# Human-Robot Collaboration Through a Multi-Scale Graph Convolution Neural Network With Temporal Attention

Zhaowei Liu, *Member, IEEE*, Xilang Lu, Wenzhe Liu, Wen Qi, *Member, IEEE*, and Hang Su, *Member, IEEE*

***Abstract*—Collaborative robots sensing and understanding the movements and intentions of their human partners are crucial for realizing human-robot collaboration. Human skeleton sequences are widely recognized as a kind of data with great application potential in human action recognition. In this letter, a multi-scale skeleton-based human action recognition network is proposed, which leverages a spatio-temporal attention mechanism. The network achieves high-accuracy human action prediction by aggregating multi-level key point features of the skeleton and applying the spatio-temporal attention mechanism to extract key temporal information features. In addition, a human action skeleton dataset containing eight different categories is collected for a human-robot collaboration task, where the human activity recognition network predicts skeleton sequences from a camera and the collaborating robot makes collaborative actions based on the predicted actions. In this study, the performance of the proposed method is compared with state-of-the-art human action recognition methods and ablation experiments are performed. The results show that the multi-scale spatio-temporal graph convolutional neural network has an action recognition accuracy of 94.16%. The effectiveness of the method is also verified by performing human-robot collaboration experiments on a real robot platform in a laboratory environment.**

***Index Terms*—Human-robot collaboration, intention recognition, skeleton, graph convolutional neural network.**

## I. Introduction

ROBOTICS advancements have made it possible for machines to function with ease in challenging environments, effectively addressing the issue of a labor shortage. However, robots still rely heavily on pre-programming to perform specific tasks, which can limit their flexibility and ability to perform complex actions that humans can easily perform. As the demand for robots to perform delicate and flexible tasks in complex environments continues to grow, collaborative robots are needed to bridge the gap in human-robot interaction. Collaborative robots can help overcome the limitations of pre-programming by working alongside humans and using their cognitive abilities to adapt to changing situations and perform tasks that are difficult for robots to perform on their own.

It is essential for collaborative robots to understand human activity intentions quickly and effectively. During the past few years, many researchers have conducted extensive research on the problem of human intent recognition for human-robot collaboration, and these studies fall into four main categories: biosignal-based human intent recognition method; image-based human intent recognition method; point cloud-based human intent recognition method; and natural language processing (NLP)-based human intent recognition method. Rapetti et al. [1] proposed a collaborative human-robot control method based on human detection with full-body wearable sensors and interaction modeling with coupled rigid-body dynamics, which enables a human and a robot to collaborate in lifting a payload. However, this type of approach requires additional sensors and has limitations in real industrial scenarios. With the advancement of computer vision, scientists have begun to investigate techniques for person recognition based on images. For the purpose of recognizing human activities, Poulose et al. [2] developed a method based on human image thresholding and R-CNN [3], but such approaches are prone to inaccuracy because of the impact of the picture background. Point cloud data can represent the position information of an object in 3D space, thus avoiding misjudgment of pictures due to environmental colors. Yang et al. [4] extracted point cloud data of five movements of human hands and trained the model by point cloud network so as to classify the poses of human hands and realize the task of item transfer between human and robot. However, when human hand recognition is extended to human body recognition, the increase in the amount of point cloud data will occupy a large amount of computer resources and affect the computational efficiency and real-time performance. Language, as a means of human communication, is

Manuscript received 20 October 2023; accepted 29 December 2023. Date of publication 18 January 2024; date of current version 29 January 2024. This letter was recommended for publication by Associate Editor S. Schneider and Editor A. Peer upon evaluation of the reviewers' comments. This work was supported in part by the School and Locality Integration Development Project of Yantai City (2022), in part by the National Nature Science Foundation of China under Grant 62303187, in part by the Fundamental Research Funds for the Central Universities, in part by the Guangdong Provincial Key Laboratory of Human Digital Twin under Grant 2022B1212010004, and in part by the Project of Chunhui Planning of the Ministry of Education under Grant HZKY20220103. *(Corresponding authors: Wen Qi; Hang Su.)*

Zhaowei Liu, Xilang Lu, and Wenzhe Liu are with the School of Computer and Control Engineering, Yantai University, Yantai 264005, China (e-mail: lzw@ytu.edu.cn; luxilang@s.ytu.edu.cn; 202200358006@s.ytu.edu.cn).

