IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 3 March 2024, accepted 12 March 2024, date of publication 19 March 2024, date of current version 26 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3379449

RESEARCH ARTICLE

A Key Skeleton Points Guided Classroom Action Recognition Method Based on Multimodal Symmetry Fusion

ZEFANG CHEN, YANG GAO, AND QIUYAN YAN
School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China
Corresponding author: Zefang Chen (TS21170041A31@cumt.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 62277046.

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

ABSTRACT Recently, there has been growing interest in utilizing skeleton data for human action recognition due to its compact size and ability to capture action characteristics effectively. However, in complex classroom scenarios, student actions encounter challenges such as high inter-class similarity, differentiation difficulty, and redundancy, which hinder effective differentiation using existing unidirectional feature splicing multimodal methods. Therefore, we propose a key skeleton points guided classroom action recognition method based on multimodal symmetry fusion. This method is primarily characterized by several innovations. Firstly, we utilize a method called Variable Series Mean to select the most significant key skeleton points of actions. Then, these points are input into a model to learn the relevant weight values, guiding the generation of salient regions in RGB images. Finally, in the data fusion stage, we utilize the Symmetric Multi-Modal optimization function to integrate the three data streams, addressing bias issues arising from unidirectional feature splicing methods. We conducted comprehensive experiments on two datasets: NTU 60 and Classroom. Synthesizing results of multiple methods, our method achieves state-of-the-art performance on the NTU 60 dataset and the second-best performance on the private Classroom dataset. Despite not attaining the highest recognition accuracy on the Classroom dataset, this approach offers substantial benefits in terms of time and storage, providing a real-time solution for recognizing student actions in the classroom. Therefore, our method effectively captures and integrates the representation information from different modalities, enabling accurate recognition of student actions in the classroom.

INDEX TERMS Action recognition, multimodal, skeleton data, classroom action.

I. INTRODUCTION

Currently, the common skeleton-based action recognition methods [1] take the 3D coordinates of the skeleton points as the input to the model and mine the spatial information carried by the original skeleton data through the values of the x, y, z three coordinates to characterize the skeleton topology. However, there are problems when such methods are applied to recognize students' actions in real classroom scenarios. Firstly, if only the spatial information of skeleton is considered for action recognition, achieving accurate recognition for similar action segments that occur within a specific timeframe is challenging. For instance, Figure 1 illustrates a series of two actions: standing up and sitting down. As shown in Figure 1, the spatial structures of these two skeleton points are similar in the third and fourth frames of the action sequence. Accurately distinguishing between the two postures based solely on the spatial information of the skeleton points is impossible. Secondly, suppose the time dimension is introduced into the action recognition

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

of skeleton data by characterizing the action changes on the consecutive timestamps. In that case, the difficulty of distinguishing similar actions between classes can be partially solved. Due to limitations in sensor accuracy, the acquired skeleton data may exhibit unstable offset of point coordinates when the observed subjects undergo dynamic changes. This instability hinders the differentiation between genuine changes in skeleton coordinates caused by actions and coordinate instability induced by the device. Lastly, the multimodal methods for RGB data address the issues of inaccurate data and poor stability in unimodal mode. Whereas the conventional multimodal approach only involves simple feature splicing. Additionally, the unidirectional loss function, causing data bias, fails to capture deep semantic information. These significantly affect the action recognition accuracy.

Aiming at the above problems, this paper puts forward the following solution ideas: firstly, the skeleton data (X_{ske}) is represented as two parts: skeleton point data (X_{joints}) and skeleton bone data (X_{bones}), and the temporal and spatial relationship modeling and feature extraction are carried out respectively. At the same time, the extraction of the "key point" method, named the Variable Series Mean (VSM) method, to capture the skeleton points with the most significant change, to preliminarily locate the region of interest of the action of the skeleton data. Secondly, image data is introduced to learn the region of interest of the skeleton data, and the weights of the key skeleton points are used to guide the RGB data to enrich further the features of the region of interest of the action. Finally, the Symmetric Multi-Modal (SMM) loss function is designed to realize the bi-directional fusion of the skeleton data and the RGB data to improve the accuracy of action recognition. The contributions of this paper are as follows:

(1) We propose a novel data selection method, called the Variable Series Mean (VSM) method, which aims to analyze change sequences. Selecting the most representative skeleton points before inputting them into the network enhances computing efficiency for recognizing student actions in the classroom while providing accurate spatiotemporal information for the skeleton data.

(2) We propose a classroom action recognition technique that combines skeleton data and RGB data to enhance the quality of feature representation for both modalities. This technique utilizes graph convolution to capture the key skeleton data points, guiding the generation of the Action Focused Area (AFA) in the corresponding RGB image. Additionally, incorporating the RGB modality, which provides semantic information about the objects surrounding the action, enables accurate recognition of actions with high inter-class similarity at a fine-grained level.

(3) We propose a Symmetric Multi-Modal loss function (SSM). This paper proposes a crossover loss function considering three modalities: skeleton points stream, skeleton bones stream, and RGB stream. The technique aims to address the problem of fusion bias caused by insufficient data in unidirectional fusion to achieve a balanced representation of multiple modalities in action recognition tasks.

FIGURE 1. Example diagram of students' continuous action to stand up and sit down.

II. RELATED WORK

A. VISION-BASED UNIMODAL ACTION RECOGNITION

In the field of computer vision, approximately 80% of the data originates from the visual image modality, considered the most direct and effective way to acquire relevant features and effectively capture valuable information. For the recognition of visual information actions, Convolutional Neural Networks [2], [3], [4], [5], [6], [7], [8], [9], [10], [11] are currently widely utilized. In 2015 Tran [7] proposes the Convolutional 3D architecture (C3D).The same year Tran [8] also combined the residual network with the C3D network and proposed the Res3D network. The Res3D network further enhances the performance of the network, running twice as fast as C3D with half the model size. Wu et al. [9] extends 3D CNN to depth and pose data beyond RGB data by introducing spatio-temporal attention in 3D convolution, visualizing the spatial configuration of the body parts to evaluate its capability for spatio-temporal multimodal learning for video action recognition. Davoodikakhki and Yin [11] introduces a multilayer attention network into a 3D volume framework with hierarchical classification of source datasets, network pruning, and skeleton-based preprocessing to improve the robustness and performance of the model.

