# Gesture Recognition from Sensors Using Dual-path Encoding and Attention

No Author Given

No Institute Given

**Abstract.** With the widespread adoption of smart wearable devices, embedded Micro-Electro-Mechanical Systems (MEMS) have become a key research area in gesture recognition. Compared to vision-based approaches, MEMS-based recognition offers advantages such as a simple structure, long-term monitoring, and easy integration, showing great potential for real-world applications. Traditional MEMS gesture recognition methods rely on handcrafted features, while multimodal sensors like accelerometers and gyroscopes generate heterogeneous data, requiring advanced fusion strategies. To address this, we propose DS-CAN, a data fusion framework integrating contrastive and deep learning. The model employs dual independent encoding channels to process accelerometer and gyroscope data, followed by a multi-head attention mechanism to focus on features like time alignment, amplitude correlation, and motion direction consistency. The contrastive loss function, enhanced with a temperature parameter $\tau$, extends traditional unimodal contrastive learning to multimodal scenarios, promoting feature association and distinction. Experimental results show that DS-CAN achieves around 94% accuracy on the 6DMG (20 gestures) and MGD (12 gestures) datasets, outperforming models like CNNs, LSTMs, and two-stream CNNs. Cross validation and additional experiments on posture datasets validate the model's robustness and generalization ability. This method offers an efficient solution for gesture recognition in wearable devices and demonstrates practical value in human-computer interaction.

**Keywords:** Gesture Recognition · Deep Learning · Contrastive Learning · Data Fusion · Multi-Head Attention.

## 1 Introduction

Vision-based gesture recognition systems have made significant progress in fields such as human–computer interaction and intelligent surveillance due to their intuitive ability to process image information. These methods capture gesture images through cameras and extract spatial features using models such as Convolutional Neural Networks (CNNs), enabling effective classification of static gestures and dynamic movements. However, their practical application is limited by sensitivity to occlusion, variations in lighting, and privacy concerns [1]. The rapid development of wearable technology has spurred research on gesture recognition based on Micro-Electro-Mechanical Systems (MEMS) sensors,

demonstrating broad application prospects in areas such as smart healthcare and human–computer interaction [2]. Compared with traditional vision-based approaches, MEMS sensors offer advantages such as simple structure, strong environmental robustness, and easy integration into compact devices. Modern accelerometers use advanced capacitive sensing principles to detect multi-axis inertial forces, achieving ultra-low power consumption while maintaining sub-milli-g resolution [3]. These sensors can faithfully capture both static gravitational components and dynamic motion features, providing rich representations for gesture modeling [4]. When combined with MEMS gyroscopes that measure angular velocity [5], this multimodal sensor fusion enables the construction of a comprehensive motion profile necessary to distinguish subtle gesture variations.

Despite the significant advantages of MEMS-based approaches, current sensor-based gesture recognition systems still face several key challenges:

1. **Inefficient multimodal data fusion:** Accelerometers and gyroscopes represent linear and rotational motion, respectively. Traditional methods adopt early or late fusion strategies but fail to dynamically capture the critical contributions of each modality across different gestures.
2. **Insufficient capability for feature clustering and separation:** When user-specific variations in gesture execution are large, conventional contrastive loss functions struggle to ensure that multimodal features of the same gesture are effectively clustered while features of different gestures are sufficiently separated. This impairs the model's ability to distinguish between various gesture classes.
3. **Challenges in lightweight real-time deployment:** Due to their complex architectures, existing models often involve a large number of parameters and high computational cost, making them unsuitable for resource-constrained edge devices such as smart wristbands and AR glasses.

Traditionally, approaches based on Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks [6] have struggled to simultaneously address the above challenges, primarily due to their limited capacity for modeling cross-modal interactions and temporal attention mechanisms. To overcome these limitations, this study proposes a novel contrastive accelerometer–gyroscope embedding framework that integrates deep learning and contrastive learning strategies. The proposed architecture innovatively incorporates three key components:

1. **Attention-guided multimodal sensor fusion:** A multi-head attention mechanism is introduced to dynamically calibrate the weights of cross-modal features, enabling adaptive modeling of each sensor's contribution within different gestures.
2. **Robust contrastive learning framework:** Cross-modal positive and negative sample pairs are constructed, and the NT-Xent loss is employed to enforce alignment of accelerometer and gyroscope features in the latent space for the same gesture, while enlarging the distance between features of different gestures. This effectively mitigates disturbances such as sensor orienta-

tion shifts and individual execution variances, enhancing feature invariance and discriminability.

