Attention mechanism based multimodal feature fusion network for human action recognition

Attention mechanism based multimodal feature fusion network for human action recognition

Xu Zhao [a], Chao Tang [a,*], Huosheng Hu [b], Wenjian Wang [c], Shuo Qiao [a], Anyang Tong [d]

[a] School of Artificial Intelligence and Big Data, Hefei University, Hefei, China
[b] School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ England, UK
[c] School of Computer and Information Technology, Shanxi University, Taiyuan, China
[d] School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China

Keywords:
Human Action Recognition
Feature Extraction and Fusion
Attention Mechanism
Multimodal Features

Current human action recognition (HAR) methods focus on integrating multiple data modalities, such as skeleton data and RGB data. However, they struggle to exploit motion correlation information in skeleton data and rely on spatial representations from RGB modalities. This paper proposes a novel Attention-based Multimodal Feature Integration Network (AMFI-Net) designed to enhance modal fusion and improve recognition accuracy. First, RGB and skeleton data undergo multi-level preprocessing to obtain differential movement representations, which are then input into a heterogeneous network for separate multimodal feature extraction. Next, an adaptive fusion strategy is employed to enhance the integration of these multimodal features. Finally, the network assesses the confidence level of weighted skeleton information to determine the extent and type of appearance information to be used in the final feature integration. Experiments conducted on the NTU-RGB + D dataset demonstrate that the proposed method is feasible, leading to significant improvements in human action recognition accuracy.

1. Introduction

In recent years, advancements in AI and computer vision have driven noteworthy progress in human action recognition (HAR), leading to impactful applications in public safety[1,2], medical monitoring[3], human–computer interaction[4], and self-driving cars[5]. HAR relies on capturing human body data through sensors and applying algorithmic models to extract and recognize human behavioral features[6,7]. However, traditional approaches have primarily focused on understanding RGB data, mimicking human visual perception. This reliance on RGB data introduces several limitations, such as sensitivity to point-of-view, background interference, and lighting conditions. These challenges highlight the urgent need for research into multimodal HAR, which can integrate diverse data sources to overcome the inherent limitations of single-modality approaches.

Compared with traditional RGB modal data, multimodal data can provide richer representational information for action recognition. For example, RGB modal data contains spatial and temporal information of human actions but is very sensitive to dynamically changing environments and complex.

scenes. Skeleton modal data is robust to changes in the background and environments of human actions, simple in structure and rich in information, but lacks the spatial information of the action. The fusion of RGB data and skeleton data is an effective strategy as their strengths and weaknesses can be comple-mentary in the main features such as appearance, posture, geometry, illumination and viewing angle [8,9,10,11,12,13,14,15].

Currently, there are many multimodal Human action recognition methods that use multimodal data fusion for decision making. Li et al [16] proposed an innovative dual-stream network architecture with three core components: a Spatial-Temporal Graph Convolutional Network (ST-GCN) for extracting skeleton features from skeleton data, an R(2 + 1)D network and a multimodal network that realizes feature complementarity between the two modal information at the semantic feature level. Duan et al[17] innovatively stacked the skeleton data to generate a 3D heat map body instead of the original skeleton data, and then extracted the RGB features and heat map body features through a two-stream 3D Convolutional Neural Network (3D-CNN) framework to realize the two types of visual features, respectively.

However, these methods rely heavily on the spatial information in the RGB modality, which has the following two problems: i) The RGB modality contains a large amount of background information which may

introduce a large amount of negative noise. It is difficult to construct corresponding strong correlation information between the features of these backgrounds and the actions, ii) The motion correlation information in the skeleton modality is not fully explored. To overcome the above challenges, this paper proposes an attention-based recognition network for multimodal feature fusion.

First, the proposed AMFI-Net accomplishes the alignment of two features and the dynamic fusion of two features through an efficient attention network. Secondly, it projects the features into the RGB features and the skeleton features to obtain the more expressive fusion features. Thirdly, it dynamically learns the importance of each modality and assigns the weights to them through a gate control unit. Fourthly, the importance of each modality is dynamically learned by the gate control unit and weights are assigned. Finally, the two features are combined and a SoftMax classifier is used to achieve accurate classification of different classes of actions. The main research work of this paper includes:

- An attention-based multimodal feature fusion recognition network architecture is proposed for extracting RGB and skeleton features and fusing two modal features for attention weighting. The complementarity of the two modal features is realized and the richness of the features is enhanced.
- A feature fusion network architecture with dynamically assigned weights is proposed to dynamically assign the confidence weights of each modality according to the reliability of the modality to fully utilize the complementarity of the two modalities.
- A novel joint loss function is proposed for network model training, which realizes the information sharing among networks and effectively promotes better collaborative learning among networks.
- A series of ablation and comparison experiments are successfully conducted by using the public dataset NTU-RGB + D to verify the effectiveness of the proposed method.

The rest of the paper is organized as follows. Section 2 briefly reviews some research works related to this research. In Section 3, a novel human behavior recognition method is proposed to better extract and fuse two different modal features. Experiments are conducted in Section 4 to verify the feasibility and performance of the proposed approach. Finally, a brief conclusion and future work are given in Section 5.

2. Related works

2.1. RGB modality based human action recognition

RGB modality refers to images or video sequences captured by an RGB camera, which are used for reproducing the visual content as the human eyes do. The two-stream network framework is widely used for RGB modality-based human behavior recognition, which typically consists of two convolutional neural network branches that process different feature inputs extracted from RGB videos. It was proposed by Simonyan et al[18], aiming at capturing the appearance and motion features of a video. Its result is obtained as a behavior recognition output through an integration strategy.

Feichtenhofer et al[19] improved the fusion strategy by changing the practice of fusing the two-stream network at the end. The fusion was implemented right at the middle layer of the network, which not only improved the network performance but also reduces its parameter size. Wang et al [20] enumerated several advanced CNN architectures, including GoogleNet [21] and VGG-16[22], and analyzed their performance and accuracy in a dual-stream network environment. Moreover, Chen et al[23] integrated the semi-coupling concept into a dual-stream network architecture and achieved good results in the application of very low-resolution behavior recognition. Tu et al[24] designed a multi-stream network based on the dual-stream network architecture, which can integrate multiple semantic information from RGB modalities and optical flow modalities for human behavior recognition. It achieved excellent performance with a limited number of training samples.

3D-CNN is a direct extension of 2D-CNN, which can effectively model information in both spatial and temporal dimensions simultaneously through an additional dimension. A 3D-CNN was proposed by Ji et al [25], which took advantage of 3D convolution to simultaneously extract features in both spatial and temporal dimensions to efficiently capture and analyze motion information in successive frames. Tran et al[26] proposed a Convolutional 3D (C3D) deep network architecture in which all the convolutional kernels have a size of $3 \times 3 \times 3$. As it has many parameters, C3D required a large amount of training data. Moreover, Qiu et al[27] decomposed the 3D convolution into a 2D spatial convolutional layer and a 1D temporal convolutional layer, for the decomposition of spatio-temporal convolutional neural networks. It reduced the number of parameters and improved the computational efficiency. Tran et al[28] designed the $R(2 + 1)D$ network with fewer parameters than the standard 3D convolution for the capture of temporal dynamic features between neighboring frames. Hara et al[29] proposed Resnet 3D by extending the Resnet residual network from 2D convolution to 3D convolution to obtain accurate recognition results.

Recent Transformer-based human behavior recognition has achieved good results, ActionForme proposed by Zhang et al[30] combines the spatio-temporal attention mechanism with Transformer to directly model the global spatio-temporal relationships of video sequences. Bertasius et al[31] proposed a TimeSformer in the TimeSformer-based Transformer's video classification framework, which handles RGB modalities through chunked spatio-temporal attention, performs well on the Kinetics dataset. Although the RGB modality based human behavior recognition has achieved satisfactory results, some problems such as complex background interference and overfitting remain to be solved.

2.2. Skeleton modality based human action recognition

The human skeleton structure is naturally a graph in non-Euclidean space, and skeleton sequence data is a graph sequence. Skeleton data can provide body structure and pose information and is robust to background changes. In comparison with the traditional Recurrent Neural Network (RNN)-based approaches[32,33,34,35], Graph Convolutional Network (GCN) has a superior and successful application in the field of human motion recognition due to the great advantages of Graph Convolutional Network (GCN) for processing non-Euclidean data.

The core of the GCN model based human action recognition task is to model the human skeleton using the graph structure and extract the spatio-temporal features of joints through the graph convolution operation. Yan et al[36] pioneered the ST-GCN model, which consists of two parts: a graph convolution network and temporal convolution network. By fusing these two networks, it effectively captured the spatio-temporal features of human actions. Li et al[37] proposed Actional-Structural Graph Convolution Network (AS-GCN) to enhance the recognition of complex actions by modeling the higher-order relationships of joint points in the human skeleton to capture the higher-order dependencies of joint points.

Shi et al[38] proposed a two-stream adaptive graph convolutional network (2 s-AGCN) by combining the adaptive graph mechanism of the spatial graph convolution module and the information of the two data streams by end-to-end approach. In this way, it could learn the adaptive graph and use decision fusion for category prediction. Furthermore, Shi et al[39] proposed the Directed Graph Neural Network (DGNN), which can fully extract the spatiotemporal features of skeleton sequence data and further improve the performance of the dual-stream framework.

In recent years, graph neural networks (GNN) have made progress in the field of human action recognition. Liu et al[40] modeled long-distance dependencies by multiscale spatio-temporal mapping convolution in multiscale gated 3D ConvNets (MS-G3D), which significantly improves the performance of recognition of complex actions; Chen et al [41] introduced a channel topology optimization mechanism into the

CTR-GCN in CTR-GCN, which dynamically adjusts the relationship weights between joints and performs well on the NTU-RGB + D dataset. Although skeleton-based human action recognition has many advantages, its performance is unsatisfactory in some subtle categories of hand movements.