With the abundance of vision-based RGB data, near-human recognition accuracy has been achieved using a single RGB modality. However, multi-stream convolutional networks stack convolutional layers repeatedly to obtain a large temporal sensory field, which introduces a large number of parameters, leading to a dramatic increase in memory consumption and computation; at the same time, convolutional networks usually only consider relatively short time intervals and are not able to capture the information of a long span of time.

B. SKELETON-BASED UNIMODAL ACTION RECOGNITION

Graph Convolutional Networks utilize the graph topology of natural connections between skeleton joints to represent 3D spatial relationships, which allows for the natural preservation of skeleton action information [1], [12], [13], [14]. Yan et al. [13] considers that the dynamic information of the human skeleton is crucial for action recognition. In response, he proposed a network model called ST-GCN, which can

effectively capture the implicit spatial and temporal patterns in the data.

Shi et al. [1] focused on the network topology and proposed Two-Stream Adaptive Graph Convolutional Networks (2s-AGCN) in a way that learning can be integrated either uniformly or individually by backpropagation algorithms that take into account first-order and second-order features such as the length and orientation of skeleton. Chen et al. [14] proposes CTR-GCN, a novel topology-optimized graph convolutional network for dynamically learning different topologies and efficiently aggregating joint features across channels. This is achieved by learning a shared topology as a common prior for all channels and refining the channel topology using channel-specific correlations.

The Transformer-based methods [15] do not depend on the human structure and can model the relationship between all points compared with the methods mentioned above. Considering this advantage, Transformer based method [16], [17], [18], [19] is proposed for skeleton action recognition tasks. Plizzari et al. [17] introduces self-attention into graph convolution and uses a spatial and temporal self-attention module to model the correlation of intra-frame and inter-frame joints respectively. Qiu et al. [16] proposes a Spatio-Temporal Tuple Transformer (STTFormer), which can establish the relationship between different joints in consecutive frames and has a relatively strong ability to distinguish similar actions to state-of-the-art results.

However, the existing Transformer-based methods cannot accurately capture the correlation of different skeleton points between frames, which the correlation is beneficial since the extreme similarity of different skeleton points between adjacent frames in a classroom.

C. MULTIMODAL ACTION RECOGNITION

Multimodal data is extensively employed in action recognition because it can fuse data from various sources such as audio, video, and text. This flexibility allows capturing action characteristics from multiple perspectives, providing more prosperous and comprehensive information, and accommodating diverse action recognition scenarios. By fusing multimodal data, the uncertainty inherent in individual data sources can be mitigated, leading to improved accuracy and robustness in recognition. Das et al. [10] designed the VPN, whose two key components are spatial embedding and attention network. The spatial embedding projects 3D poses and RGB cues into a common semantic space, and the attention network provides weighted processing information that enables action recognition frameworks to learn spatiotemporal features better using both modalities. In 2021, Bruce designed a novel multimodal fusion network for indoor scenes called a Teacher-Student Multimodal Fusion model (TSMF) [20] that fuses the skeleton and RGB modalities at the model level for action recognition. In the second year, Bruce proposed a Model-based Multimodal Network(MMNet) [21], which uses the attentional weight values in the skeleton domain to guide the later modal fusion, achieving the best recognition results at that time.

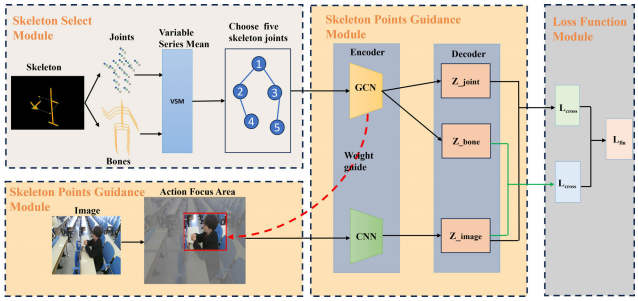

FIGURE 2. Overall framework of the proposed method.

Multimodal approaches are popular because of the complementary information that multiple-modal data can provide. However, most of the above multimodal methods are based on standard datasets for iterative optimization, which dilute the design module for discriminative features in the face of the problems of minor class differences, high inter-class similarity, and single data in the data in complex real classroom scenarios. Even though methods attempt to address the incompleteness of information within a single data domain by incorporating multiple modalities, they often rely on simple feature-splicing methods. These methods limit the exploration of multiple modal features' full potential and fail to consider the potential bias introduced by the unidirectional multiplicative fusion.

III. METHOD

A. OVERALL FRAMEWORK

This chapter mainly focuses on the overall framework of our proposed method in this paper, as shown in Figure 2, which mainly consists of the following modules: 1) Skeleton Select Module; 2) Skeleton Points Guidance Module; 3) Loss Function Module.

B. MODEL INPUTS

The input training sample data of this model, starting from $t = 1$ and ending at time T. The primary purpose is to model and characterize the temporal and spatial relationship between the skeleton frame data acquired during this time interval. The initial input sequence of the model skeleton data is: $X_{ske} = \{x^1_{ske}, x^2_{ske}, \ldots, x^N_{ske}\} \in R^{N\times C\times V\times T}$, where N denotes the batch size, C is the number of channels, V is the number of skeleton points in each frame, and the skeleton data acquired on the Kinect sensors [22] $V = 25$, and T is the number of time frames.