Wen Qi is with the School of Future Technology, South China University of Technology, Guangzhou 510641, China (e-mail: wenqi@scut.edu.cn).

Hang Su is with Paris-Saclay University, 91190 Paris, France (e-mail: hang.su@ieee.org).

This letter has supplementary downloadable material available at https://doi.org/10.1109/LRA.2024.3355752, provided by the authors.

Digital Object Identifier 10.1109/LRA.2024.3355752

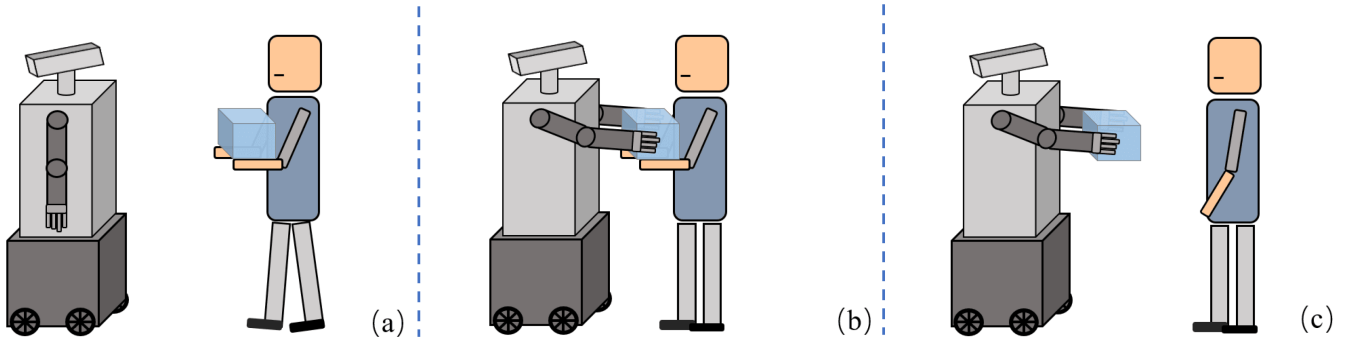


Fig. 1. Object transfer task in human-robot collaboration. (a) When a human holds a box towards the robot, the robot extracts the human skeleton from the images captured by the camera; (b) the human activity recognition model based on the skeleton predicts the human intention and the robot reacts interactively; (c) the robot selects a suitable pose to receive the object according to the pose of the human passing the box.

also used in interactions with robots, especially with the advent of the Transformer [5], which has brought new dynamism to NLP technology. Zinchenko et al. [6] developed a voice control method for an endoscope holder that navigates the endoscope through the operator's voice commands. However, this method has the obvious limitation that it cannot be used in a noisy environment like a factory. Human skeletal data has slowly risen to the forefront of human activity recognition in recent years. By processing and analyzing human skeleton data, motion information such as human motion trajectory, pose, velocity and acceleration can be extracted, thus helping robots to understand human behavioral intentions more accurately. Compared with image data, skeleton data refers to data consisting of key points of the human body (such as joints), which is little affected by the background. Moreover, compared with the data used by other methods, using skeleton data directly as input requires less computation and faster processing. When using neural networks to process skeleton data, key points of the human body can be directly used as input, which will greatly reduce the amount of data that needs to be computed. It can improve the real-time performance and accuracy of the robot. As a result, skeleton data will be more widely used in human-robot collaboration because it will enable robots to better understand human activity and perform collaborative tasks with humans more effectively.