2.3. Human action recognition based on RGB and skeleton fusion modalities

As RGB data and skeleton data are complementary in nature, some recent works have deployed deep learning architectures to fuse RGB and skeleton data for human behavior recognition. Zhao et al[9] used RNN to extract joint motion information from skeleton data and CNN to process the RGB data to extract appearance information, i.e., a two-stream network to fuse the RGB data and the skeleton data. Zolfaghari et al[10] used a three-stream 3D-CNN to process pose, motion, and raw RGB images, and then utilized a Markov chain model to fuse the spatio-temporal information and dynamic features to further enhance the abstraction and representation of the action features.

Das et al[42] used a graphic CNN to extract the skeleton information and a 3D-CNN to extract the RGB modal features by introducing an attention mechanism for information extraction and a weight calculation for capturing the correlation and importance between different modal data more efficiently. Joze et al[43] proposed a Multimodal Transfer Module (MMTM) with reference to SENet[44] for integrating 3D-CNN processed RGB modalities and two-layer feature maps of 2D skeleton modalities to obtain attention weights and matrix for early fusion. Moon et al[45] used 2D-CNN and temporal convolution to model RGB and 2D skeleton modalities, and applied the attention mechanism and weighted summation to perform feature fusion to achieve a deeper level of fusion of multimodal data.

Das et al [46] designed an end-to-end spatio-temporal attention network using RGB data and 3D skeleton data as inputs to realize the combination of spatial and temporal attention. Xu et al[47] designed a human action recognition model based on multimodal data to extract features through a deep learning model. They utilized a multimodal fusion and alignment network, namely Bilinear Pooling and Attention Network (BPAN), for feature fusion. Unlike these approaches, this paper extends both GCN and CNN networks to the overall multimodal feature fusion recognition network architecture to explore the complementary nature of feature fusion between heterogeneous networks. Compared with Vision Transformer Bilinear Pooling and Attention Network (VT-BPAN) proposed by Sun et al[48], AMFI-Net constructs global feature associations through nonlocal mean operations and preserves the initial feature morphology by combining residual connections, which effectively solves the VT-BPAN which effectively solves the problem of insufficient local dependency between modalities in VT-BPAN. Similar to the Multimodal Audio-Image and video Action Recognizer (MAiVAR) framework recently proposed by Shaikh et al[49], AMFI-Net and MAiVAR effectively reduce the computational complexity in the processing of RGB modalities.

3. Research Methodology

To better extract and fuse two different modal features, a novel human behavior recognition method based on a multimodal fusion attention network is proposed, as shown in Fig. 1. It consists of three modules, which are described in this section. The pre-processed RGB and skeleton data are inputted into Resnet 3D and AGCN networks, respectively, to extract the appearance features of the RGB and the motion features of the skeleton. Then, the feature association module is used to fuse the features and to solve the problem of weak feature characterization. Feature weighting is used to make the fusion features more representative, and the modal information is decided by feature confidence to improve the complementarity of different modes. Finally, Softmax is used to classify the fusion features for decision making.

3.1. Data preprocessing module

In the RGB video classification task, the presence of noise and background affects the classification accuracy and increases the computational burden and risk of overfitting. In the RGB video classification task, the presence of noise and background affects the classification accuracy and increases the computational burden and risk of overfitting. To cope with these challenges, we perform a series of preprocessing operations, which not only focuses attention on human targets in RGB videos, reduces computational memory consumption, and improves the efficiency and utility of the algorithm. Moreover, experimental results in resource-constrained devices by Yun et al[50] show that using preprocessing techniques to optimize the operations and integrating them into the real-time processing flow does not lead to significant performance degradation. The operations of the data

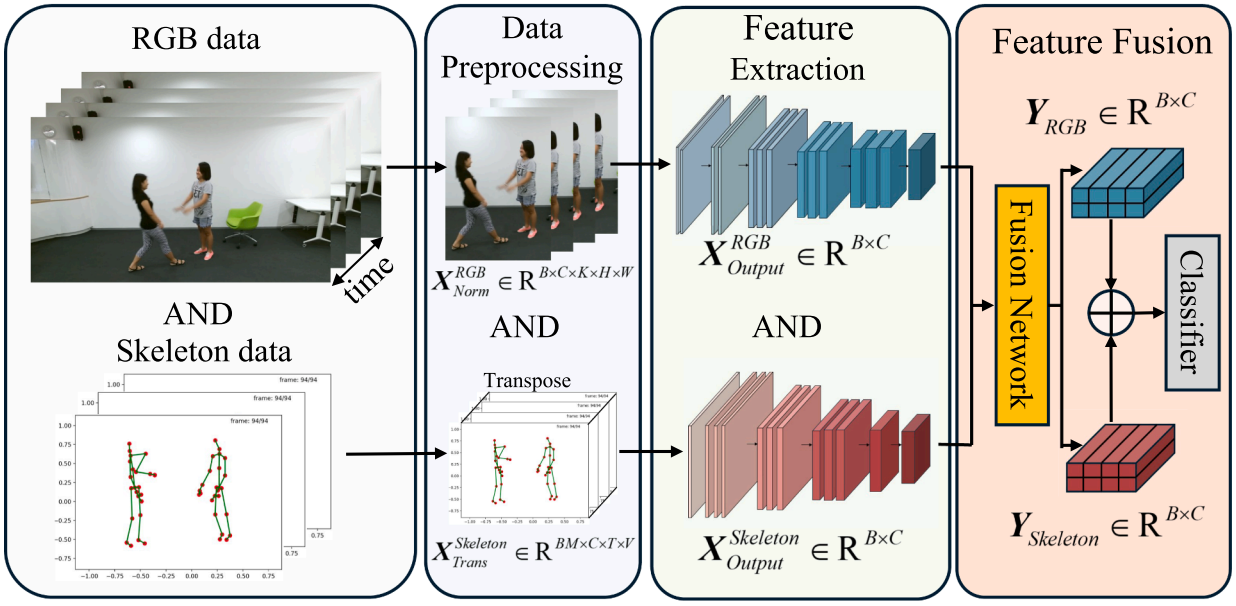


Fig. 1. Illustration of the proposed network architecture.

preprocessing module are as follows.

(1) The video V is cropped at equal time intervals and only the K frames related to the human body are retained. The resulting set of video frames is denoted as $\boldsymbol{X}^{RGB} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_K\}$, the total frame number of the video V is T, and the default frame extraction

interval is $\lfloor T/K \rfloor$ frames, and a sequence $\boldsymbol{X}_{Input}^{RGB} \in \mathrm{R}^{B\times C\times K\times H\times W}$ of images with the height of H, the width of W, and the number of channels of C is obtained, where B is the number of batches of inputs into the network at a time.

(2) To ensure that all the samples of the dataset input into the network focus on the character's actions and reduce the interference of useless background as much as possible, this paper uses the Crop$(\cdot)$ method to crop the task-centered activity region of all the image sequences, and the size of the cropped data image is 1000 × 1000, the processing method is as follows.

$$\boldsymbol{X}_{Crop}^{RGB} \in \mathrm{R}^{B\times C\times K\times H\times W} \leftarrow \mathrm{Crop}(\boldsymbol{X}_{Input}^{RGB}) \tag{1}$$

(3) All the image sequences in the dataset are then processed by the Resize$(\cdot)$ method, and the size of the image sequences is adjusted to the size dimensions required by the deep learning network, and the sample image data is processed to a size of 112 × 112 as shown below:

$$\boldsymbol{X}_{\mathrm{Resize}}^{RGB} \in \mathrm{R}^{B\times C\times K\times H\times W} \leftarrow \mathrm{Resize}(\boldsymbol{X}_{Crop}^{RGB}) \tag{2}$$

(4) To effectively improve the overall training process speed and the stability of the model, as well as accelerate the convergence of the model, the data needs to be normalized after adjusting the size. In this paper, we use the maximum-minimum normalization for the normalization operation, which is a linear transformation of the processed data $\boldsymbol{X}_{\mathrm{Resize}}^{RGB}$ of the same size, and set $\boldsymbol{X}_{\min}$ and $\boldsymbol{X}_{\max}$ as the minimum and maximum values of the attributes, respectively, and map the original value of $\boldsymbol{X}_{\mathrm{Resize}}^{RGB}$ to

the value of the interval [0,1] through the maximum-minimum normalization, as shown below:

$$\boldsymbol{X}_{Norm}^{RGB} \in \mathrm{R}^{B\times C\times K\times H\times W} = \frac{\boldsymbol{X}_{\mathrm{Resize}}^{RGB} - \boldsymbol{X}_{\min}}{\boldsymbol{X}_{\max} - \boldsymbol{X}_{\min}} \tag{3}$$

(5) To alleviate the overfitting of data in the RGB training set, this paper generates new sample images

by data augmentation, which is performed by randomly rotating the training set of the image data counterclockwise by 90°, 180°, 270°, flipping.

horizontally or vertically, and scaling at a certain ratio ∂(0–1). Its generalization ability can be improved while avoiding overfitting and improving model robustness.

(6) To obtain a better extraction of the skeleton features, this paper adopts a method, which is similar to 2 s-AGCN[38], to process the skeleton data $\boldsymbol{X}^{Skeleton} \in \mathrm{R}^{B\times C\times T\times N\times M}$, where B denotes the number of batches, C denotes the number of channels, T denotes the frames in the sequence of skeleton data, N denotes the number of joints, and M denotes the skeleton in the frame. The skeleton data dimensions are first adapted by combining the B and M dimensions as a new dimension $BM = B \times M$.

$$\boldsymbol{X}_{Trans}^{Skeleton} \in \mathrm{R}^{BM\times C\times T\times N} \leftarrow \mathrm{Transpose}(\boldsymbol{X}^{Skeleton}) \tag{4}$$

3.2. Feature extraction module

The feature extraction network model consists of two main streams: processing RGB data stream and processing skeleton data stream. The Resnet3D 18[29] network is used to extract the appearance information of the RGB data and the AGCN[38] network to extract the motion information of the skeleton data, respectively.