Specifically, the input to the temporal skeleton point is in the form of: $X_{joints} = \{X^i(t)|i = 1, 2, \ldots, V; t = 1, 2, \ldots, T\}$, the value of i is the index value corresponding to the skeleton point, t is the selected time frame. At time t, the ith skeleton point representation is obtained from the Kinect sensor capturing the three-dimensional coordinates of the

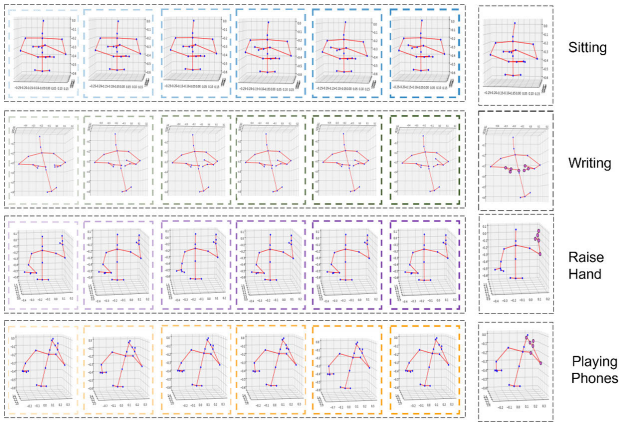

FIGURE 3. Visualization of the largest nodes for different actions and their variations.

classroom student's actions, i.e., $X^i(t) = \left(x_t^i, y_t^i, z_t^i\right) \in R^3$. As for the spatial skeleton bone, the bone sequence is obtained from the skeleton joint points transformation, so the input formula for the bone is: $X_{bones} = \left\{X^i(t) - X^j(t) \,|i, j = 1, 2, \ldots, V; t = 1, 2, \ldots, T\right\}$. Similarly, the i, j refer to the ith and the jth skeleton point, respectively, and t is the selected time frame.

For the visual modality, the RGB data is collected by Kinect at the same time as the sampling of skeleton data. This sampling strategy facilitates the alignment and fusion of feature information extracted from both skeleton and RGB data, allowing key skeleton points to guide RGB generation of action focus areas at any time in the classroom. The specific input data sequence of visual RGB is: $X_{img} = \left\{x_{img}^1, x_{img}^2, \ldots, x_{img}^N\right\} \in R^{N \times C \times H \times W}$, where N indicates the batch size, C is the number of channels, the RGB data acquired on the Kinect sensor $C = 3$, H indicates the height of each image, W is the width of the image.

C. DEFINITION OF KEY SKELETON POINTS DATA

Relying on the statistical generalization study and analysis of the skeleton data of students' actions in the classroom, it is found that the categories of students' actions occurrences are generally concentrated in seven categories, such as raising hands, sitting squarely, sleeping on the table, writing, standing up, sitting down, and playing with cell phones. Further digging deeper into the semantic information carried by actions in the classroom context, this paper finds that the key areas of these action occurrences are focused on the localized position of the students' upper body, as shown in Figure 3. By observing the results of the action visualization and comparing the different action categories to the baseline action (sitting down), We observed that the skeleton points exhibiting the highest rate of change varied among different actions. This indicates that certain important skeleton points contain semantics that influence and determine the characteristics of the discriminative action.

This paper proposes a Variable Series Mean (VSM) method for selecting the key skeleton points. By appling the VSM method, the 25 skeleton points can be further subdivided, and the input data can be localized to the points with the most significant changes in the skeleton data points of the actions, as shown on the right side of Figure 3 (highlighted with a red circle on the far right side).

In Skeleton Select Module, the aim is to use the VSM method to capture the skeleton points in the upper body with the largest variation. Specifically, for the input points data X_{joints} and bone data X_{bones}, lightweight processing of the data is carried out while ensuring accuracy. The output of the skeleton point branch is $X'_{joints} = \{X^i(t) | i = 1, 2, \ldots, K; t = 1, 2, \ldots, T\}$, and the output of the bone branch is $X'_{bones} = \{X^i(t) - X^j(t) | i, j = 1, 2, \ldots, K; t = 1, 2, \ldots, T\}$. Where K is a hyperparameter is the agent task according to the skeleton data input judgment, pick out the amplitude of the action of the largest change in the first points, the experiment to select the first five points, that is $K = 5$. The output X'_{joints}, X'_{bones} is compared with the original input data X_{joints}, X_{bones}, the spatial structure of the skeleton point information and skeleton information after the output selection will not change. The only change is the number of skeleton sampling points, which can remove a large amount of redundant information contained in the original information and, at the same time, can maintain the temporal and spatial information contained in the original skeleton data to a great extent.

D. DEFINITION OF ACTION FOCUS AREAS FOR RGB GENERATION GUIDED BY SKELETON DATA

As shown in Figure 3, after experiencing the data selection, in the Skeleton Points Guidance Module, for the skeleton data stream, this paper adopts the Spatial Temporal Graph Convolution Neural Network framework (ST-GCN) to encode the skeleton data information features. Using a spatial graph structure to represent the skeleton data topology allows for better adaptation to the action. Moreover, the graph's strong ability to handle non-Euclidean data enables the practical mining of spatial and temporal relationships between skeletons. Therefore, applying this approach to classroom student actions recognition aligns well with the data's inherent characteristics. However, the input skeleton point is processed indiscriminately for the existing graph convolution methods, so the model introduces a large amount of redundancy. Meanwhile, no direct operation is available for the existing ST-GCN network to select the key features of the skeleton points. Based on this, we improve the data input style by selecting the most representative skeleton data as the initial region of interest for the ST-GCN network to provide the key information. Unlike the previous work on introducing attentional weights in graph convolutional networks, this paper addresses the problem of graph convolutional networks that cannot "focus on action focused area" and are prone to extracting redundant features.

FIGURE 4. RGB image action focus area generation map.

The specific skeleton modality processing mechanism is as follows: the data output X'_{joints}, X'_{bones} from the VSM module will be input into the ST-GCN, and the graph convolution compiler will be used to extract the feature representation information Z_{fea}, the detailed operation can be divided into two steps, for the skeleton point branch, the input X'_{joints} will generate the feature representation information Z_{joint}, which can be formulated in Equation 1:

$$Z_{joint} = f_{joint}(X'_{joint}) \tag{1}$$

f_{joint} is the encoder for the skeleton point branch in the model of the spatio-temporal graph convolutional network, and Z_{joint} is the feature representation. Similarly, the processing of the skeleton bone branch can be computed using Equation 2:

$$Z_{bone} = f_{bone}(X'_{bone}) \tag{2}$$

In order to efficiently capture key action recognition information and reduce redundant computations, this paper proposes the construction of an Action Focus Area (AFA). The AFA is generated based on the weight guidance of skeleton points data to locate the occurrence area students' actions in images quickly. Figure 4 illustrates the AFA process for generating image guided by skeleton point weights. This method can significantly reduce the data volume of RGB images input by mapping the action occurrence skeleton points information to the corresponding image information, while retaining the semantic information of the action in the image. In other words, the AFA will accurately screen the data as the input to the neural network.