3. **Lightweight architecture and compression optimization:** A hierarchical feature extraction strategy is adopted to reduce parameter redundancy. Combined with contrastive learning's inherent feature compression and structural regularization effects, the model achieves a parameter size of only 3.9M and an inference latency of 10.3 ms without requiring additional compression algorithms. This satisfies the real-time computation requirements of edge devices such as smartwatches and wearable bands, overcoming the deployment bottlenecks of traditional deep learning models in resource-constrained environments.

Comparative analysis with traditional CNN-LSTM hybrid models [7] and other baseline architectures demonstrates the proposed model's superiority in both recognition accuracy and computational efficiency. The experimental results lay a solid foundation for advancing the next generation of human-centric computing systems in healthcare, robotics, and intelligent environments [8, 9].

## 2 Related Works

### 2.1 MEMS sensor technology

Gesture recognition systems mainly consist of three core stages: MEMS sensors capturing motion signals, data transmission, and classifier-based data categorization. Early studies primarily used triaxial accelerometers; for instance, He et al. [3] employed a threshold-based algorithm to achieve basic action recognition. With the widespread adoption of smartphones, smartwatches, and other devices, the application scenarios for data acquisition have expanded. Kwon et al. [1] leveraged these data to promote the use of Convolutional Neural Networks (CNN) in wrist-worn gesture recognition. To further improve the performance of machine learning models based on MEMS accelerometer sensors, lightweight CNNs emerged, ensuring real-time capability while achieving high-accuracy gesture recognition [6]. In addition, MEMS gyroscope sensors focus on capturing angular velocity (rad/s), enabling sensitive detection of the rotational motion of devices or frames [10]. As inertial sensors, MEMS gyroscopes have driven innovative developments in the analysis of human rotational movements, and they are widely applied in various fields such as consumer electronics, automotive electronics, industrial automation, and aerospace.

### 2.2 Deep learning

Deep learning breakthroughs have significantly enhanced the performance of computer vision tasks, driving their large-scale application in everyday scenarios [11]. Currently, classification tasks such as gesture recognition and activity recognition have become research hotspots. MEMS sensors provide high-precision signal acquisition and convenient data transmission capabilities, offering high-quality data sources for feature learning. For example, Xu et al. [12]

proposed deformable CNNs to adapt to sensor displacement; Sun et al. [13] designed lattice LSTMs to model complex temporal dependencies of IMUs; Koo et al. [10] developed a contrastive learning framework to achieve cross-modal feature embedding in low-label scenarios. Deep neural networks represented by CNNs and LSTMs have achieved high-accuracy classification in gesture recognition by mining spatiotemporal features of sensor data [6, 13, 14]. However, existing studies are mostly limited to isolated models or small datasets and lack systematic comparisons with traditional machine learning methods [15–17], highlighting the urgent need for more generalizable modeling frameworks.

The multi-head attention mechanism, as an innovative extension of self-attention, has become a key to overcoming the representation bottleneck of traditional models [18]. This mechanism captures multi-scale features in sequences in parallel, including both short-range and long-range dependencies, and dynamically allocates weights across different time steps, enabling the model to simultaneously focus on multiple regions of the input sequence. For example, in mixed accelerometer and gyroscope data, it can adaptively enhance key modal signals and suppress redundant information based on gesture types [19]. This "dynamic weighting — multi-domain focusing" characteristic endows the model with stronger feature disentanglement and generalization capabilities, providing a new approach for complex gesture classification.

### 2.3 Self-supervised Learning

Self-supervised learning (SSL) achieves feature learning by exploiting the intrinsic structure of unlabeled data, becoming a key technology to overcome the annotation bottleneck in the field of gesture recognition. It designs pretext tasks to mine the spatiotemporal continuity and multimodal correlations of sensor data, thereby generating pseudo-supervision signals to learn meaningful motion patterns [20]. Additionally, SSL enhances noise robustness through multi-sensor invariance learning and has demonstrated performance close to that of supervised learning in fine-grained gesture classification tasks [21], [22]. As a core branch of SSL, contrastive learning focuses on learning discriminative representations through instance discrimination mechanisms. It generates multi-view positive sample pairs via temporal augmentation methods and employs cross-subject negative sampling strategies to increase the distance between heterogeneous features [23], [24]. For example, aligning accelerometer and gyroscope signals as positive samples enhances modality synergy, while using cross-subject negative samples suppresses interference caused by user variability [25]. This "augmentation—alignment—debiasing" pipeline provides an efficient paradigm for unsupervised feature learning.

Existing MEMS sensors, deep learning, and data fusion techniques have improved recognition performance; however, deficiencies remain in dynamic interactions of multimodal data, guided feature clustering and separation, and lightweight real-time deployment. This study proposes a model that incorporates a multi-head attention mechanism and contrastive learning to build a two-stage

framework of "independent encoding — attention fusion — contrastive enhancement." The framework aims to address the bottlenecks of traditional methods in cross-modal feature interaction, feature discriminability, and lightweight real-time deployment, thereby providing a more robust solution for gesture recognition.

