNN with Stationary Wavelet Packet Transform. In Proceedings of the 2022 4th International Conference on Advances in Computer Technology, Information Science and Communications (CTISC), Suzhou, China, 22 April 2022; pp. 1–6.
26. Fan, X.; Zou, L.; Liu, Z.; He, Y.; Zou, L.; Chi, R. CSAC-Net: Fast Adaptive sEMG Recognition through Attention Convolution Network and Model-Agnostic Meta-Learning. *Sensors* **2022**, *22*, 3661. [CrossRef]
27. Rahimian, E.; Zabihi, S.; Asif, A.; Farina, D.; Atashzar, S.F.; Mohammadi, A. Hand Gesture Recognition Using Temporal Convolutions and Attention Mechanism. In Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022.
28. Hu, Y.; Wong, Y.; Wei, W.; Du, Y.; Kankanhalli, M.; Geng, W. A Novel Attention-Based Hybrid CNN-RNN Architecture for sEMG-Based Gesture Recognition. *PLoS ONE* **2018**, *13*, e0206049. [CrossRef]
29. Han, L.; Zou, Y.; Cheng, L. A Convolutional Neural Network with Multi-Scale Kernel and Feature Fusion for sEMG-Based Gesture Recognition. In Proceedings of the 2021 IEEE International Conference on Robotics and Biomimetics (ROBIO), Sanya, China, 27 December 2021; pp. 774–779.
30. Shen, S.; Gu, K.; Chen, X.-R.; Lv, C.-X.; Wang, R.-C. Gesture Recognition Through sEMG with Wearable Device Based on Deep Learning. *Mob. Netw. Appl.* **2020**, *25*, 2447–2458. [CrossRef]
31. Jiang, B.; Wu, H.; Xia, Q.; Xiao, H.; Peng, B.; Wang, L.; Zhao, Y. Gesture Recognition Using sEMG Based on Multi-Scale Fusion Convolution and Channel Attention. *SSRN* **2023**, preprint. [CrossRef]
32. Eichhorn, M.; Pollnau, M. Spectroscopic Foundations of Lasers: Spontaneous Emission Into a Resonator Mode. *IEEE J. Sel. Top. Quantum Electron.* **2015**, *21*, 486–501. [CrossRef]
33. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
34. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014.
35. Pan, X.; Ge, C.; Lu, R.; Song, S.; Chen, G.; Huang, Z.; Huang, G. On the Integration of Self-Attention and Convolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
36. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.