The computational process of AFA is represented by Equation 3. The design concept involves combining the joint weights W_{joints} (calculated as shown in Equation 4, where the magnitude indicates the degree of relevance of a specific skeleton region to the action) learned from the skeleton joint branch using the ST-GCN model with the RGB input data X_{imgs} (i.e., X'_{imgs} obtained through dot-multiplication operation). The weight information obtained from the skeleton points drives the RGB data to construct AFA regions, enabling feature fusion between the skeleton model and the visual RGB modality.

$$X'_{imgs} = X_{imgs} \cdot W \tag{3}$$

$$W_{joints} = \frac{1}{c \cdot t} \sum_{1}^{c} \sum_{1}^{t} \sqrt{\left(X'_{joints}\right)^2} \tag{4}$$

Then X'_{imgs} is input into the backbone network by the ResNet-50 convolutional neural network encoder f_{image} to generate the feature representation information based on the visual image Z_{image}, which is calculated by the Equation 5:

$$Z_{image} = f_{image}(X'_{imgs}) \tag{5}$$

The Z_{joint}, Z_{bone} and Z_{image} obtained from the skeleton point branch, the bone branch, and the visual RGB branch in the encoding stage are fed into the decoding stage. Finally, the decoding results are fed into the design loss function module and the real labeling information in the dataset.

E. SYMMETRIC MULTIMODAL FUSION LOSS FUNCTION

This paper is an improvement strategy tailored to the classroom for the common Cross-Entropy loss (CE) in the field of deep learning. CE is a measure of the distance between the true value of the data y and the model predicts the value of the two probability distributions of $\widehat{y}$, which is calculated in the form of $A \bullet B$, *i.e.* $A \bullet (b_1, b_2, b_3, \ldots b_n)$. The specific calculation steps are as follows: first, the tensor B is disaggregated and then projected onto the space of tensor A to determine the similarity relationship between tensor B and tensor A. The linear transformation from tensor B to tensor A is achieved by considering the changes in coordinate data. This transformation preserves all the feature information in tensor A while discarding the information from tensor B. Consequently, the outcome of model fusion heavily relies on the feature space of tensor A, which may result in insufficient data fusion.

To address the issues above, this paper proposes a novel loss function that incorporates a linear overlap between the product of tensor A and B and the product of B and A. By leveraging symmetric product, this approach aims to mitigate the fusion bias problem that arises from conventional loss functions.

In summary, symmetric multimodal fusion is specified as follows: after the encoder-decoder in obtaining the feature representations Z_{joint}, Z_{bone} and Z_{image} of the skeleton point data, skeleton bone data, and RGB data, the cross-point product operation is performed on the representations to compute the similarity representations, and two intermodal cross-loss functions based on the $L^{cross}_{joint \Leftrightarrow image}$ and the $L^{cross}_{bone \Leftrightarrow image}$ are designed. Specifically, the cross-loss function representation for skeleton points with RGB is shown in Equation 6:

$$L^{cross}_{joint \Leftrightarrow image} = \alpha_1 \cdot \exp(Z_{joint} \cdot Z_{image}) + \beta_1 \cdot \exp(Z_{image} \cdot Z_{joint}) \tag{6}$$

where $L^{cross}_{joint \Leftrightarrow image}$ represents the extracted features on the skeleton point branch and the RGB branch, after going through the compiler to compute the cross-modal feature representation cross-similarity computation. $\alpha_1, \beta_1 \in [0, 1)$ is the correlation coefficient, reflecting the effect of Z_{joint}, Z_{image} and Z_{image}, Z_{joint} on $L^{cross}_{joint \Leftrightarrow image}$ taken from the

skeleton point branch and the RGB branch, respectively. In subsequent experiment, α_1 and β_1 were set to 0.5.

For the similar multimodal information fusion operation done on the bone branch and the RGB branch, the cross-loss function performance is shown in Equation 7:

$$L^{cross}_{bone \Leftrightarrow image} = \alpha_2 \cdot \exp(Z_{bone} \cdot Z_{image}) + \beta_2 \cdot \exp(Z_{image} \cdot Z_{bone}) \quad (7)$$

where $L^{cross}_{bone \Leftrightarrow image}$ represents the information extracted from the bone branch and the RGB branch after compilation and decoding for cross-modal feature characterization cross-similarity calculation. $\alpha_2, \beta_2 \in [0, 1)$ is the correlation coefficient, reflecting the influence of Z_{bone},Z_{image} and Z_{image},Z_{bone} on $L^{cross}_{bone \Leftrightarrow image}$ taken from the bone branch and the RGB branch respectively. In subsequent experiment, α_2 and β_2 were set to 0.5.

Up to this point, the final cross-modal loss function of the model L_{fin} consists of a linear combination of the two cross-modal loss functions described above, which is calculated as shown in Equation 8:

$$L_{fin} = \phi \cdot L^{cross}_{joint \Leftrightarrow image} + \varphi \cdot L^{cross}_{bone \Leftrightarrow image} \quad (8)$$

In order to balance the relationship between different levels of magnitude, we add a weight coefficient for each stage. Where $\phi, \varphi \in (0, 1)$ is a model parameter indicating the correlation coefficient of each type of loss, reflecting the influence of the two cross-comparison loss functions $L^{cross}_{joint \Leftrightarrow image}$ and $L^{cross}_{bone \Leftrightarrow image}$ on the final objective loss function L_{fin}. In subsequent experiment, ϕ, and φ were set to 0.5. The final design objective loss function is back-propagated iteratively optimized under the overall framework of the model.

IV. EXPERIMENTS

In this paper, we utilize a graph convolutional network to capture the skeleton points weight information to guide the RGB generation of the action focused regions (AFA) and compute symmetric multimodal fusion loss function L_{fin}, which is validated on the NTU 60 dataset to verify the scientific validity of the two design solutions proposed in the method. Then, the proposed method is compared with the current method in terms of performance and theoretical analysis on the NTU 60 dataset and Classroom Dataset captured in a real classroom environment.

The experimental sequence is organized as follows: in section A, a brief description of the NTU 60 Dataset and the Classroom Dataset, which mainly includes the collection of student movements, data processing, and movement category archiving. The following section B describes the experimental environment and the related parameter settings. In section C, the proposed method is compared and analyzed with other state-of-the-art methods on the above two datasets. Finally, in section D, the implementation details of the ablation experiment are described in detail, and the results are analyzed and demonstrated scientifically and theoretically.