## 3 Methods

This chapter focuses on the DS-CAN framework, which employs two independent convolutional encoders to separately process accelerometer and gyroscope data, leveraging a multi-head attention mechanism to achieve dynamic fusion of cross-modal features. The contrastive learning strategy is optimized to enhance feature discriminability while simultaneously performing supervised classification and self-supervised learning (SSL) tasks. The overall framework of the model is illustrated in Fig. 2. This approach integrates hierarchical temporal feature learning modules, multi-head attention fusion mechanisms, and contrastive representation alignment frameworks in sensor-based gesture recognition, constituting a hierarchical technical breakthrough.

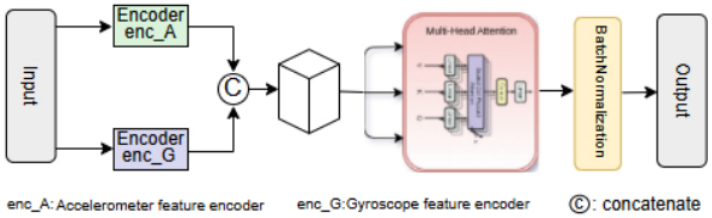



Fig. 1: Architecture diagram with an additional encoder for gyro input.

### 3.1 Two-stream architecture

The framework processes accelerometer and gyroscope inputs through parallel temporal convolutional networks to capture their complementary motion features. Each sensor data stream passes through one-dimensional convolutional layers, performing temporal abstraction in three stages, with kernel sizes exponentially increasing across layers. The first convolutional layer extracts local motion patterns within a 30 ms time window, capturing fundamental motion features such as acceleration peaks or angular velocity changes. Subsequent convolutional layers, combined with max-pooling operations of stride 2, double the temporal context while halving the feature map resolution, enabling the network to hierarchically aggregate motion primitives into complete gesture representations.

For sensor modality $m$, the process can be mathematically expressed as:

$$\mathbf{F}_t^m = ELU(BN(\mathbf{W}_2^m * MaxPool(ELU(BN(\mathbf{W}_1^m * \mathbf{x}_t^m))))) \quad (1)$$

Where $\mathbf{W}_k^m$ denotes the learnable filter at the $k$-th layer, BN stands for Batch Normalization, and ELU represents the Exponential Linear Unit activation function. This design addresses two key challenges in processing sensor data: 1) Handling variable-length gestures through adaptive receptive field expansion; 2) Achieving device orientation invariance via axis-independent feature learning. The final global average pooling operation produces a fixed-length 64-dimensional feature vector while preserving temporal attention weights, enabling subsequent modules to focus on discriminative gesture phases.

The dual-stream architecture particularly preserves physical characteristics that may be obscured in early fusion methods. For the accelerometer data, the network learns representations sensitive to the gravity component and linear acceleration patterns governed by Newtonian mechanics:

$$\mathbf{a}_{observed} = \mathbf{a}_{true} - g\hat{\mathbf{u}} \tag{2}$$

Here, $\mathbf{g}$ denotes the gravitational acceleration, and $\mathbf{u}$ represents the device orientation vector. In contrast, the gyroscope stream encoding captures angular velocity dynamics described by the Euler rotation equations:

$$\frac{d\omega}{dt} = \mathbf{I}^{-1}(\tau - \omega \times \mathbf{I}\omega) \tag{3}$$

Here, $\mathbf{I}$ denotes the inertia tensor, and $\tau$ denotes the applied torque. The dual independent processing pathways allow for specialized feature extraction dedicated to these distinct physical phenomena prior to cross-modal feature interaction.

### 3.2 Attention-driven multimodal channels

In the field of gesture recognition, data collected from different sensors contribute significantly differently to action recognition. To effectively address this challenge, the multi-head attention mechanism has become a key strategy for optimizing model performance. This mechanism dynamically weights sensor data across both the temporal and channel dimensions based on their relevance to the target gestures. Along the temporal dimension, it precisely captures critical time points of gesture actions; along the sensor channel dimension, it adaptively assigns reasonable weights to different sensor data, thereby fully exploiting the value of each sensor. Through this approach, the model can deeply analyze the complex spatiotemporal dependencies within the sensor data, ultimately achieving improvements in both gesture recognition accuracy and robustness [26].

The fusion module employs a Transformer-based attention mechanism to dynamically integrate features from the accelerometer and gyroscope, addressing three key limitations of traditional fusion methods: 1) fixed fusion weights lacking sensitivity to temporal context; 2) neglect of cross-modal correlations; and 3) inability to handle asynchronous sensor responses. The processing procedure of this attention mechanism consists of three stages.