In this letter, a human activity recognition method for human-robot collaboration is proposed. Specifically, in this letter, a skeleton dataset of human-computer interactions in the working environment is collected for training a human action recognition network. Many human actions in human-computer collaboration require cooperative movements between multiple joints, while ordinary deep graph convolutional neural networks only aggregate information in the local range, ignoring the connections between distant nodes. In this letter, we propose a, multi-scale graph convolutional neural network with spatio-temporal attention for the human-robot collaboration, which achieves local feature extraction by aggregating the multi-order neighborhood information of the skeleton keypoints in the spatial dimension and extracts the features of the skeleton in the spatial dimension by using the spatial attention mechanism to achieve high-precision recognition of human activities. The robot reacts to the predictions of the model to achieve human-robot interaction, and Fig. 1 shows an example of a human and robot passing items. The proposed method is compared to other cutting-edge techniques of identical nature on the collected dataset. The experiments illustrate that the proposed method is more accurate. Finally, this letter conducts collaboration experiments with a robot in a laboratory setting to showcase the efficacy of the proposed method.

This letter's primary contributions are:

1) A multi-scale graph convolutional neural network based on temporal attention mechanism is proposed for human-robot collaboration tasks.
2) Collected a human skeleton dataset for network training, containing eight action classes in real scenarios.
3) Experiments confirmed the effectiveness of the proposed human-robot collaboration method on a real robot platform.

## II. Related Works

### A. Human Acitivity Recognition

Deep learning-based human activity recognition (HAR) commonly utilizes various data modalities, including images, skeletons, depth maps, and point cloud [7]. Skeleton data offers valuable structural and pose information about the body and is more robust to clothing texture, environmental changes, viewpoint alterations, and other sources of noise. Consequently, skeleton-based HAR shows vigorous ability. Early skeleton-based HAR mainly obtains the motion states of joints through hand-crafted features [8], [9]. Skeleton-based HAR benefits from the powerful feature learning ability of deep learning, which has seen rapid development in recent years. These methods primarily rely on Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), Graph Neural Network (GNN), Graph Convolutional Neural Network (GCN), and Transformer models. Skeleton data can be naturally transformed into graph structure data. Compared with the methods based on the graph structure, the vector sequence generated by RNN and the image mapping information generated by CNN do not capture well the interconnections between the different joints of the body and the temporal correlation of the same joint. Conversely, the graph structure naturally portrays the relationship between joints. Therefore, GCN has gained more attention in skeleton-based