For RGB video streaming, the Resnet3D 18 network framework is shown in Fig. 2(a). The first three layers of the network have 64 output channels, the middle two layers have 128 output channels, the next two layers have 256 output channels, and the last two layers have 512 channels. Compared to deeper networks (e.g., ResNet3D 50), ResNet3D-18 has only 33 M parameters, which significantly reduces the risk of overfitting. The structure of the Resnet3D 18 network is similar to that of the 2D ResNet, with each layer containing a residual.

structure similar to that of the 2D ResNet[51] network. The Resnet3D 18 network uses Basic Block as the residual block. As shown in Fig. 2(b), the Basic Block consists of two 3 × 3 × 3 convolutional layers, each followed by a bulk normalization (BN) layer and a ReLU activation function layer, whose time-dimensional convolution extracts local temporal features[51]. The feature vector $\boldsymbol{X}_{Norm}^{RGB} \in \mathrm{R}^{B\times C\times K\times H\times W}$ of the input Basic Block output is characterized by:

$$\boldsymbol{X}_{Out}^{RGB} = f(\boldsymbol{X}_{Norm}^{RGB}) + \boldsymbol{X}_{Norm}^{RGB} \tag{5}$$

where $f(\cdot)$ is a residual module, the data preprocessed $\boldsymbol{X}_{Norm}^{RGB}$ input into the continuous Basic Block residual block to get the final output:$\boldsymbol{X}_{Output}^{RGB} \in \mathrm{R}^{B\times C}$ features, where C denotes the number of RGB channels, the output is a 1D vector of 512 features.

Fig. 3(a) shows the overall network framework of AGCN, which has 64 output channels in the first three layers, 128 output channels in the middle three layers, and 256 output channels in the last three layers. AGCN dynamically adjusts the connection weights between joints in a data-driven way to improve the robustness to noisy skeleton data[38]. To ensure that the skeleton data features are consistent with the RGB features, an adaptive graph convolution block with 512 channels is added. AGCN is a classical model based on the ST-GCN[36] model, which replaces the original spatial graph convolution with an adaptive graph convolution block to model the joint motion in consecutive frames, and the temporal convolution kernel in the adaptive graph convolution block slips along the frame sequences to learn the changes of the joint trajectories. Fig. 3(b) shows the adaptive convol-ution blocks, each adaptive graph convolution block consists of adaptive convolution layers Convs, BN.

Relu, and Dropout. The adaptive graph convolution layer Convs optimizes the topology of the skeleton joints and overcomes the limitation of the traditional ST-GCN that relies on fixed physical connections. Fig. 3(c) shows the ada-ptive convolutional layer, the data-driven network layer, which combines the similarity adjacency matrix with the original adjacency matrix to jointly follow the network for training and updating parameters. The feature vector of the input network is $\boldsymbol{X}_{Trans}^{Skeleton} \in \mathrm{R}^{BM\times C_{in}\times T\times N}$, and the output network features are:

$$\boldsymbol{X}_{Out}^{Skeleton} = \sum_{k=1}^{K} W_k \boldsymbol{X}_{Trans}^{Skeleton} (A_k + B_k + C_k) \tag{6}$$

The AGCN network adopts the spatial division strategy in ST-GCN [36] to divide the joint K into three parts: root, centripetal, and centrifugal nodes. Each of the three parts undergoes parameter updating and possesses different importance during the training process, which is finally superimposed as the output network features. Each part of the

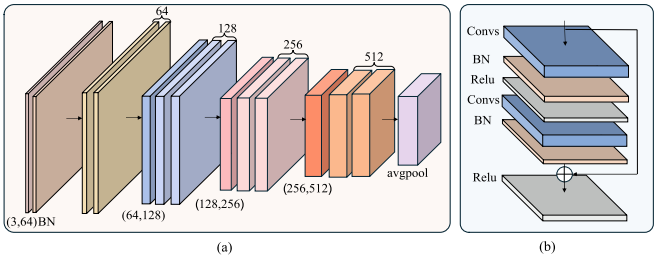


Fig. 2. Resnet3D 18 network structure. (a) Overall network framework of Resnet3D 18 (b) Basic Block structure in Resnet3D 18.

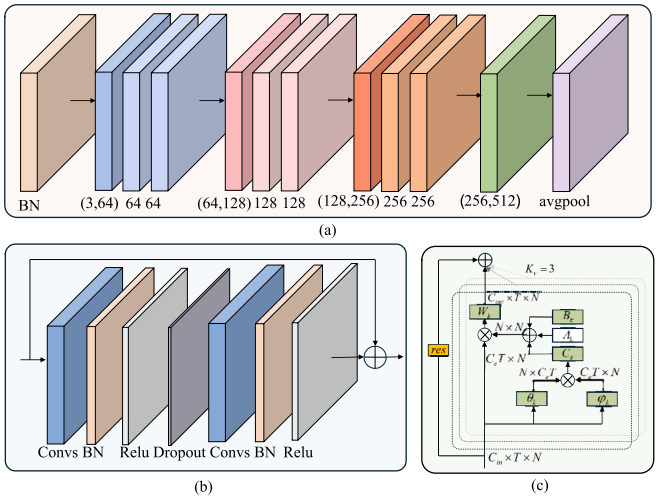


Fig. 3. AGCN network structure. (a) AGCN overall network framework (b) AGCN adaptive convolution block (c) AGCN adaptive convolution layer.

parameter update contains three adjacency matrices A_k, B_k and C_k.A_k represents the adjacency matrix of the physical information of the human body, where the weights are not updated. B_k is also an $N \times N$ matrix with all zero initial values, where the elements are optimized along with the other parameters during the training process. C_k is a data correlation graph, which learns a unique data correlation graph for each sample. Similar to the Resnet 3D network, the $\boldsymbol{X}_{Pile}^{Skeleton}$ was input into the AGCN network to obtain the $\boldsymbol{X}_{Output}^{Skeleton} \in \mathrm{R}^{B\times C}$. After the Resnet3D 18 and AGCN networks and the dimensionality transformation, the RGB features, and the skeleton features have the same dimensions.

3.3. Amfi-net module

The AMFI-Net module takes the RGB feature $\boldsymbol{X}_{Output}^{RGB} \in \mathrm{R}^{B\times C}$ and the skeleton feature $\boldsymbol{X}_{Output}^{Skeleton} \in \mathrm{R}^{B\times C}$ processed by the feature extraction module as input. First, the feature maps of the two different modalities are used to generate new global association features. Second, the global association features are used to calculate the attention feature weights. Third, the attention feature weights are used to weigh the original features. Fourth, the modulated features are dynamically allocated with the confidence of the two features through the gating mechanism. Finally, the two confidence-allocated features are summed up and classified for action recognition. The AMFI-Net module has three main components, as shown in Fig. 4: feature association, feature weighting, and feature confidence.

3.3.1. Feature association

To integrate the information from various sources into one feature vector, this paper firstly adopts the splicing method to reduce the computational complexity during the model training process and retain the information contained in two different modal features as much as possible.

$$\boldsymbol{X}_{Fusion} \in \mathrm{R}^{B\times M} = \mathrm{Concatenation}(\boldsymbol{X}_{Output}^{Skeleton}, \boldsymbol{X}_{Output}^{RGB}) \tag{7}$$

where Concatenation$(\cdot)$ is the feature splicing operation and $M = 2 \times C$.$\boldsymbol{X}_{Output}^{Skeleton}, \boldsymbol{X}_{Output}^{RGB} \in \mathrm{R}^{B\times C}$ denotes the skeleton features processed by the feature extraction module and the RGB features, respectively.

The information output from Resnet3D 18 and AGCN networks is feature extraction of local regions, which can reduce the number of parameters. However, the extracted information is limited by its perceptual field and lacks global information, which is not favorable for constructing remote dependencies. Therefore, this paper creates a feature correlation module based on the nonlocal mean operation [52,53] to dynamically capture the long-range spatio-temporal dependencies by calculating the similarity weights of all positional feature pairs and correlating each feature with all feature data. It computes the correlation by calculating the output of each pixel location through neighborhood computation and all positions in the image. The computed response values are then used to reassign feature weighting to the original features as shown in Fig. 5.

The operation for the current position feature vector $\boldsymbol{x}_i \in \boldsymbol{X}_{Fusion}$ is:

$$\boldsymbol{y}_i = \frac{1}{c(\boldsymbol{x})} \sum_{\forall j} f(\boldsymbol{x}_i, \boldsymbol{x}_j) g(\boldsymbol{x}_j) \tag{8}$$

where $\boldsymbol{x}_j \in \boldsymbol{X}_{Fusion}$ is the feature vector of all possible other locations,$\boldsymbol{y}_i$ denotes the output feature,$\boldsymbol{y}_i$ and $\boldsymbol{x}_i$ have the same feature size,$g(\cdot)$ performs a convolutional transform operation of 1×1 on the features at each location $\boldsymbol{x}_j$, and $f(\cdot)$ is an embedded Gaussian function that computes the correlation between $\boldsymbol{x}_j$ and $\boldsymbol{x}_i$:

$$f(\boldsymbol{x}_i, \boldsymbol{x}_j) = e^{\theta(\boldsymbol{x}_i)^T \beta(\boldsymbol{x}_j)} \tag{9}$$

$$c(\boldsymbol{x}) = \sum_{\forall j} f(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{10}$$

where $\theta(\boldsymbol{x}_i) = W_\theta \boldsymbol{x}_i \beta(\boldsymbol{x}_j) = W_\beta \boldsymbol{x}_i, \theta(\cdot)$ and $\beta(\cdot)$ are the convolutional transform operations of $1 \times 1, W_\theta$ and W_β are the weight matrices.