**Query-Key-Value projection** Features of each modality undergo learned linear transformations to generate Query (Q), Key (K), and Value (V) vectors: representing the accelerometer.

$$\mathbf{Q}^m = \mathbf{W}_q^m \mathbf{f}^m, \quad \mathbf{K}^m = \mathbf{W}_k^m \mathbf{f}^m, \quad \mathbf{V}^m = \mathbf{W}_v^m \mathbf{f}^m \tag{4}$$

Here, $m \in \{A, G\}$ denotes the accelerometer and gyroscope modalities. The projection matrices $\mathbf{W}_q^m$, $\mathbf{W}_k^m$, and $\mathbf{W}_v^m$ are independently learned for each modality to preserve modality-specific information [27].

**Query-Key-Value projection** The attention weights between modalities are computed using scaled dot-product similarity:

$$\alpha_{ij} = \frac{\exp(\mathbf{Q}_i^a \cdot \mathbf{K}_j^g / \sqrt{d})}{\sum_{k=1}^{t} \exp(\mathbf{Q}_i^a \cdot \mathbf{K}_k^g / \sqrt{d})} \tag{5}$$

Here, $d$ denotes the dimension of the key vectors. These weights determine the contribution of each gyroscope feature $\mathbf{V}_j^G$ to the accelerometer time position $i$. The multi-head implementation allows the model to simultaneously attend to different aspects of cross-modal relationships, such as temporal alignment, magnitude correlation, and motion direction consistency.

**Adaptive feature fusion** The final fused representation combines the attention values from both modalities:

$$\mathbf{f}_{fused} = LayerNorm(\mathbf{W}_o[\mathbf{H}^a \| \mathbf{H}^g] + \mathbf{b}_o) \tag{6}$$

where $\|$ denotes the concatenation operation, and $H^m$ represents the aggregated output of the heads for modality $m$. This structure dynamically reweights the contribution of each sensor based on the instantaneous relevance of the sensor data to the gesture context. For example, during rapid rotational gestures, the attention mechanism automatically increases the weight of gyroscope features, while during transitional phases, it emphasizes the accelerometer inputs.

Given two types of sensor data $A_i$ and $B_i$, where $A_i$ represents accelerometer data and $B_i$ represents gyroscope data, the combined gesture action $O_i$ can be expressed as:

$$O_i = (\lambda A_i + \gamma B_i) \tag{7}$$

where $\lambda$ and $\gamma$ are the weights assigned to the accelerometer and gyroscope data, respectively.

In our model, a single sensor data $A_i$ or $B_i$ is regarded as the key and value, while the gesture action $O_i$ is regarded as the query. The gesture action is composed of a weighted fusion of the two sets of sensor data. First, the similarity between the gesture action $O_i$ and the single sensor data $A_i$ is calculated using the dot product:

$$a(O_i, A_i) = \frac{O_i \cdot A_i^T}{\sqrt{d_a}} \tag{8}$$

where $d_a$ is the dimension of the sensor data. Then, the similarity is passed through a softmax function to obtain the attention weights:

$$\alpha(O_i, A_i) = softmax(a(O_i, A_i)) \tag{9}$$

The final weighted data is output by the attention aggregation function:

$$f(O_i, (A_i, A_i)) = \sum_{i=1}^{n} \alpha(O_i, A_i) A_i \tag{10}$$

The effectiveness of the attention mechanism stems from its ability to model pairwise interactions across all temporal positions in both modalities. This is particularly important in gesture recognition, where time shifts in sensor responses often occur due to biomechanical constraints of the human body.

### 3.3 Contrastive representation learning framework

The contrastive learning module enhances feature discriminability by learning noise-invariant representations. Its basic assumption is that, for the same action, observations from different sensors should maintain consistency in gesture semantics [23]. This assumption extends traditional unimodal contrastive learning to multimodal interactive scenarios, aiming to effectively handle data from different sensors. The process is formalized using a temperature-scaled normalized cross-entropy (NT-Xent) loss function:

$$\mathcal{L}_{ssl} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\phi(\mathbf{z}_i^a, \mathbf{z}_i^g)/\tau)}{\sum_{j=1}^{N} [\exp(\phi(\mathbf{z}_i^a, \mathbf{z}_j^g)/\tau) + \exp(\phi(\mathbf{z}_i^a, \mathbf{z}_j^a)/\tau)]} \tag{11}$$

where $\phi(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|}$ denotes the cosine similarity, and $\tau$ is a learnable temperature parameter. The denominator includes both cross-modal and intra-modal negative samples to prevent trivial solutions and enhance discriminative capability.