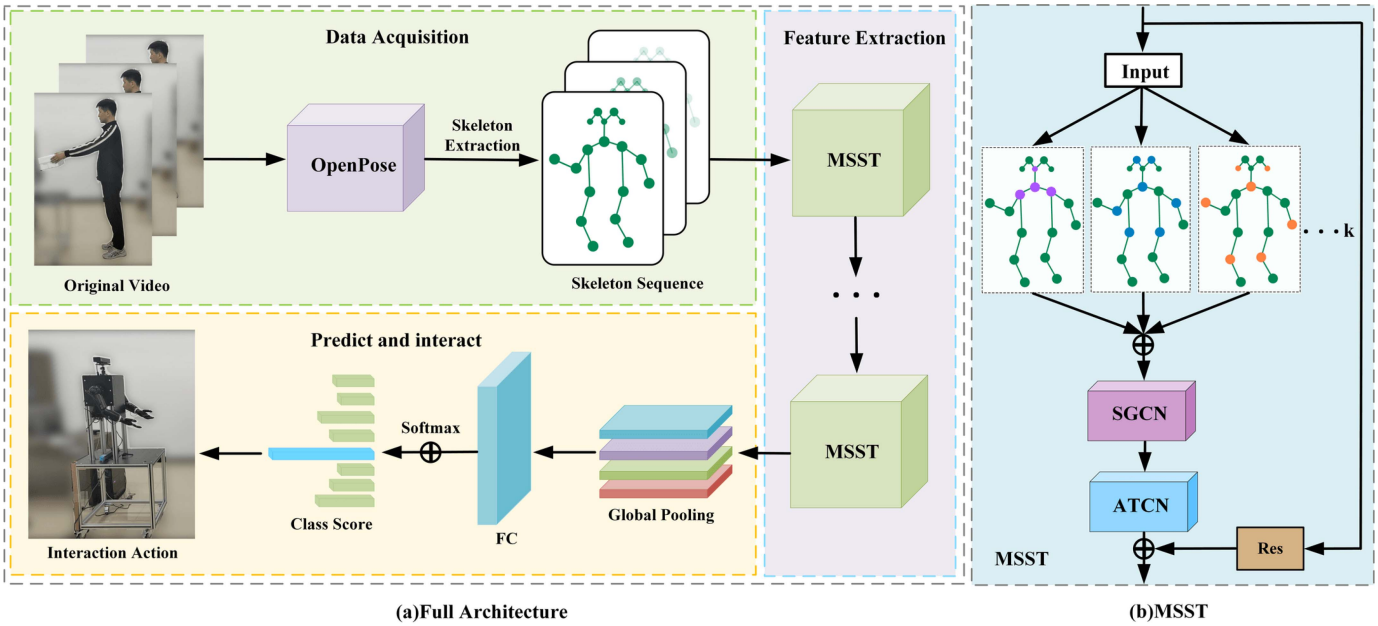


Fig. 2. Multi-scale spatio-temporal graph convolutional neural network. By performing multi-scale feature extraction for each skeleton in the spatial hierarchy, each key point k-order (k = 3 in the figure) neighbor is extracted separately for aggregation, and the aggregated features are fed into SGCN separately for feature extraction, then the features are connected, and the category scores are then output by a standard Softmax classifier.

HAR. Human activity recognition based on GCN usually builds spatio-temporal models to aggregate joint node relationships in time and space [10], [11], [12]. Transformer-based methods often do not depend on the structure of the human body, and instead use spatio-temporal modeling to extract contextual semantic information from spatio-temporal joints [13], [14].

### B. Human-Robot Collaboration With Machine Learning

Fundamental components of human-robot collaboration (HRC) model development involve the perception of human motion trajectory, the recognition of human activity, and the provision of robot feedback [15]. Machine learning enables robots to make decisions autonomously through learning because of its powerful feature extraction and function approximation capabilities. HRC driven by different algorithms plays an important role in industrial, medical, and other fields [16], [17], [18].

Laplaza et al. [19] used contextual information and human intentions to predict human actions based on a multi-head attention structure for interactions between humans and social robots. Using deep learning to model the corresponding scene problem is a good choice. Deep learning models capturing intrinsic rules and hierarchical information between samples have achieved significant results in enhancing robot adaptive motion accuracy [15], [20], [21], [22] and enabling collaborative human-robot handling [23] and robotic surgery [24]. Human-machine collaboration based on deep learning is widely used in industrial, medical, and other industries.

## III. Methodology

Understanding human behavioral intent is crucial for effective human-robot collaboration. This study aims to develop a human-robot interaction framework that uses skeletal human recognition, as shown in Fig. 2. This framework first utilizes OpenPose [25] to extract human bone data from camera data, trains the model to recognize human activities from the bone data, and enables robots to interact based on the recognized human activities to complete collaborative tasks. This section first describes the details of the collection of the dataset and then describes in detail the implementation of the neural network.

### A. Data Collection

The goal of this work is to realize robots assisting their human partners in various object handling tasks, thus reducing the human labor burden, which requires that the recognized actions need to be specific and collaborable. Although human action recognition datasets such as Kinetics [26] exist, the actions in these datasets do not meet the needs of this letter. For this reason, this letter categorizes the actions of handling tasks, and finally selects eight types of actions that have the value of human-robot collaboration, and creates an action dataset, including the actions of humans picking up objects, transferring objects in different postures, and pushing and pulling the handling tool, which are very common in daily work, and can be accomplished by robots collaboratively. Moreover, in order to avoid the robot from appearing when it is no longer needed, human activities that do nothing are included in the dataset. Fig. 3 shows the categories and visualizations of the collected human-robot collaboration skeleton dataset.