To include the feature similarity information after the nonlocal mean operation without destroying the initial feature morphology and information, we insert the residual connection in the nonlocal mean operation:

$$\boldsymbol{z}_i = W_z \boldsymbol{y}_i + \boldsymbol{x}_i \tag{11}$$

where $''+\boldsymbol{x}_i''$ denotes the residual connection,W_z is its weight matrix,$\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ are given in Eq. (6). The features $\boldsymbol{z}_i$ are obtained by processing $\boldsymbol{x}_i$ corresponding to i at all locations. Then features are stacked to obtain $\boldsymbol{Z}_{Fusion} \in \mathrm{R}^{B\times M}$.

3.3.2. Feature weighting

Since features in different regions have different importance in the recognition task, appropriate weighting of $\boldsymbol{Z}_{Fusion} \in \mathrm{R}^{B\times M}$ is essential. Feature weighting is achieved by introducing a channel attention mechanism, which is inspired by the Squeeze-and-Excitation Networks (SENet) of Hu et al[54] generating channel-level attention weights through a fully connected layer and a Sigmoid activation function, which enables the model to dynamically adjust the importance of each modality according to the real-world situation. As shown in Fig. 6, by weighting the features, the model can be more flexible to adapt to different tasks and make the fused features more representative and

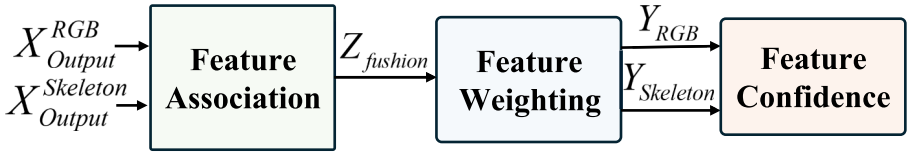


Fig. 4. Feature fusion module framework.

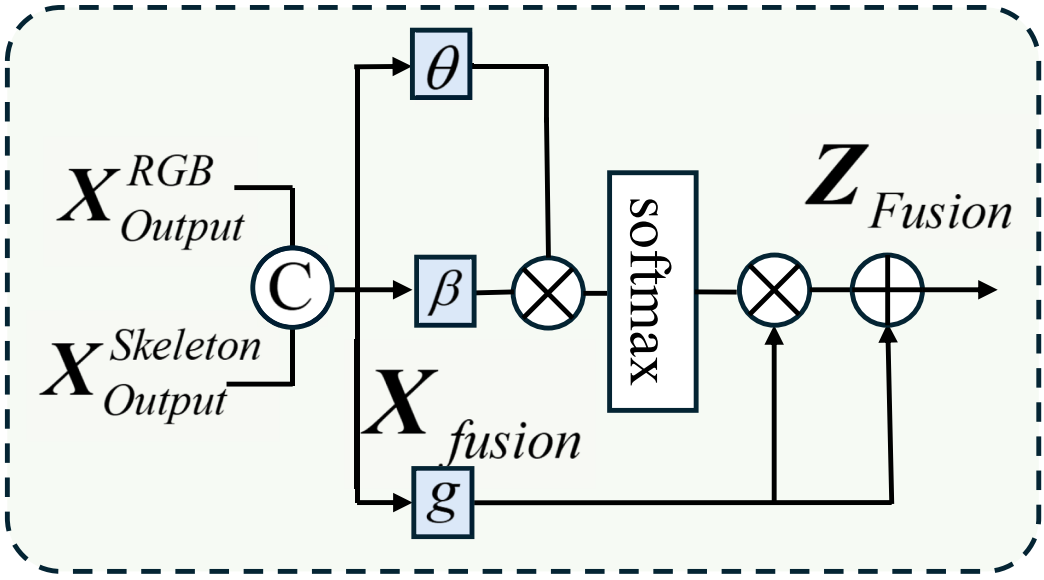


Fig. 5. Feature association module in which ⊕ denotes the residual connection of elements ⊗ denotes matrix multiplication.

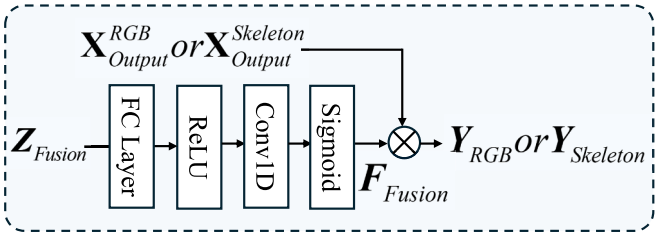


Fig. 6. Feature weighting module in which ⊗ denotes matrix multiplication.

distinguishable.

To eliminate the different magnitudes and ranges between the two distinctive features, and speed up the model convergence and improve the model stability, the fusion feature $\boldsymbol{Z}_{Fusion}$ is first normalized by L2 regularization:

$$\boldsymbol{Z}_{Fusion} = \text{sign}(\boldsymbol{Z}_{Fusion})\sqrt{|\boldsymbol{Z}_{Fusion}|} \in \mathrm{R}^{B\times M} \tag{12}$$

$$\boldsymbol{Z}_{Fusion} = \frac{\boldsymbol{Z}_{Fusion}}{\|\boldsymbol{Z}_{Fusion}\|_2} \in \mathrm{R}^{B\times M} \tag{13}$$

Then $\boldsymbol{Z}_{Fusion}$ is put into the fully connected layer and the ReLU layer to obtain new fusion features that are more expressive:

$$\boldsymbol{Z}_{Fusion} = \text{ReLU}(\text{FC}(W_f\boldsymbol{Z}_{Fusion})) \in \mathrm{R}^{B\times C} \tag{14}$$

Where $\text{FC}(\cdot)$ denotes the fully connected layer and W_f is its weight matrix. $\boldsymbol{Z}_{Fusion}$ is then convolved using one-dimensional convolution to compute the attention weight $\boldsymbol{F}_{Fusion}$:

$$\boldsymbol{F}_{Fusion} = \sigma(\text{Conv1}D(W_C\boldsymbol{Z}_{Fusion})) \in \mathrm{R}^{B\times C} \tag{15}$$

where $\sigma(\cdot)$ denotes the sigmoid activation function and $Conv1D(\cdot)$ denotes one-dimensional convo-lution.

Finally, the attention weights $\boldsymbol{F}_{Fusion}$ are linearly multiplied with the two input feature vectors to output the attention-weighted feature maps $\boldsymbol{Y}_{RGB}$ and $\boldsymbol{Y}_{Skeleton}$, respectively.

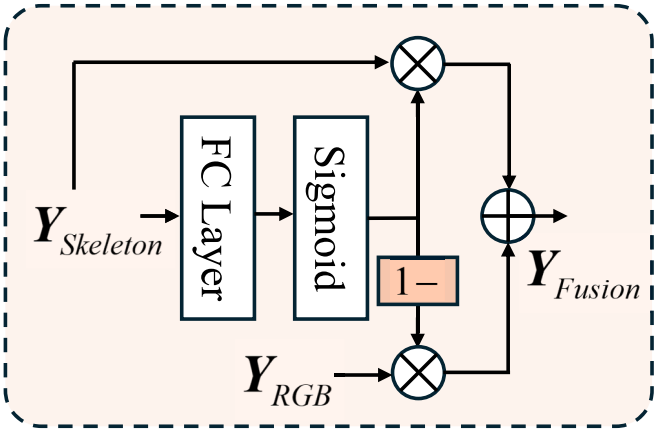


Fig. 7. Feature confidence module in which ⊕ denotes elementwise addition ⊗ denotes elementwise multiplication.

$$\boldsymbol{Y}_{RGB} = \boldsymbol{F}_{Fusion} \times \boldsymbol{X}_{Output}^{RGB} \in \mathrm{R}^{B\times C} \tag{16}$$

$$\boldsymbol{Y}_{Skeleton} = \boldsymbol{F}_{Fusion} \times \boldsymbol{X}_{Output}^{Skeleton} \in \mathrm{R}^{B\times C} \tag{17}$$

3.3.3. *Feature confidence*

Since different heterogeneous features do not have the same feature reliability at the same location, the direct addition of two heterogeneous features $\boldsymbol{Y}_{RGB}$ and $\boldsymbol{Y}_{Skeleton}$ as fusion features may result in not being able to fully utilize the complementary characteristics of the heterogeneous features. Therefore, in this paper, the two heterogeneous features are weighted with dynamic confidence, which combines the dynamic confidence learning idea of gated recurrent unit (GRU)[55], generates modal confidence weights through Sigmoid function, and dynamically adjusts the weights of modal features in the decision-making process according to the reliability of each modal feature in the current situation.

This is done by having a certain modal feature as the decision-making modality, and the reliability of that modal data will directly affect its contribution in feature integration and the degree of using the

other modal information[45]. For example, when the skeleton information itself is already accurate enough to independently accomplish the action recognition task, it will reduce the fusion network's dependence on the RGB modal features and pay more attention to the human motion information in the skeleton features. The specific framework structure is shown in Fig. 7.