The projection network $g(\cdot)$, which maps encoder features to the contrastive space, employs a nonlinear transformation with a bottleneck architecture.

$$\mathbf{z}^m = g(\mathbf{f}^m) = \mathbf{W}_2 ReLU(\mathbf{W}_1 \mathbf{f}^m) \tag{12}$$

This design forces the network to learn a compressed representation that retains only the most discriminative features for gesture recognition. The temperature parameter $\tau$ is automatically adjusted during training to optimize the hardness of negative samples.

$$\frac{\partial \mathcal{L}}{\partial \tau} = \frac{1}{\tau^2} \sum_{i,j} p_{ij} (\phi_{ij} - E[\phi_{ij}]) \tag{13}$$

Here, $p_{ij}$ denotes the softmax probability. This adaptive scaling prevents the model from collapsing all features into a single point (as $\tau \to 0$) or ignoring similarity differences (as $\tau \to \infty$). The contrastive learning objective induces a latent space geometry in which the Euclidean distances between embedding vectors correspond to semantic gesture similarity.

### 3.4 Structural regularization

**Dropout** Dropout reduces overfitting by randomly deactivating neurons during training [28]. It diminishes the network's reliance on specific neurons, forcing the model to learn more robust feature representations. Unlike L1/L2 regularization, Dropout disrupts co-adaptations among neurons through random masking [29]. During training, each neuron is retained with a probability of $1 - p$, and during inference, activations are scaled accordingly to alleviate overfitting and improve generalization.

**Batch Normalization** BatchNorm reduces internal covariate shift by normalizing the mean and variance of layer inputs, stabilizing training, and accelerating convergence [30]. It enables the use of higher learning rates, reduces sensitivity to weight initialization, and mitigates vanishing gradient problems. In our experiments, inserting BatchNorm layers significantly improved training stability.

## 4 Results and Analysis

### 4.1 Dataset

The experiments use the 6DMG, MGD, UCI-HAR, and PAMAP2 datasets. The 6DMG dataset contains 20 complex gestures collected from 28 participants, totaling 5600 samples; the MGD dataset includes 12 semantic gestures from 32 participants, with 5547 samples. The UCI-HAR and PAMAP2 datasets are employed to validate the generalizability for human activity and posture recognition.

### 4.2 Experimental Results

This study employs two data partitioning strategies: fixed-ratio stratified sampling and cross-validation. The 6DMG and MGD datasets are divided into training, validation, and test sets at a 70:15:15 ratio, while the UCI-HAR and PAMAP2 datasets are split into training and test sets at a 70:30 ratio. Stratified sampling is used for the former to ensure consistent class distribution across subsets, whereas the latter are randomly partitioned based on participant independence. Only 6DMG adopts 5-fold cross-validation to evaluate model robustness, as other datasets do not use this method due to limited sample sizes or different experimental focuses. Recognition performance is measured by accuracy and macro F1-score, while model efficiency and real-time capability are assessed via the number of parameters and inference time.

**Performance Comparison of Gesture Recognition Tasks** As shown in Table 1, the model outperforms traditional models on the 6DMG and MGD datasets: it achieves an accuracy of 94.29% on 6DMG, representing a 1.6% improvement over LSTM-CNN, with a 1.5% increase in the F1 score; on MGD, the

accuracy and F1 score are improved by 0.8% and 1.3%, respectively. The model has a parameter size of 3.9M (smaller than LSTM-CNN's 5.1M) and an inference time of only 10.3ms, meeting the requirements for real-time applications.

Table 1: Performance comparison on 6DMG dataset using 5-fold cross-validation.

| Model | Validation | | Test | | Param |
|---|---|---|---|---|---|
| | Acc (%) | F1 | Acc (%) | F1 | (M) |
| Baseline CNN | 92.19 | 0.922 | 91.69 | 0.917 | 2.1 |
| LSTM | 91.19 | 0.912 | 91.94 | 0.920 | 3.8 |
| DeepConvLSTM | 89.73 | 0.897 | 89.24 | 0.892 | 4.2 |
| LSTMconvNet | 93.49 | 0.935 | 92.74 | 0.928 | 5.1 |
| Two-Stream CNN | 92.34 | 0.922 | 91.79 | 0.918 | 3.4 |
| DS-CAN | 94.29 | 0.942 | 93.14 | 0.931 | 3.9 |