A. INTRODUCTION TO THE DATASET

1) NTU 60

The dataset is the largest human action recognition dataset proposed by the Rose Lab at Nanyang Technological University (NTU). The dataset contains 60 categories of actions, with a total of 56,880 samples, of which 40 are daily behavioral actions, 9 are health-related actions, and 11 are two-person interaction actions. These actions were performed by 40 subjects aged 10 to 35 by Microsoft Kinect V2 sensors from three different angles, and the data collected were in the form of depth information, 3D skeleton information, RGB frames, and infrared IR sequences. The NTU dataset has two standard evaluation metrics: cross-subject (CS) and cross-view (CV). In the CS experimental setup, the samples will be divided into a training dataset and a test dataset based on the person's ID; in the CV experimental setup, the training dataset and the test dataset will be divided based on the data captured by the camera in different viewpoints.

2) CLASSROOM DATASET

The private Classroom dataset, a Kinect V2 depth camera sensor device developed by Microsoft, was used to record student movement data occurring in the classroom environment to obtain the final output constituting the dataset used in the experiment. The Kinect V2 sensor can acquire RGB images, depth maps, depth maps, skeleton data, and infrared data; this experiment only utilizes skeleton data and RGB data. For the skeleton data collected by the device, 25 frames per second were set, and the collected data were categorized according to the NTU 60 standard. The Classroom dataset collects 7 action categories, namely: raising hands, writing, playing with phones, sitting, sleeping, standing up, and sitting down, based on students' actions in a real classroom environment, as shown in Figure 5. Based on the RGB information and the visualization of skeleton information from the Classroom dataset, it is evident that students primarily engage in upper body action in the classroom. When analyzing the data, inputting all the skeleton information into the network leads to two issues. Firstly, it results in information redundancy. Secondly, the presence of tables and chairs in the classroom affects the action data of the students. Specifically, the posture estimation of the leg skeleton points may exhibit significant deviation, resulting in low experimental result accuracy. To better capture the students' action information in the classroom, this experiment simplifies the 25 skeleton points obtained from Kinect. Unlike NTU 60 dataset, which considers whole body data,it focuses solely on analyzing the skeleton data about the upper body.

B. EXPERIMENTAL SETTINGS

The ST-GCN and C3D networks are used as backbone networks to implement the proposed method using the PyTorch deep learning framework. All experiments are deployed on a device with a GPU of NVIDIA GTX 4090, 32G of graphics memory, and 32G of RAM. Stochastic Gradient

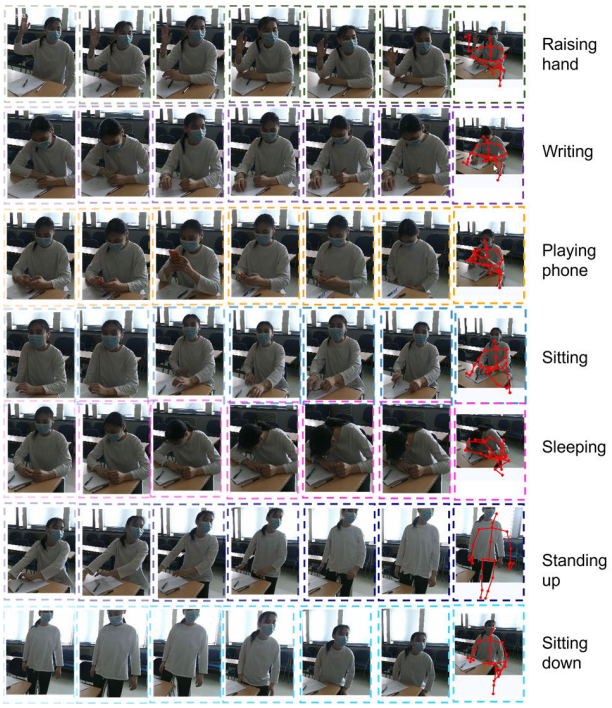

FIGURE 5. Visualization of the seven action categories on the classroom dataset.

TABLE 1. Comparison of recognition accuracy of existing methods on NTU 60 and Classroom dataset.

Method	Time	S	R	N_XSub	N_XView	C_Xsub	C_Xview
ST-GCN	2018	✓	-	81.60%	88.30%	50.21%	54.78%
2s-AGCN	2019	✓	-	88.50%	93.10%	52.45%	57.94%
MS-G3D	2020	✓	-	89.50%	95.37%	53.92%	54.45%
CTR-GCN	2021	✓	-	90.40%	96.20%	59.84%	65.72%
STTFormer	2022	✓		91.20%	96.50%	**62.37%**	**68.51%**
C3D	2015	-	✓	63.56%	70.39%	40.21%	46.38%
I3D	2017	-	✓	77.02%	80.12%	43.17%	48.22%
Glimpose Clouds	2018	-	✓	86.60%	93.25%	46.83%	50.19%
VPN(I3D)	2020	✓	✓	91.50%	96.25%	54.78%	57.62%
TSMF	2021	✓	✓	**92.50%**	96.50%	56.74%	60.45%
MMNet	2022	✓	✓	90.71%	96.78%	58.69%	62.45%
Ours	2023	✓	✓	92.29%	**97.16%**	60.94%	65.22%

Descent (SGD) is used to optimize the parameter settings, and the learning rate is initially set with a weight of 0.0004 to train the model. We use 80 epochs to train all the models and select and save the best model parameter values for subsequent inference.

C. COMPARISONS WITH STATE-OF-THE-ART METHODS PERFORMANCE

In order to verify the validity of our proposed method, we made a comparison experiment between our method and the contemporary unimodal and multimodal methods on the NTU dataset and Classroom dataset; in the table 1, S stands for the input information is skeleton data, R is the input data RGB data, N_XSub and N_XView are the two kinds of evaluation indexes on the NTU dataset, and C_XSub and C_XView are the model's evaluation indexes in the dataset of Classroom. The specific results of the experiments are shown in the Table 1:

Table 1 compares our proposed method with the state-of-the-art methods on the NTU 60 and private Classroom datasets. Comparing multimodal and unimodal results show that the multimodal approach is superior to most of the unimodal approaches. We believe that the input of multimodal data, on the one hand, provides rich semantic information for the model and, simultaneously, provides a complementary role for the information under the single domain. The above experiments can effectively prove the validity of the choice of multimodality for action recognition, thus verifying the correctness of solving students' classroom action recognition from the perspective of a multimodal approach.