In this paper, the skeleton modality is used as the decision-making modality. The input of the feature confidence module is the output of the feature weighting module $\boldsymbol{Y}_{Skeleton}$, which first passes through a fully connected layer and then a sigmoid activation function to obtain the gating matrix $\boldsymbol{Y}'_{Skeleton}$ such that its every element is located in [0,1] as shown below.

$$\boldsymbol{Y}'_{Skeleton} = \sigma(\mathrm{Fc}(W \cdot \boldsymbol{Y}_{Skeleton})) \in \mathrm{R}^{B \times C} \tag{18}$$

It is used for gating operations on $\boldsymbol{Y}_{Skeleton}$ and $\boldsymbol{Y}_{RGB}$: $\boldsymbol{Y}_{Skeleton} \odot \boldsymbol{Y}'_{Skeleton}$ and $\boldsymbol{Y}_{RGB} \odot (1 - \boldsymbol{Y}'_{Skeleton})$, where $\odot$ denotes element-by-element multiplication. Finally, the gating features of RGB and skeleton are simply combined by element-by-element summation to produce the composite action feature as follows:

$$\boldsymbol{Y}_{Fusion} = \boldsymbol{Y}_{Skeleton} \odot \boldsymbol{Y}'_{Skeleton} + \boldsymbol{Y}_{RGB} \odot (1 - \boldsymbol{Y}'_{Skeleton}) \tag{19}$$

In order that the feature confidence module can play a role in deciding which modality to use and how much modal information to use in the whole fusion network, a loss function L_{gate} is set here to train the gating module to force the learning of the decision-making modality's priority.

$$L_{gate} = -\log(1 - \boldsymbol{Y}'_{Skeleton}) \tag{20}$$

3.3.4. Combined training

In this paper, we use two deep learning models to extract deep features, i.e., $\boldsymbol{X}^{Skeleton}_{Output}$ and $\boldsymbol{X}^{RGB}_{Output}$, from the RGB stream and the skeleton stream, respectively, for characterizing the implicit semantics of the same class of actions corresponding to different modal data. The $\boldsymbol{X}^{Skeleton}_{Output}$ and $\boldsymbol{X}^{RGB}_{Output}$ are then fed into the AMFI-Net module for training to obtain the final fusion feature $\boldsymbol{Y}_{Fusion}$. Afterward, the predictions are normalized to a probability distribution by means of a SoftMax function. Algorithm 1 gives the pseudo-code for the implementation of our method. To facilitate the cooperative training of two deep learning networks with different modalities, we adopt the joint loss function to comprehensively optimize the whole fusion network framework.

$$L_{total} = L_{fushion} + \lambda L_{gate} \tag{21}$$

where $L_{fushion}$ is the cross-entropy loss function, and λ denotes the equilibrium factor with the initial value 1. This design can make full use of the complementary information between different modes to maximize the performance of the model during the training process.

Table 1
Action categories of the NTU RGB + D dataset.

Action number	Action category	Action number	Action category	Action number	Action category
1	drink water	21	take off a hat/cap	41	sneeze/cough
2	eat meal	22	cheer up	42	staggering
3	brush teeth	23	hand waving	43	falling
4	brush hair	24	kicking something	44	headache
5	drop	25	reach into pocket	45	chest pain
6	pick up	26	hopping	46	back pain
7	throw	27	jump up	47	neck pain
8	sit down	28	phone call	48	nausea/vomiting
9	stand up	29	play with phone/tablet	49	fan self
10	clapping	30	type on a keyboard	50	punch/slap
11	reading	31	point to something	51	kicking
12	writing	32	take a selfie	52	pushing
13	tear up paper	33	check time (from watch)	53	pat on back
14	put on jacket	34	rub two hands	54	point finger
15	take off jacket	35	nod head/bow	55	hugging
16	put on a shoe	36	shake head	56	giving object
17	take off a shoe	37	wipe face	57	touch pocket
18	put on glasses	38	salute	58	shaking hands
19	take-off glasses	39	put palms together	59	walking towards
20	put on a hat/cap	40	cross hands in front	60	walking apart

Table 2
Experimental environment and configuration.

Device	Version
Operating system	Windows10
Python version	3.8
Torch version	1.9.0
CPU	Inter i9-11900 K
GPU	NVIDIA GeForce RTX3080

Algorithm 1

Input: $\boldsymbol{X}^{RGB}$, $\boldsymbol{X}^{skeleton}$, Softmax, *Epoch*
Output: Action categories *Prediction*
obtain $\boldsymbol{X}_{Norm}^{RGB}$ and $\boldsymbol{X}_{Trans}^{skeleton}$ using (1) (2) (3) (4) with $\boldsymbol{X}^{RGB}$, $\boldsymbol{X}^{skeleton}$
obtain $\boldsymbol{X}_{train}^{RGB}$, $\boldsymbol{X}_{test}^{RGB}$, $\boldsymbol{X}_{train}^{skeleton}$ and $\boldsymbol{X}_{test}^{skeleton}$ under the benchmark
initialize Resnet 3D parameters $\boldsymbol{w}^{rgb}$.
initialize AGCN parameters $\boldsymbol{w}^{skeleton}$.
for $i = 1, 2, \ldots, Epoch$ **do**
train Resnet 3D with $\boldsymbol{X}_{train}^{RGB}$.
test Resnet 3D with $\boldsymbol{X}_{test}^{RGB}$.
obtain $Accuracy_i^{RGB}$ using (19).
obtain parameters $\boldsymbol{w}_i^{rgb}$.
if $i > 1$ **and** $Accuracy_i^{RGB} > Accuracy_{i-1}^{RGB}$
update $\boldsymbol{w}^{rgb} = \boldsymbol{w}_i^{rgb}$
end
for $i = 1, 2, \ldots, Epoch$ **do**
train AGCN with $\boldsymbol{X}_{train}^{skeleton}$.
test AGCN with $\boldsymbol{X}_{test}^{skeleton}$.
obtain $Accuracy_i^{skeleton}$ using (19).
obtain parameters $\boldsymbol{w}_i^{skeleton}$.
if $i > 1$ **and** $Accuracy_i^{skeleton} > Accuracy_{i-1}^{skeleton}$
update $\boldsymbol{w}^{skeleton} = \boldsymbol{w}_i^{skeleton}$
end
obtain $\boldsymbol{X}_{Output}^{RGB}$ and $\boldsymbol{X}_{Output}^{skeleton}$ using Resnet 3D and AGCN with parameters $\boldsymbol{w}^{rgb}$ and $\boldsymbol{w}^{skeleton}$.
initialize **FusionNetwork** parameters $\boldsymbol{w}^{fusion}$.
for $i = 1, 2, \ldots, Epoch$ **do**
train **FusionNetwork** with $\boldsymbol{X}_{Output}^{RGB}$ and $\boldsymbol{X}_{Output}^{skeleton}$.
obtain $Accuracy_i^{fusion}$ using (19).
obtain parameters $\boldsymbol{w}_i^{fusion}$.
if $i > 1$ **and** $Accuracy_i^{fusion} > Accuracy_{i-1}^{skeleton}$
update $\boldsymbol{w}^{fusion} = \boldsymbol{w}_i^{fusion}$
end
obtain $\boldsymbol{Y}_{Fusion}$ using **FusionNetwork** with parameters $\boldsymbol{w}^{fusion}$.
obtain *Prediction* using Softmax.

4. Experiment and analysis

4.1. Datasets

NTU-RGB + D dataset: this dataset[56] is one of the most widely used datasets in the field of human action recognition. It contains 60 categories of actions performed by 40 subjects, totaling 56,880 action sequences. Its 60 action categories are categorized into two types, one is for single-person actions and the other is for two-person interaction actions. Each action contains RGB data, depth data, and skeleton data for 25 joint points. The dataset is evaluated in two ways, Cross Subject (CS) and Cross View (CV). The CS benchmark divides the 40 subjects into two groups, the first group has 40,320 action sequences for the training samples, and the second group has 16,560 action sequences for the test samples. The CV benchmark uses action sequences recorded by cameras 2 and 3 for the training samples, totaling 16,560 action sequences. Sequences were used for the training samples, totaling 37,920 samples, and the action sequences recorded by camera 1 were used for the test samples, totaling 18,960 samples. Table 1 shows the specific action categories in the NTU RGB + D dataset.

UTD-MHAD dataset: UTD-MHAD[57] is a classic multimodal dataset in the field of human action recognition, which was released in 2015. This dataset contains 27 types of actions, performed by 8 subjects, with each action repeated 4 times, and a total of 861 valid samples are collected. Its action types cover daily actions (such as waving, clapping) and body movements (such as jumping, kicking), all of which are single-person actions. The data modalities include RGB videos, depth sequences, skeleton data, and inertial sensor data. The one used in this paper is the cross-subject protocol[57], specifically, the data from subjects numbered 1,3,5,7 are used for training and the data from subjects numbered 2,4,6,8 are used for testing.

4.2. Training details

The experimental environments and configurations are listed in Table 2. The training process was as follows: firstly, an RGB modality-

Table 3
Effect of feature fusion scheme on accuracy rate.

Group	Method					Accuracy_(%)
	Average	Multiplication	Max	Concatenation	**Ours**	
A	√					94.65
B		√				94.53
C			√			94.63
D				√		94.99
E					√	**95.82**

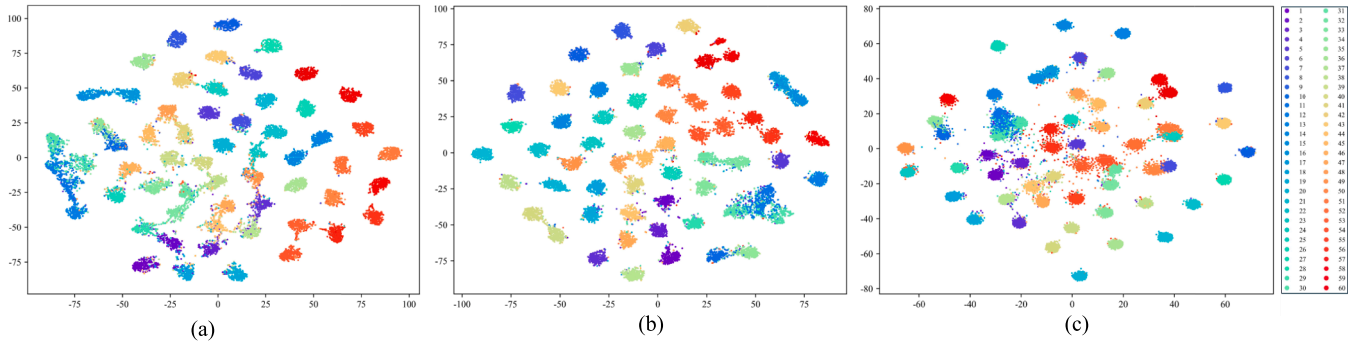


Fig. 8. *t*-SNE visualization based on single and fusion modes: (a) *t*-SNE visualization of RGB modes; (b) *t*-SNE visualization of skeleton modes; (c) *t*-SNE visualization of fusion features.