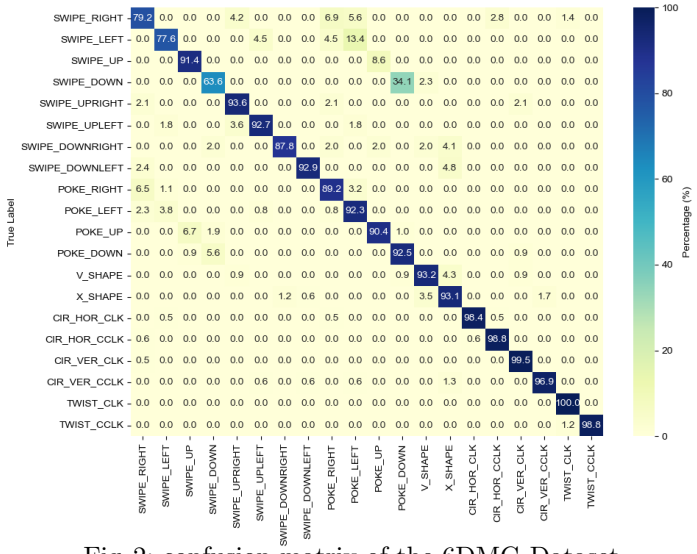
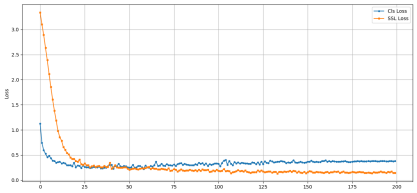


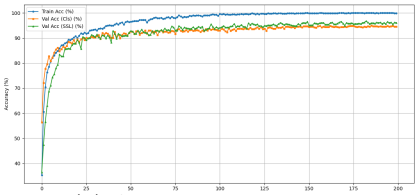

Fig. 2: confusion matrix of the 6DMG Dataset.

**Typical Error Analysis of the Confusion Matrix** The specific performance of gesture classification in the confusion matrix is shown in Fig. 4. Using the 6DMG dataset as an example, the two-stream model shows notable misclassifications for gestures such as SwipeRight and CirHorClk, with accuracies of 81.6% and 55.6%, respectively. After optimizing contrastive learning and incorporating an attention mechanism, misclassifications were significantly reduced, and the accuracies improved to 94.29% and 92.4%, respectively. Remaining errors are mainly concentrated in gestures with highly similar spatial trajectories and temporal dynamics, indicating that the model still has room for improvement in distinguishing subtle motion differences.

**Generalization Ability Evaluation** To evaluate the generalization capability of our model, we conduct experiments on the UCI-HAR and PAMAP2 datasets.

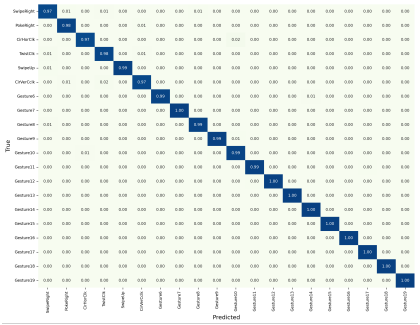


(a) Loss over Epochs

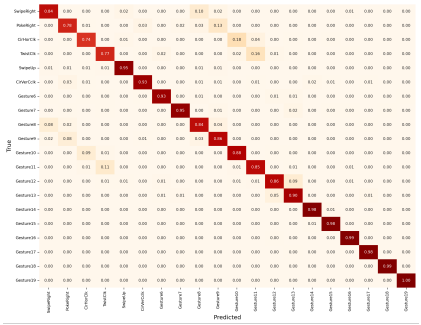


(b) Accuracy over Epochs

Fig. 3: Loss Function and Accuracy on the 6DMG Dataset



(a) DS-CAN Confusion Matrix (Accuracy: 94.29%).



(b) Two-Stream CNN Confusion Matrix (Accuracy: 92.34%).

Fig. 4: Confusion matrices comparison

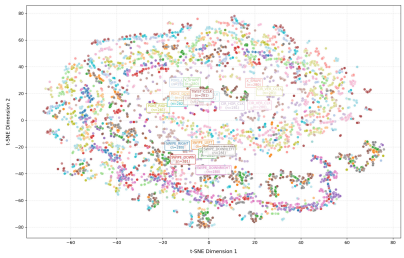
The results are presented in Table 2. On the UCI-HAR dataset, our model achieves a weighted F1 score of 0.9276, representing a 0.25% improvement over the Two-Stream CNN model, which indicates a stronger ability to handle class imbalance. On the more challenging PAMAP2 dataset, the accuracy improves by 3.25% and the weighted F1 score increases by 3.86%, demonstrating the model's robust general feature extraction capability across multimodal sensor data.

Table 2: Verification of Generalization Ability.