We can see that we are still competitive by further validating the method's effectiveness and comparing our proposed method with existing multimodal methods. Our method outperforms VPN by 0.79% and 0.91% in the X-Sub and X-View evaluation schemes on the NTU 60 dataset, respectively, and outperforms our baseline method MMNet by 1.58% and 0.38%, respectively. For the best-performing multimodal method TSMF, our method is 0.21% lower in the X-Sub evaluation metric but 0.66% higher in the X-View metric. Combining the performance of recognition accuracy of all the methods on the NTU dataset, our proposed method achieves the SOTA action recognition accuracy. For our private Classroom dataset, our method outperforms most of the SOTA methods only in the unimodal STTFormer comparisons in the X-Sub and X-View evaluation scenarios, which outperform them by 1.43% and 3.29%, respectively. We analyze that the Transform-based STTFormer method uses powerful contextual relationships to build models that can learn the temporal action sequences of skeleton data, coupled with the fact that our private Classroom dataset has a single class of actions and is acquired in a single time segment. Therefore, STTFormer has a better performance under these conditions. Besides that, our proposed model has better result enhancement than other methods.

D. ABLATION EXPERIMENT

1) MULTIMODAL FEATURE JOINT REPRESENTATION

The experiment consists of two parts, and the specific setup of the experiment operates as follows: the experiment is deployed on two datasets, NTU 60 and Classroom. In Experiment 1, the skeleton data was feature-extracted using a graph convolutional network, while the image information was extracted using a convolutional neural network in a separate branch. Subsequently, the extracted features were concatenated to create the final feature information that was input into the classification network. In Experiment 2, the process begins with the skeleton input information. The skeleton information is then split into skeleton point data and bone data, incorporating temporal and spatial information. Subsequently, both the skeleton point data and bone data undergo feature extraction using graph convolutional networks separately. The skeleton point weight matrix is obtained during the skeleton point data processing.

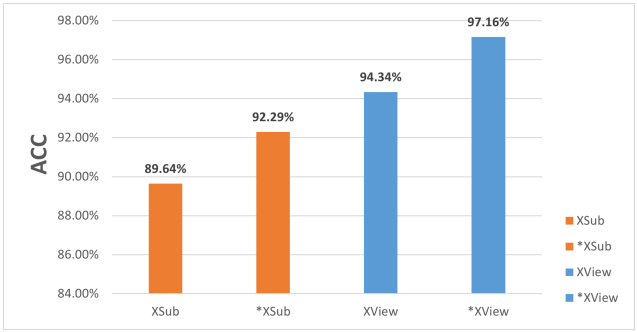

FIGURE 6. W/O AFA recognition accuracy for NTU dataset.

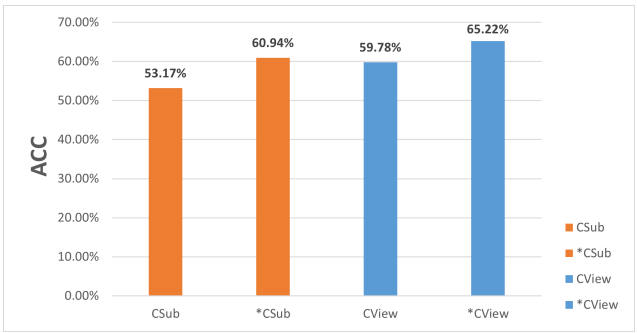

FIGURE 7. W/O AFA recognition accuracy for classroom dataset.

This weight matrix is used to process the complete visual input, integrating the weight information that describes the skeleton point data's importance into the image data's layout. Finally, the AFA, which represents the RGB domain's spatial region that captures the student's movement, is determined. The specific experimental results are shown in Figures 6 and 7.

In Figure 6, XSub and XView are the two accuracy evaluation metrics for the NTU dataset. At the same time, *XSub and *XView are the model accuracy evaluation metrics for the NTU dataset after skeleton guided RGB selection region. Similarly, in Figure 7, CSub and CView are two accuracy evaluation metrics for Classroom dataset. At the same time, *CSub and *CView are model accuracy evaluation metrics for Classroom dataset after skeleton guided RGB selection region.

From the experimental results shown in Figure 6 and Figure 7, we can observe the following findings. In Experiment 1, the multimodal features are fused after extracting information from the respective networks. On the NTU 60 dataset, the accuracy of XSub is 89.64% and XView is 94.34%. However, on the Classroom dataset, the accuracy of CSub is 53.17% and CView is 59.78%. In contrast, Experiment 2 utilizes the learned weighting information from skeleton points to guide the generation of AFAs for RGB images. On the NTU dataset, the accuracy of *XSub is 92.29% and *XView is 97.16%. On the Classroom dataset, the accuracy of *CSub is 60.94% and *CView is 65.22%. Comparing Experiment 1 and Experiment 2, the main difference lies in whether the weight matrix information of the skeleton points is used to guide the selection of the action focus areas in RGB images. Experiment 2 demonstrates improved recognition accuracy on both XSub and XView evaluation metrics, indicating the effectiveness of using skeleton point weights to guide the selection of action focus regions in RGB images.

TABLE 2. Experimental accuracy of inter-modal loss functions.

Number	evaluation	RGB+Skeleton+CE	RGB+Skeleton+SMM
1	N_XSub	89.97%	92.29%
	N_XView	95.28%	97.16%
2	C_XSub	57.45%	60.94%
	C_XView	51.59%	65.22%

TABLE 3. Recognition accuracy in NTU dataset with complete skeleton and VSM operation.