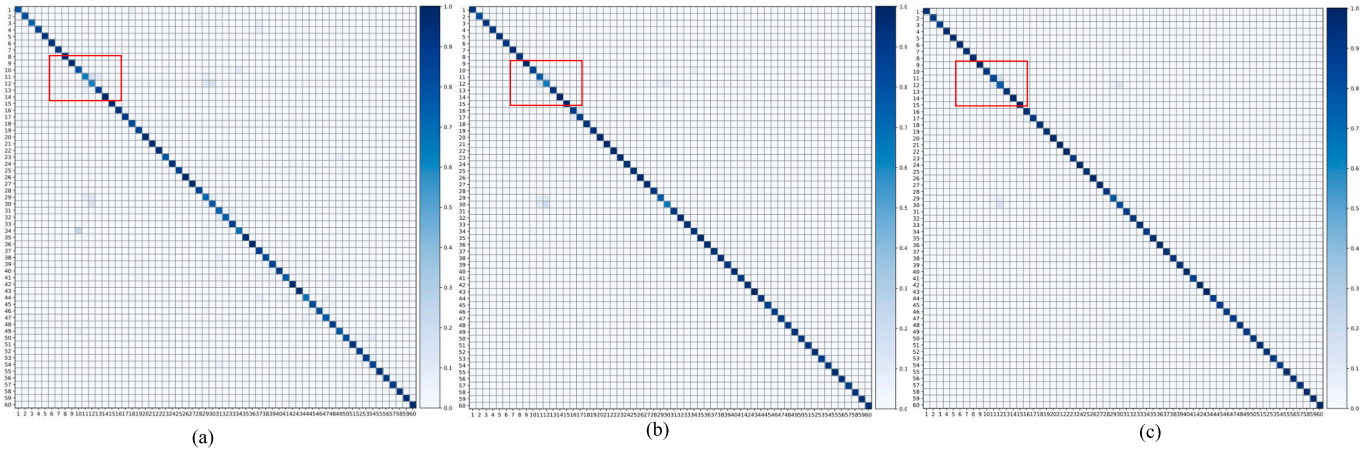


Fig. 9. Confusion matrix based on single and fusion modes. (a) Confusion matrix for RGB modes (b) Confusion matrix for skeleton modes (c) Confusion matrix for fusion features.

Table 4
Effect of multimodality on accuracy.

Group	Method			Accuracy_(%)
	Only RGB	Only skeleton	Ours	
A	√			88.67
B		√		93.24
C			√	**95.82**

Table 5
Effect of Weighting Factors on Accuracy.

Group	Method				Accuracy_(%)
	$\lambda = 1$	$\lambda = 1.5$	$\lambda = 3$	$\lambda = 5$	
A	√				95.22
B		√			**95.82**
C			√		95.71
D				√	95.09

Table 6
Impact of decision module on accuracy.

Group	Method		Accuracy_(%)
	RGB	skeleton	
A	√		88.26
B		√	95.82

only model and a skeleton modality-only model were trained. For the RGB stream, the RGB frame size was adjusted to 112*112, the learning rate and weight decay were set to 0.1 and 0.0005, respectively. The length of the video sequence was set to 16. For the skeleton stream, the learning rate and weight decay were set to 0.1 and 0.0005, respectively. The network was trained using the SGD optimizer, and the optimal training model was only used for feature extraction of the RGB stream data and the skeleton data, respectively. The training was done by minimizing the loss $L_{fushion}$ for classification only.

Two more modal pre-trained RGB-only and skeleton-only models were loaded for initializing the fusion network. The pre-training portion of the network was frozen while training the fusion network, and then the entire feature fusion portion was trained by minimizing the entire loss L_{total}. The number of training rounds for the fusion network was 6, and the learning rate was set to 0.01 and divided by 10 at the 3rd epoch. The overall model complexity of this paper is 43.26 M#Params, 5.67GFLOPs. In this paper, the accuracy rate was chosen as the

Table 7
Comparison of the model on the NTU-RGB + D dataset with prior art.

Methods	Multi-model	CS	CV
BI-LSTM[58]	Yes	85.4	91.6
FUSION[59]	Yes	91.8	94.9
MSAF[60]	Yes	92.24	–
Separable STA[46]	Yes	92.2	94.6
PoseMap[61]	Yes	91.7	95.2
MMTM[43]	Yes	**91.9**	95.3
SGM-Net[16]	Yes	89.1	**95.9**
AMFI-Net	Yes	**88.97**	**95.82**

Table 8
Comparison of the model on the UTD-MHAD dataset with prior art.

Methods	Multi-model	Top-1 Accuracy
JDM-CNN[62]	Yes	88.10
MCRL[63]	Yes	93.02
PoseMap[61]	Yes	**94.51**
AMFI-Net	Yes	**93.21**

evaluation index of the model performance, which can intuitively express the generalization ability of the model, as shown below:

$$Accuracy = \frac{TP}{Total} \times 100\% \tag{22}$$

where *TP* denotes the number of correctly predicted samples and *Total* denotes the total number of samples.

4.3. Ablation study

The performance of the proposed fusion network framework was validated by performing ablation experiments on cross-view benchmarks on the NTU-RGB + D dataset. The ablation validation was performed, and the best accuracy of the model was recorded in the training cycle for multimodal, fusion programs, weighting factors, and decision modules, respectively.

4.3.1. Impact of multimodality

To verify the effect of different modalities on the accuracy, an exhaustive comparative analysis was conducted to compare the difference in the performance of unimodal and multimodal features in terms of recognition accuracy. From the data in Table 3, it can be clearly observed that the multimodal feature fusion scheme proposed in this paper improved the recognition accuracy by 7.15 % and 2.58 % compared to the single RGB modality and skeleton modality, which significantly improved the recognition accuracy, and the results fully demonstrated the effectiveness of multimodal feature fusion in improving the recognition performance.

For the test samples, the feature vectors $\boldsymbol{Y}_{Fusion}$,$\boldsymbol{X}^{RGB}_{Output}$ and $\boldsymbol{X}^{Skeleton}_{Output}$ of the fusion features, RGB modality and skeleton modality were visualized using *t*-distributed stochastic neighborhood embedding (*t*-SNE), respectively. It can be found that the feature aggregation of the fusion features is better than that of the RGB modality and skeleton modality, and the visualization results are shown in Fig. 8.

Fig. 9 shows the confusion matrices of multimodal feature fusion, RGB modality only and skeleton modality, in which the darker color indicates the higher accuracy. From the confusion matrix, it can be observed that both RGB modalities and skeleton modalities can more accurately recognize actions with significant differences such as “throw”, “stand up” and “sit down”. Skeleton modality can recognize hand movements such as “reading” and “writing,” “playing cell phone” and “tapping keyboard” with high accuracy. The accuracy was not high for subtle actions, while the rich appearance information of RGB modality can better complement the similar actions of skeleton, which can be seen in the multimodal feature fusion confusion matrix for “reading” and “writing”. From the multimodal feature fusion confusion matrix, the correct rate of “reading” and “writing”, “playing cell phone” and “tapping keyboard” were higher than that of individual modes. The multimodal feature fusion proposed in this paper can realize the fusion of two complementary modes better.

4.3.2. Impact of feature fusion programs

To explore the contribution of the feature fusion techniques to the accuracy in this paper, some of the traditional fusion techniques for the fusion of RGB features and skeleton features were also used in this paper, including averaging, summing, multiplying, maximizing, and tandem operations. Table 4 reveals the specific effects of different fusion strategies on the recognition accuracy. According to the data in Table 4, it can be observed that compared with the average, multiplication, maximum, and concaten-ation in the traditional fusion, the recognition accuracy of the proposed fusion scheme in this paper is improved by 1.17 %, 1.3 %, 1.19 % and 0.83 %, respectively, which is a result that fully confirms the superiority of the fusion scheme proposed in this paper over the traditional fusion scheme.

4.3.3. Impact of weighting factors

The network consists of three serial modules used for correlation of skeleton and RGB features, weighting of features, and confidence determination of features, respectively. Two loss functions $L_{fushion}$ and L_{gate} were defined, where $L_{fushion}$ denotes the loss of fusion modal features and L_{gate} denotes the loss of confidence modal confidence. The total loss function was defined as $L_{total} = L_{fushion} + \lambda L_{gate}$, and λ was the weighting factor to regulate the relative importance between the two losses. When λ was set to different values, it slightly affected the overall accuracy. Table 5 shows the effect of different λ settings on the fusion accuracy.

4.3.4. Impact of the decision module

To verify the role played by the decision-making modes in 3.3.3 Partial Feature Confidence, the RGB modes and the skeleton modes were adopted as the decision-making modes to conduct the verification experiments. Table 6 shows the results of the experiments. It can be seen the accuracy of the skeleton modes as the decision-making modes is 7.56 %, which is higher than that of the RGB modes. The appropriate decision-making modes played a significant role in the whole fusion process.