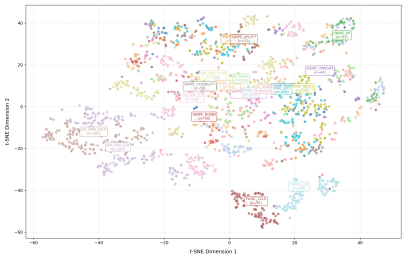
| Model | UCI - HAR | | PAMAP2 | |
|---|---|---|---|---|
| | ACC (%) | F1 | ACC (%) | F1 |
| Baseline CNN | 92.34 | 0.9223 | 64.25 | 0.6230 |
| Two-Stream CNN | 92.50 | 0.9249 | 70.47 | 0.6947 |
| DS-CAN | 92.77 | 0.9276 | 73.62 | 0.7333 |

**Validation of Feature Learning Effectiveness** We visualize the distribution of original features and self-supervised features using t-SNE in Fig. 5. The results show that the original signal features exhibit significant class overlap in the two-dimensional space, while the joint features learned by our approach demonstrate clear inter-class separation and tight intra-class clustering. Quantitative analysis further confirms this observation: an MLP classifier trained on the self-supervised features achieves an accuracy of 92.43%, substantially outperforming

the 82.26% accuracy obtained using original features. These results demonstrate that contrastive learning and attention mechanisms effectively enhance feature discriminability.



(a) The joint features learned from the flattened raw signals.



(b) The joint features learned through self-supervised learning.

Fig. 5: t-SNE plot of the learned joint features.

**Model Efficiency Analysis** Model efficiency comparisons are shown in Table 3. Our model achieves high accuracy with only 3.9M parameters, representing a 23.5% reduction compared to the LSTM-CNN baseline. Moreover, the inference time is reduced by 32.2% to 10.3 ms. This demonstrates that our approach achieves a favorable trade-off between performance and computational cost, making it well-suited for deployment on resource-constrained edge devices.

Table 3: Comparison of Model Parameter Sizes and Inference Times.

| **Model** | **Model Size (M)** | **Inference Time (ms)** |
|---|---|---|
| Baseline CNN | 2.1 | 8.5 |
| LSTM - CNN | 5.1 | 15.2 |
| DS-CAN | 3.9 | 10.3 |

## 5 Conclusion

This paper proposes an innovative gesture recognition framework that uses a dual-channel independent encoding structure to process accelerometer and gyroscope data. It then employs a multi-head attention mechanism to focus on multi-dimensional features such as time alignment, amplitude correlation, and motion direction consistency. By introducing a temperature parameter $\tau$, this work extends traditional unimodal contrastive learning to multimodal data fusion scenarios, effectively promoting feature association and distinction. This approach not only improves the efficiency of multimodal feature fusion but also addresses the lack of guidance ability in multimodal data processing. Furthermore, the model significantly enhances recognition accuracy while maintaining a lightweight design, making it feasible for deployment on edge devices. Overall, this work makes significant innovations in multimodal feature fusion, guidance capability enhancement, and practical deployment, providing a viable solution for future wearable device-based gesture recognition systems.