Epoch	Accuracy(ALL)	Accuracy(VSM)
65	80.89%	82.76%
70	81.03%	**82.83%**
75	**81.60%**	80.74%
80	81.27%	82.74%

2) LOSS FUNCTION BASED ON INTER-MODAL

In this module, the validation is carried out on the NTU and Classroom datasets, respectively. In Experiment 1, RGB and Skeleton are input into the model together, and feature extraction is carried out on RGB and Skeleton to form a multimodal information representation. However, the cross-entropy objective loss function(CE) is simply used for function iteration in the final objective function design stage. In Experiment 2, while maintaining the multimodal design concept of Experiment 1, model optimization using Symmetric Multi-Modal loss function (SMM)at the three basins of Joints, Bone, and RGB was also introduced. The specific experimental results are shown in Table 2, where N_XSub and N_XView are the two evaluation metrics on the NTU dataset, and C_XSub and C_XView are the evaluation metrics of the model on the Classroom dataset.

According to Table 2, the introduction of the Symmetric Multi-Modal loss function on the NTU dataset gives the method an improvement of 2.32% and 1.88% on N_XSub and N_XView metrics, respectively. Moreover, there is an improvement of 3.49% and 2.63% on the two evaluation metrics C_XSub and C_XView for Classroom dataset. We analyze that the performance on the NTU dataset is because the NTU dataset is large and complete. The model can achieve the desired effect through multiple iterations at the cost of consuming a large amount of computational space and spending a large amount of time and cost. At the same time, this method has higher requirements on the dataset size and labels. For better performance on the Classroom dataset, we analyze the introduction of the Symmetric Multi-Modal loss function to make up for the lack of a private dataset in scale and accuracy. Comparing the difference in the

TABLE 4. Time and storage after complete skeleton and VSM operation on NTU dataset.

Methods	Times	Memory
Skeleton(ST-GCN)	30h 49min	82.31%
Skeleton(ST-GCN+VSM)	6h 10min	21.43%

objective function between two times, the enhancement of the experimental results side by side verifies that the inter-modal cross entropy function has significant advantages.

3) SELECTION OF KEY SKELETON POINTS IN CHANGING SEQUENCE

Experiment 1: In the unimodal method, ST-GCN was used as the main network framework, and its inputs were all the skeleton data and the key skeleton points data selected based on the Variable Series Mean (VSM), respectively. The recognition accuracies for the two experimental setups are shown in Table 3, and the time and storage are shown in Table 4.

In Tables 3 and 4, we utilize the ST-GCN network as the backbone network and employ two different strategies for skeleton input. One strategy, "ALL" in the table, involves directly inputting the complete skeleton data information. With this approach, we do not perform any skeleton quantity processing. We input all 75 3D coordinate data points obtained from the sensors, corresponding to the 25 skeleton points, into the ST-GCN model. Subsequently, we conduct the superposition of RGB and skeleton modal information finally output the recognition accuracy. The second strategy involves using the VSM strategy. We preprocess the 75 original skeleton point coordinates obtained from the sensors. By applying the principle of mean change, we identify the skeleton point that exhibits the most significant variation during a specific action among the 5 skeleton points. Subsequently, we proceed with the subsequent operations following the same procedure as the first strategy. As can be seen from Table 3, after adding the VSM module, the model reaches the optimal state at the 70th epoch, which is five iterations faster than that of the other methods. There is also a 1.23% improvement in accuracy, which indicates that our method can recognize the movements efficiently and can achieve a high recognition accuracy with a low number of iterations. In this regard, we believe that these selected skeleton points contain key information about the occurrence of this type of action, and such processing, on the one hand, acts as a kind of purification for the skeleton data, eliminating a large amount of redundant information and ensuring a high degree of accuracy of the information input into the model framework. For Table 4, our method has a significant advantage over the original method in terms of time spent and logistics storage, which is almost 20%–25% of the original method, significantly optimizing the time and space complexity of the initial method. We analyze that the VSM reduces the amount of data for input bone points, simplifies the model to process the input bones, and reduces the model's expenditure on time and physical storage. Such a deployment implementation allows subsequent real-time detection and recognition of actions.

TABLE 5. Recognition accuracy after complete data and VSM operation under NTU dataset.

Epoch	Accuracy(ALL)	Memory(ALL)	Accuracy(VSM)	Memory(VSM)
65	89.15%	67.81%	91.62%	21.27%
70	89.48%	68.15%	**92.29%**	21.31%
75	**90.71%**	68.21%	91.66%	21.37%
80	90.05%	68.19%	91.63%	21.35%

TABLE 6. Time, storage after complete data and VSM operation under NTU dataset.

Methods	Times	Memory
Our-VSM	99h 28min	89.32%
Our	61h 19min	21.43%

TABLE 7. Recognition accuracy, time, and storage for different k values under NTU dataset.

k_value	Accuracy	Times	Memory
3	91.31%	64h 34min	21.13%
4	91.73%	69h 21min	21.37%
5	92.29%	71h 19min	21.43%
6	91.82%	72h 01min	27.68%
7	91.16%	72h 25min	30.01%
8	89.84%	75h 03min	33.71%

Experiment 2: In the multimodal method, the proposed method is used with inputs consisting of all the skeleton data and the key skeleton point data selected based on the Variable Series Mean. Recognition accuracies for the two experimental setups are shown in Table 5, and the time and storage are shown in Table 6.

As shown in Table 5, our method achieves a best action recognition accuracy of 90.71% on the NTU dataset. After adding the designed VSM module, the model achieves a higher accuracy of 92.29%. This experimental result demonstrates that the VSM module can effectively improve action recognition accuracy, validating its scientific and practical nature. Considering the storage parameter (Table 6), the original network processes all the 3D coordinates of the 25 skeleton points acquired by Kinect as input during the model learning process. This approach consumes a significant amount of computer storage space, occupying additional physical space during the model iteration process. However, our proposed VSM module effectively addresses this issue by selecting the most representative action skeleton points based on the mathematical arguments discussed in the previous chapter. This approach retains the original and accurate action features while minimizing the waste of physical space caused by redundant data input.