The raw data for creating the 2D skeleton is video from various angles, which can be done with just a cell phone, with no order of extra sensors and other equipment. The video collection was performed by five volunteers (four men and one woman) ranging

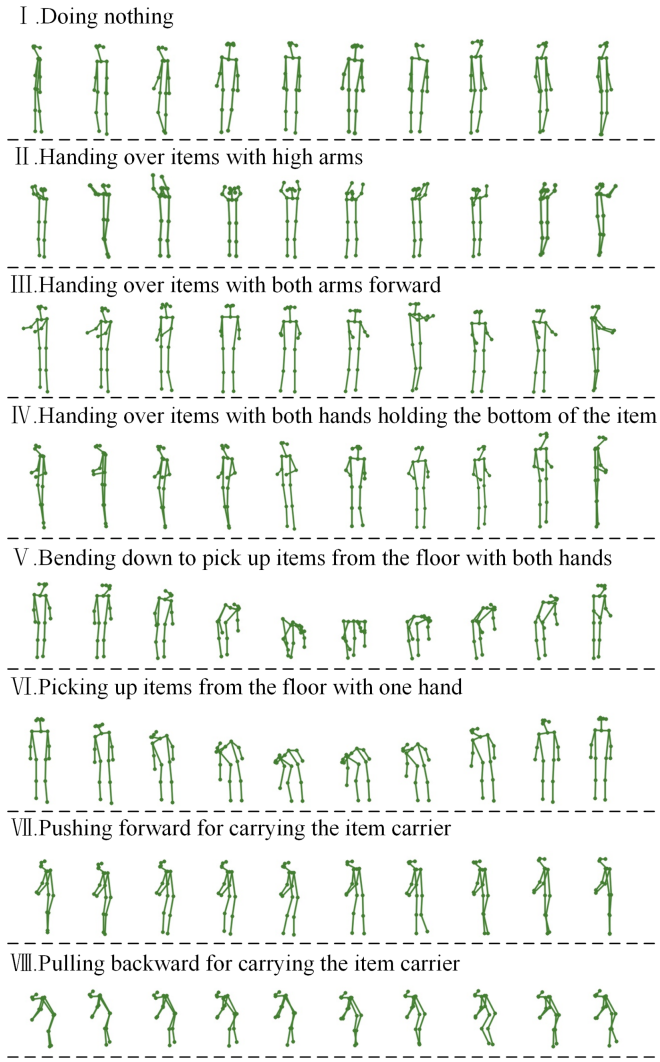


Fig. 3. Example of human-robot collaboration skeleton dataset. Figures (1) to (8) show the eight action categories of the dataset, respectively. Each row is a skeleton extracted from video frames drawn in chronological order from a single video.

in age from 20 to 27 years old, three of whom had experience with robots, all of whom participated voluntarily and were fully informed about the research aspects of this work. The video collection site is an open space in the laboratory. During the collection process, volunteers follow instructions to perform a certain action (such as holding a box with both hands from the ground). Volunteers repeat the process by changing the angles they face (a total of 5 angles: facing the left, facing the camera, facing the right, and rotating 45 degrees from facing the camera to the left and right). The resolution of the acquired videos was reduced to 340×256, and the frame rate was 30 fps. Among these 8 action categories, the first 4 action categories mainly emphasize less differences in arm and body movements, so the volunteers only need to maintain their posture during filming, there can be appropriate moving or turning movements, and all of them have appropriate camera shake and filming position movement during filming. The data was also expanded using mirroring and other methods, resulting in a total of 2400 videos. The human skeleton data for each video was extracted using the Extract Skeleton API provided by OpenPose and labeled with the category of each skeleton, where each human skeleton contains 18 keypoints. 80% of the dataset was used for training and 20% for testing.