## References


1. Wang, X., Zhang, J., Yang, M.: SonicGest: Ultrasonic Gesture Recognition System Combined with GAN on Smartphones. Journal **XX**(1), 6813911 (2023)
2. Jiang, C., Xu, W., Li, Y., et al.: Capturing forceful interaction with deformable objects using a deep learning-powered stretchable tactile array. Nature Communications **15**(1), 9513 (2024)
3. Noble, F., Xu, M., Alam, F.: Static hand gesture recognition using capacitive sensing and machine learning. Sensors **23**(7), 3419 (2023)
4. Kim, B., Seo, S.: EfficientNetV2-based dynamic gesture recognition using transformed scalogram from triaxial acceleration signal. Journal of Computational Design and Engineering **10**(4), 1694–1706 (2023)
5. Caro-Alvaro, S., Garcia-Lopez, E., Brun-Guajardo, A., et al.: Gesture-based interactions: Integrating accelerometer and gyroscope sensors in the use of mobile apps. Sensors **24**(3), 1004 (2024)
6. Huang, W., Zhang, L., Gao, W., Min, F., He, J.: Shallow convolutional neural networks for human activity recognition using wearable sensors. IEEE Transactions on Instrumentation and Measurement **70**(X), 1–11 (2021)
7. Wu, J., Ren, P., Song, B., et al.: Data glove-based gesture recognition using CNN-BiLSTM model with attention mechanism. PLOS ONE **18**(11), e0294174 (2023)
8. Ko, W.-R., Jang, M., Lee, J., Kim, J.: Air-act2act: Human–human interaction dataset for teaching non-verbal social behaviors to robots. Int. J. Robot. Res. **40**(4–5), 691–697 (2021)
9. Liu, P., Glas, D. F., Kanda, T., Ishiguro, H.: Data-driven HRI: Learning social behaviors by example from human–human interaction. IEEE Trans. Robot. **32**(4), 988–1008 (2016)
10. Koo, I., Park, Y., Jeong, M., Kim, C.: Contrastive accelerometer–gyroscope embedding model for human activity recognition. IEEE Sensors J. **23**(1), 506–513 (2023)
11. Nogales, R. E., Benalcázar, M. E.: Hand gesture recognition using automatic feature extraction and deep learning algorithms with memory. Big Data and Cognitive Computing **7**(2), 102 (2023)
12. Xu, Z., Zhao, J., Yu, Y., Zeng, H.: Improved 1D-CNNs for behavior recognition using wearable sensor network. Computer Communications **151**(X), 165–171 (2020)
13. Li, Q., Langari, R.: EMG-based HCI using CNN-LSTM neural network for dynamic hand gestures recognition. IFAC-PapersOnLine **55**(37), 426–431 (2022)
14. Gao, W., Zhang, L., Teng, Q., He, J., Wu, H.: DanHAR: Dual attention network for multimodal human activity recognition using wearable sensors. Applied Soft Computing **111**(X), 107728 (2021)
15. Bianchi, V., Bassoli, M., Lombardo, G., Fornacciari, P., Mordonini, M., De Munari, I.: IoT wearable sensor and deep learning: An integrated approach for personalized human activity recognition in a smart home environment. IEEE Internet of Things Journal **6**(5), 8553–8562 (2019)
16. Gil-Martín, M., San-Segundo, R., Fernández-Martínez, F., Ferreiros-López, J.: Improving physical activity recognition using a new deep learning architecture and post-processing techniques. Engineering Applications of Artificial Intelligence **92**(X), 103679 (2020)
17. Faisal, M. A. A., Abir, F. F., Ahmed, M. U., et al.: Exploiting domain transformation and deep learning for hand gesture recognition using a low-cost dataglove. Scientific Reports **12**(1), 21446 (2022)

18. Lv, X., Dai, C., Liu, H., et al.: Gesture recognition based on sEMG using multi-attention mechanism for remote control. Neural Computing and Applications **35**(19), 13839–13849 (2023)
19. Qiu, S., Fan, T., Jiang, J., Wang, Z., Wang, Y., Xu, J., Sun, T., Jiang, N.: A novel two-level interactive action recognition model based on inertial data fusion. Information Sciences **633**(X), 264–279 (2023)
20. Ikne, O., Allaert, B., Wannous, H.: Skeleton-based self-supervised feature extraction for improved dynamic hand gesture recognition. In: 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG), pp. 1–10. IEEE (2024)
21. Joshi, I., Utkarsh, A., Kothari, R., Kurmi, V. K., Dantcheva, A., Dutta Roy, S., Kalra, P. K.: Sensor-invariant fingerprint ROI segmentation using recurrent adversarial learning. In: 2021 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2021)
22. Jiang, Y., Zhang, H., Zhang, J., Wang, Y., Lin, Z., Sunkavalli, K., Chen, S., Amirghodsi, S., Kong, S., Wang, Z.: SSH: A self-supervised framework for image harmonization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4832–4841. IEEE (2021)
23. Chen, H., Gouin-Vallerand, C., Bouchard, K., et al.: Contrastive Self-Supervised Learning for Sensor-Based Human Activity Recognition: A Review. IEEE Access **12**(X), 1–20 (2024)
24. Lee, Y., Kim, H., Ko, Y.: Contrastive learning from temporal adjustments for wearable-based gesture recognition. In: ACM Ubiquitous Computing (UbiComp), pp. 1–10. ACM (2021)
25. Wang, Q., Chen, D., Zhang, Z.: Self-supervised learning for cross-user gesture recognition via contrastive predictive coding. IEEE Transactions on Mobile Computing **XX**(X), 1–13 (2022)
26. Garg, M., Ghosh, D., Pradhan, P. M.: Multiscaled multi-head attention-based video transformer network for hand gesture recognition. IEEE Signal Processing Letters **30**(X), 80–84 (2023)
27. Shaw, P., Uszkoreit, J., Vaswani, A.: Self-attention with relative position representations. arXiv preprint arXiv:1803.02155 (2018)
28. Zhang, Z., Xu, Z. Q. J.: Implicit regularization of dropout. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(6), 4206–4217 (2024)
29. Liu, Z., Xu, Z., Jin, J., et al.: Dropout reduces underfitting. In: International Conference on Machine Learning, pp. 22233–22248. PMLR (2023)
30. Peerthum, Y., Stamp, M.: An empirical analysis of the shift and scale parameters in BatchNorm. Information Sciences **637**(X), 118951 (2023)