4.3.5. Interpretability analysis

Restricted by the experimental conditions, we validate the model decision logic through theoretical analysis and statistical methods. The top 10 % regions of the calculated RGB modal attention

weights overlap with the labeled key human body parts (e.g., hand, head) by 72 %, indicating that the model effectively focuses on the action-related regions; the average attention weight of the hand joints in the skeleton modal is 0.68, which is also significantly higher than that of the other joints; in the test set, the variance of the skeleton gating weights was only 0.03, indicating robustness to background interference.

4.4. Comparison experiments

In order to verify the validity and sophistication of the methods proposed in this paper, we conducted an exhaustive comparative analysis of the NTU-RGB + D and UTD-MHAD datasets. We selected a variety of current state-of-the-art methods for comparison. As shown in Table 7 and Table 8, on the NTU-RGB + D dataset, the dynamic weight assignment mechanism introduced in this paper is more effective than the simple feature splicing to achieve fusion in Liu et al[58]; the non-local feature association module in this paper enhances the feature representation capability compared to the limitation of global dependency between modalities in De et al[59]; the recognition rate of complex actions is poor in Su et al[60], the fusion strategy in this paper enhances the complex feature representation capability; compared to Li et al[16] AMFI-Net achieves comparable performance with fewer parameters. Similarly, on a small-scale dataset like UTD-MHAD, this paper can still achieve good performance. These comparative analysis results fully demonstrate the superiority of the fusion strategy proposed in this paper, and its accuracy is significantly improved.

5. Conclusion

In this paper, we proposed an innovative multimodal feature integration recognition network leveraging an attention mechanism. During

the feature extraction stage, differential representations of actions from RGB and skeleton data were obtained through multi-level preprocessing and then used for multimodal feature extraction via ResNet3D 18 and AGCN networks. An adaptive fusion strategy was employed to enhance multimodal feature charact-erization. Finally, the confidence level of weighted skeleton information guided the integration of appearance information, optimizing feature fusion. Evaluations on the NTU-RGB + D dataset demonstrated the superiority and practicality of our method compared to current state-of-the-art approaches.

For future work, we aim to explore the application of our approach in real-world scenarios such as security surveillance, fall detection, sports training and judging, as well as dance training and judging.

6. Advantages,Hypothesis, and Limitations

The multimodal fusion framework based on RGB and skeleton proposed in this paper has advantages in the field of behavior recognition: firstly, the remote dependency of cross-modal features is enhanced by the spatio-temporal feature alignment mechanism; secondly, the adaptive ability of the dynamic weight allocation mechanism to the fluctuation of data quality, especially the skeleton data shows good robustness in realistic scenes due to its insensitivity to the background perturbation and illumination changes; finally, the present method has important engineering applications in various fields such as medical and health monitoring, intelligent security, natural human–computer interaction, fall warning and motion posture analysis. However, this study still has deficiencies:

1. Lightweight deployment: Current preprocessing introduces computational redundancy, necessity-yating adaptive keyframe sampling and hardware acceleration., and it is proposed to explore lightweight preprocessing algorithms, such as adaptive keyframe sampling, and investigate hardware acceleration schemes to meet the real-time scene deployment requirements.
2. Limitations of topological modeling capability: Although the performance of multimodal fusion is better than traditional methods, there is still a performance gap in modeling the topological relationship of the human body compared with Graph Neural Network (GNN)[40,41] based on pure skeleton modality.
3. Model structure sensitivity analysis: systematic ablation experiments are planned, including but not limited to (a) backbone network depth comparison; (b) research on spatio-temporal feature extractor alternatives, such as replacing AGCN with GNN variants such as ST-GCN[36], MS-G3D[40], or visual Transformers such as TimeSformer [31]; and (c) evaluating the model's performance in terms of accuracy, computational efficiency and parameter count dimensions of the trade-off relationship.
4. Small Sample Learning Ability Enhancement[49,64,65,66]: to address the problem of scarcity of labeled data in practical applications, it is proposed to explore a cross-domain adaptive approach based on Meta-Learning (ML), combined with generative data augmentation and comparative pre-training strategies, in order to enhance the generalization ability of the model under limited sample conditions.

CRediT authorship contribution statement

Xu Zhao: Writing – original draft. **Chao Tang:** Writing – review & editing, Supervision, Funding acquisition. **Huosheng Hu:** Writing – review & editing. **Wenjian Wang:** Writing – review & editing, Funding acquisition. **Shuo Qiao:** Writing – review & editing. **Anyang Tong:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported in part by the National Natural Science Foundation (62076154, U21A20513), the Provincial Graduate Student Academic Innovation Project (2022xscx145, 2023xscx145, 2024xscx153), the Provincial Graduate Student Innovation and Entrepreneurship Practice Project (2023cxcysj191, 2023cxcysj192, 2024cxcysj201), the Graduate Student Innovative Research Project of Hefei University (2024Ycxsj01, 2024Yxscx01, 2024Yxscx06, 2024Yxscx07), and the Anhui Province University Student Innovation Training Project (202411059041, 202411059044, S202411059123, S202411059135, S202411059151, S202411059169).

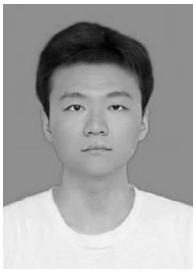
Data availability

Data will be made available on request.

References

[1] E. Hatirnaz, M. Sah, C. Direkoglu, A Novel Framework and Concept-Based semantic search interface for abnormal crowd behaviour analysis in surveillance videos, Multimed. Tools Appl. 79 (2020) 17579–17617.
[2] W. Hu, J. Zhang, B. Huang, W. Zhan, X. Yang, Design of Remote Monitoring System for Limb Rehabilitation Training Based on Action Recognition, in: In Proceedings of 2020 4th International Workshop on Advanced Algorithms and Control Engineering, 2020, pp. 546–555.
[3] A. Lentzas, D. Vrakas, Non-Intrusive human activity recognition and abnormal behavior detection on elderly people: a review, Artif. Intell. Rev. 53 (2020) 1975–2021.
[4] J. An, X. Cheng, Q. Wang, H. Chen, J. Li, S. Li, in: Human Action Recognition Based on Kinect, IOP Publishing, 2020, pp. 0120–0190.
[5] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, J. Liu, Human Action Recognition from Various Data Modalities: A Review, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2022) 3200–3225.
[6] XuXin, YuanXin, WangZheng, ZhangKai, and HuRuimin, Rank-in-Rank Loss for Person Re-identification. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) (2022).
[7] Y. Ma, R. Wang, Relative-position embedding based spatially and temporally decoupled Transformer for action recognition, Pattern Recogn. 145 (2024).
[8] F. Baradel, C. Wolf, J. Mille, Human Action Recognition: Pose-Based Attention Draws Focus to Hands, in: In Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 604–613.
[9] R. Zhao, H. Ali, P. Van der Smagt, Two-stream RNN/CNN for action recognition in 3D videos, in: In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2017, pp. 4260–4267.
[10] M. Zolfaghari, G.L. Oliveira, N. Sedaghat, T. Brox, Chained Multi-Stream Networks Exploiting Pose, Motion, and Appearance for Action Classification and Detection, in: In Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2904–2913.
[11] S. Song, C. Lan, J. Xing, W. Zeng, J. Liu, Skeleton-Indexed Deep Multi-Modal Feature Learning for High Performance Human Action Recognition, in: In Proceedings of the 2018 IEEE International Conference on Multimedia and Expo, 2018, pp. 2934–2946.
[12] W. Liu, X. Jia, X. Zhong, K. Jiang, X. Yu, and M. Ye, Dynamic and static mutual fitting for action recognition. Pattern Recognition.
[13] D. Guo, K. Li, B. Hu, Y. Zhang, M. Wang, Benchmarking Micro-action Recognition: Dataset, Methods, and Applications, IEEE Trans. Circuits Syst. Video Technol. (2024).
[14] X. Wang, S. Zhang, J. Cen, C. Gao, Y. Zhang, D. Zhao, N. Sang, CLIP-guided Prototype Modulating for Few-shot Action Recognition, Int. J. Comput. Vis. 132 (2024).
[15] X. Yuan, X. Xu, X. Wang, K. Zhang, L. Liao, Z. Wang, C.W. Lin, OSAP-Loss: Efficient optimization of average precision via involving samples after positive ones towards remote sensing image retrieval, CAAI Trans. Intell. Technol. 8 (2023).
[16] J. Li, X. Xie, Q. Pan, Y. Cao, Z. Zhao, G. Shi, SGM-Net: Skeleton-Guided Multimodal Network for Action Recognition, Pattern Recogn. 104 (2020).
[17] H. Duan, Y. Zhao, K. Chen, D. Lin, B. Dai, Revisiting Skeleton-Based Action Recognition, in: In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 2969–2978.
[18] K. Simonyan, A. Zisserman, Two-Stream Convolutional Networks for Action Recognition in Videos, Adv. Neural Inf. Proces. Syst. 27 (2014).
[19] C. Feichtenhofer, A. Pinz, A. Zisserman, Convolutional Two-Stream Network Fusion for Video Action Recognition, in: In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1933–1941.
[20] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, Towards good practices for very deep two-stream convnets. computer, Science (2015).