4) SELECTION OF SKELETON POINT K_VALUE FOR MEAN SELECTION OF VARIABLE SEQUENCE

In the experiment, we determine the selection of $k = 5$ skeleton points input to the network in order to verify that the selection of 5 skeleton points can best improve the high efficiency of the information at the same time, minimize

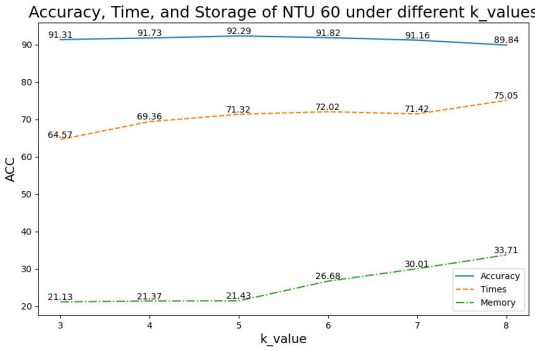

FIGURE 8. Plot of recognition accuracy, time, and storage results for different k_value of NTU dataset.

the skeleton redundant information to improve the speed of computation. The paper in the NTU dataset to take the other values of the case of the accuracy, time, and storage experiments, the specific results of the experiment are shown in Table 7:

According to Table 7 and Figure 8, as the value of k increases, our proposed model slightly increases in action recognition accuracy on the NTU 60 dataset and then shows a decreasing trend. However, it significantly increases overhead time and memory storage costs. For the analysis of the above experimental results, in the process of k_value growth, the model can extract more changes in the amount of skeleton, RGB features and can get a more accurate distinction between the semantic information of the action, which brings a wealth of characterization information to the fine-grained action recognition. However, it is undeniable that as the k_value keeps rising, the clever design of our proposed VSM module loses its effect, which is equivalent to inputting a large amount of skeleton data into the model, bringing redundant information to the model and increasing the amount of computation, which will lead to a decrease in the accuracy of the model. In summary, under the combination of model recognition accuracy, model time overhead, and memory storage share of three considerations, we choose the number of skeleton points selection $k = 5$, under the premise of ensuring the accuracy of the experiment, the most likely choice of the most valuable settings.

V. CONCLUSION

This paper addresses the challenges of accurately identifying student behavior in real classroom environments. In order to overcome the difficulties of distinguishing specific actions and processing redundant information, a key skeleton point guided classroom action recognition method based on multimode symmetric fusion is proposed. This method starts with selecting "key points" skeleton points based on VSM. These key points are representative and provide efficient spatiotemporal skeleton data information. Additionally, multimodal fusion is employed to guide the RGB data using the training weights obtained from the skeleton points data. This fusion process enhances the learning representation of the action focused area, leading to improved recognition performance. Furthermore, the symmetric multimodal fusion loss function is designed to enhance the calculation of a single loss function within the existing methods and balances the optimization problem across multiple modalities.

Overall, the method proposed in this article performs well in improving the accuracy of individual student action recognition. However, the accuracy of identifying classroom group actions could be improved. Meanwhile, since it is difficult to scale up the collection scale of private Classroom dataset, to better identify classroom student action, future research will focus on exploring methods in areas such as multi-view and unsupervised methods. This will help improve the accuracy of overall classroom actions and solve the problem of limited private dataset size, thereby further promoting research and application of classroom student action recognition.

REFERENCES

[1] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7904–7913.

[2] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014.

[3] N. Crasto, P. Weinzaepfel, K. Alahari, and C. Schmid, "MARS: Motion-augmented RGB stream for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7874–7883.

[4] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2016, pp. 20–36.

[5] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[6] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[7] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6450–6459.

[8] D. Tran, J. Ray, Z. Shou, S.-F. Chang, and M. Paluri, "ConvNet architecture search for spatiotemporal feature learning," 2017, *arXiv:1708.05038*.

[9] H. Wu, X. Ma, and Y. Li, "Spatiotemporal multimodal learning with 3D CNNs for video action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1250–1261, Mar. 2021.

[10] S. Das, S. Sharma, R. Dai, F. Brémond, and M. Thonnat, "VPN: Learning video-pose embedding for activities of daily living," in *Proc. Eur. Conf. Comput. Vis.*, Glasgow, U.K. Cham, Switzerland: Springer, 2020, pp. 72–90.

[11] M. Davoodikakhki and K. K. Yin, "Hierarchical action classification with network pruning," in *Proc. 15th Int. Symp. Vis. Comput.* San Diego, CA, USA: Springer, 2020, pp. 291–305.

[12] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, and N. Zheng, "Semantics-guided neural networks for efficient skeleton-based human action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1109–1118.

[13] M. Jiang, J. Dong, D. Ma, J. Sun, J. He, and L. Lang, "Inception spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. Int. Symp. Control Eng. Robot. (ISCER)*, Feb. 2022, pp. 208–213.

[14] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, "Channel-wise topology refinement graph convolution for skeleton-based action recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13339–13348.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
[16] H. Qiu, B. Hou, B. Ren, and X. Zhang, "Spatio-temporal tuples transformer for skeleton-based action recognition," 2022, *arXiv:2201.02849*.
[17] C. Plizzari, M. Cannici, and M. Matteucci, "Skeleton-based action recognition via spatial and temporal transformer networks," *Comput. Vis. Image Understand.*, vols. 208–209, Jul. 2021, Art. no. 103219.
[18] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Decoupled spatial–temporal attention network for skeleton-based action-gesture recognition," in *Proc. Asian Conf. Comput. Vis.*, 2020.
[19] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.
[20] X. Bruce, Y. Liu, and K. C. Chan, "Multimodal fusion via teacher–student network for indoor action recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 4, 2021, pp. 3199–3207.
[21] B. X. B. Yu, Y. Liu, X. Zhang, S.-H. Zhong, and K. C. C. Chan, "MMNet: A model-based multimodal network for human action recognition in RGB-D videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3522–3538, Mar. 2023.
[22] Z. Zhang, "Microsoft Kinect sensor and its effect," *IEEE MultimediaMag.*, vol. 19, no. 2, pp. 4–10, Feb. 2012.

ZEFANG CHEN received the B.S. degree from Nanjing University of Chinese Medicine, Nanjing, China, in 2021. He is currently pursuing the master's degree with China University of Mining and Technology, Xuzhou, China. His research interests include multimodal action recognition and human object interaction detection.

YANG GAO received the bachelor's degree in computer science and technology from China University of Mining and Technology, Xuzhou, China, in 2021, where he is currently pursuing the master's degree. His research interests include human object interaction and multimodal behavior recognition.

QIUYAN YAN received the Ph.D. degree in computer applications from China University of Mining and Technology, in 2010. She is currently a Professor with China University of Mining and Technology. Her current research interests include multimodal image action recognition, big data analytics for education, and temporal data mining.