[21] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, A. Rabinovich, Going Deeper with Convolutions, IEEE Computer Society (2014).
[22] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, in: In Proceedings of the 3rd International Conference on Learning Representations, 2014, pp. 7912–7921.
[23] J. Chen, J. Wu, J. Konrad, P. Ishwar, Semi-Coupled Two-Stream Fusion Convnets for Action Recognition at Extremely Low Resolutions, in: In Proceedings of 2017 IEEE Winter Conference on Applications of Computer Vision, 2017, pp. 139–147.
[24] Z. Tu, W. Xie, J. Dauwels, B. Li, J. Yuan, Semantic Cues Enhanced Multimodality Multistream CNN for Action Recognition, IEEE Trans. Circuits Syst. Video Technol. 29 (2018) 1423–1437.
[25] S. Ji, W. Xu, M. Yang, K. Yu, 3D Convolutional Neural Networks for Human Action Recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2012) 221–231.
[26] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning Spatiotemporal Features with 3D Convolutional Networks, in: In Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4489–4497.
[27] Z. Qiu, T. Yao, T. Mei, Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks, in: In Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5533–5541.
[28] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, M. Paluri, A Closer Look at Spatiotemporal Convolutions for Action Recognition, in: In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6450–6459.
[29] K. Hara, H. Kataoka, Y. Satoh, Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and Imagenet?, in: In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6546–6555.
[30] C. Zhang, J. Wu, Y. Li, ActionFormer: localizing moments of actions with transformers, The International Journal of Robotics Research (2022).
[31] G. Bertasius, H. Wang, L. Torresani, Is Space-time attention all you need for video understanding? International Conference on Computer Vision Workshops (2021).
[32] J. Liu, G. Wang, P. Hu, L.-Y. Duan, A.C. Kot, Global Context-Aware Attention LSTM Networks for 3D Action Recognition, in: In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1647–1656.
[33] J. Liu, A. Shahroudy, D. Xu, G. Wang, Spatio-Temporal LSTM with Trust Gates for 3D Human Action Recognition, in: In Proceedings of the 14th European Conference on Computer Vision, 2016, pp. 816–833.
[34] W. Li, L. Wen, M.C. Chang, S.N. Lim, S. Lyu, Adaptive RNN Tree for Large-Scale Human Action Recognition, in: In Proceedings of IEEE International Conference on Computer Vision, 2017, pp. 1453–1461.
[35] X. Jiang, K. Xu, T. Sun, Action Recognition Scheme Based on Skeleton Representation with DS-LSTM Network, in: IEEE Transactions on Circuits and Systems for Video Technology PP, 2019, p. 1.
[36] S. Yan, Y. Xiong, D. Lin, Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, 2018.
[37] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, Q. Tian, Actional-Structural Graph Convolutional Networks for Skeleton-Based Action Recognition, in: In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3595–3603.
[38] L. Shi, Y. Zhang, J. Cheng, H. Lu, Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition, in: In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 12026–12035.
[39] L. Shi, Y. Zhang, J. Cheng, H. Lu, Skeleton-Based Action Recognition with Directed Graph Neural Networks, in: In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7912–7921.
[40] Z. Liu, H. Zhang, Z. Chen, Z. Wang, W. Ouyang, Disentangling and Unifying graph convolutions for skeleton-based action recognition, Neural Comput. & Applic. (2020).
[41] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, W. Hu, Channel-wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition, Multimed. Tools Appl. (2021).
[42] S. Das, S. Sharma, R. Dai, F. Bremond, M. Thonnat, Vpn: Learning Video-Pose Embedding for Activities of Daily Living, in: In Proceedings of the 16th European Conference on Computer Vision, 2020, pp. 72–90.
[43] H.R.V. Joze, A. Shaban, M.L. Iuzzolino, K. Koishida, Mmtm,, Multimodal Transfer Module for CNN Fusion, in: In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 13289–13299.
[44] A. Kendall, M. Grimes, R. Cipolla, Posenet: A Convolutional Network for Real-Time 6-dof Camera Relocalization, in: In Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2938–2946.
[45] G. Moon, H. Kwon, K.M. Lee, and M. Cho, Integralaction: Pose-Driven Feature Integration for Robust Human Action Recognition in Videos, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3339-3348.
[46] S. Das, R. Dai, M. Koperski, L. Minciullo, L. Garattoni, F. Bremond, G. Francesca, Toyota Smarthome: Real-World Activities of Daily Living, in: In Proceedings of the 17th IEEE/CVF International Conference on Computer Vision, 2020, pp. 1478–1489.
[47] X. Weiyao, W. Muqing, Z. Min, X. Ting, Fusion of Skeleton and RGB Features for RGB-D human action recognition, IEEE Sens. J. 21 (2021) 19157–19164.
[48] Y. Sun, W. Xu, X. Yu, J. Gao, T. Xia, Integrating Vision Transformer-Based Bilinear Pooling and Attention Network Fusion of RGB and skeleton features for human action recognition, International Journal of Computational Intelligence Systems 16 (2023).
[49] M.B. Shaikh, D. Chai, I.N. Akhtar, Multimodal fusion for audio-image and video action recognition, Neural Comput. & Applic. 36 (2024) 5499–5513.
[50] J. Yun, D. Jiang, Y. Liu, Y. Sun, B. Tao, J. Kong, J. Tian, X. Tong, M. Xu, Z. Fang, Real-time target detection method based on lightweight convolutional neural network, Front. Bioeng. Biotechnol. 10 (2022) 861–886.
[51] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[52] A. Buades, B. Coll, J.-M. Morel, A Non-Local Algorithm for Image Denoising, in: In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, pp. 60–65.
[53] X. Wang, R. Girshick, A. Gupta, K. He, Non-Local Neural Networks, in: In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.
[54] J. Hu, L. Shen, G. Sun, S. Albanie, Squeeze-and-Excitation Networks, IEEE Trans. Pattern Anal. Mach. Intell. (2017).
[55] J. Chung, C. Gulcehre, K.H. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, J. Franklin Inst. (2014).
[56] A. Shahroudy, J. Liu, T.T. Ng, G. Wang, Ntu rgb+d,, A Large Scale Dataset for 3D Human Activity Analysis, in: In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016, pp. 1010–1019.
[57] C. Chen, Utd-mhad,, A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor, IEEE International Conference on Image Processing (2015).
[58] G. Liu, J. Qian, F. Wen, X. Zhu, R. Ying, P. Liu, Action Recognition Based on 3D Skeleton And RGB Frame Fusion, in: Proceedings of 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2019, pp. 258–264.
[59] A.M. De Boissiere, and R. Noumeir, Infrared and 3D Skeleton Feature Fusion for RGB-D Action Recognition. Institute of Electrical and Electronics Engineers Inc 8 (2020) 168297-168308.
[60] L. Su, C. Hu, G. Li, D. Cao, MSAF: multimodal split attention fusion, Multimed. Tools Appl. (2020).
[61] M. Liu, J. Yuan, Recognizing Human Actions as the Evolution of Pose Estimation Maps, in: In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1159–1168.
[62] C. Li, Y. Hou, P. Wang, W. Li, Joint distance maps based action recognition with convolutional neural networks, IEEE Signal Process Lett. 24 (2017) 624–628.
[63] T. Liu, J. Kong, M. Jiang, RGB-D action recognition using multimodal correlative representation learning model, IEEE Sens. J. (2018).
[64] X. Wang, S. Zhang, Z. Qing, Z. Zuo, C. Gao, R. Jin, and N. Sang, HyRSM++: Hybrid Relation Guided Temporal Set Matching for Few-shot Action Recognition. Pattern Recognition 2024,147:110110 (2024).
[65] L. Wang, and P. Koniusz, Flow Dynamics Correction for Action Recognition, ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2024, pp. 3795-3799.
[66] A.C. Cob-Parro, C. Losada-Gutiérrez, M. Marrón-Romera, A. Gardel-Vicente, I. Bravo-Muoz, A new framework for deep learning video based Human Action Recognition on the edge, Expert Syst. Appl. 238 (2024) 122220.



Xu Zhao was born in 1999. He is currently pursuing his master's degree in the School of Artificial Intelligence and Big Data at Hefei University. His research interests include multimodal human action recognition and machine learning.

Chao Tang received his M.E. degree from the School of Computer and Information Technology, Shanxi University, in 2009, and his Ph.D. degree from the School of Information Science and Technology, Xiamen University, China, in 2014. In 2016, he was a visiting scholar at Bloomfield College, USA. From 2017 to 2018, he was a visiting scholar at the University of Birmingham, U.K. Since December 2018, he has been an associate professor at the School of Artificial Intelligence and Big Data, Hefei University, China. His research interests include machine learning and computer vision.

Huosheng Hu is currently a Professor in the School of Computer Science and Electronic Engineering, University of Essex, UK. His research interests include robotics, human-robot interaction, mechatronics, embedded systems, and cloud computing. He has authored over 500 papers. Prof. Hu is a fellow of the Institution of Engineering and Technology and the Institute of Measurement and Control. He has been a Program Chair for many IEEE international conferences, such as IEEE ICRA, IROS, ICMA, and ROBIO. He is an Executive Editor for the Int. Journal of Mechatronics and Automation, a former Editor-in-Chief for Int. Journal of Automation and Computing and Former Editor-in-Chief for MDPI Robotics Journal.

Wenjian Wang received her Ph.D. degree from Institute for Information and System Science, Xi'an Jiaotong University in 2004. Now she is a full-time professor and Ph.D. supervisor at the Key Laboratory of Computational Intelligence and Chinese Information Processing of the Ministry of Education, Shanxi University. She has published more than 170 academic papers. Her current research interests include machine learning and computational intelligence.

Shuo Qiao was born in 2000. He is currently pursuing his master's degree in the School of Artificial Intelligence and Big Data at Hefei University. His research interests include gait recognition, behavior recognition and machine learning.

Anyang Tong was born in 1998. He received his M. E. degree from the School of Artificial Intelligence and Big Data, Hefei University, China, in 2023. He is currently pursuing a Ph.D. degree at the School of Computer Science and Information Engineering, Hefei University of Technology. His research interests include semi-supervised learning, facial expression recognition, and affective computing.