### B. Spatio-Temporal Graph Convolution

In this subsection, four aspects of spatial graph convolution, multi-scale aggregation, temporal self-attention mechanism and model implementation are presented. Where the type of graph convolution in this letter is spatial graph convolution [27].

*1) Spatial Graph Convolution:* The skeletal data extracted by openpose is a vector sequence composed of 2D coordinates of each frame of human joint points. Naturally, $G = (V, E)$ is the definition of the skeleton graph, where node set $V = \{v_1, \ldots, v_n\}$ represents the set of all N joint nodes, and bones between joints are represented by undirected edges, forming the edge set $E$. The adjacency matrix of the undirected graph is $A \in R^{n\times n}$, where $A_{i,j} = 1$ if there is an edge between $v_i$ and $v_j$ and 0 otherwise. Human movement is usually accomplished by multiple joint groups, which consist of several independent joints. The skeleton inputs are defined as matrix $X \in R^{T\times N\times C}$, where total input video frame number is represented by $T$, $N$ is the number of joints, and $C$ is the dimension of feature vectors. Then, in order to aggregate the information of neighbor nodes, the layer-wise spatial GCN at time $t$ can be defined as:

$$X_t^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}X_t^{(l)}W^{(l)}\right), \tag{1}$$

where $\tilde{A} = A + I$ is an adjacency matrix with added self-loops for keeping nodes' own features, and the diagonal matrix $\tilde{D}$ is obtained by computing the degree of the node, $\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}$ is a way to normalized $A$, and $W \in R^{C_l\times C_{l+1}}$ is a learnable weight matrix. Feature aggregation of node neighbors is achieved by term $\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}X_t^{(l)}$, the output is then obtained through the activation function $\sigma$.

*2) Multi-Scale Aggregation:* For the above GCN, only the local information around the node is aggregated, and it is difficult to capture the features that are far away from the node. For example, in the action of lifting a box, the body center node should converge more information of the hand nodes, which requires the establishment of further links to obtain this relationship. To create links further afield, nodes' higher-order neighbors are incorporated into the network. Obtaining higher-order neighbor information is implemented by $\tilde{A}_k$, where $\tilde{A}_k = A^k + I$. However, multi-scale aggregation using higher-order polynomials leads to excessive local neighbor weights of nodes due to the large value of low-order neighbors in the higher-order adjacency matrix. To solve this weight bias problem, for higher order adjacency matrices, $A_{i,j}$ greater than 1 is replaced with 1, such that the adjacency matrix is substituted with $\hat{A}_k = 1(\tilde{A}_k \geq 1)$. Therefore, Applying the multi-scale strategy to the (1) transforms as:

$$X_t^{(l+1)} = \sigma\left(\sum_{k=0}^{K}\hat{D}_{(k)}^{-\frac{1}{2}}\hat{A}\hat{D}_{(k)}^{-\frac{1}{2}}X_t^{(l)}W_{(k)}^{(l)}\right), \tag{2}$$

where the scale size is determined by $k$, and $\hat{D}_{(k)}^{-\frac{1}{2}}\hat{A}_{(k)}\hat{D}_{(k)}^{-\frac{1}{2}}$ is normalized k-adjacency, the problem of overweighting of near neighbor nodes is eliminated by (2).

*3) Temporal Self-Attention Mechanism:* Human activity is sensitive to the adjacent position of the same joint, especially the movement of some joints with large amplitude. Traditional temporal graph convolution uses a 2D CNN with kernel size $(\tau, 1)$ to act on the sequence of input and performs feature aggregation on the last $\tau$ frames of each node. For this convolution operation, the temporal relative position information feature cannot be effectively extracted.The self-attention mechanism can be applied to such temporal input sequences. Through this mechanism, the model can autonomously discover the temporal changes of joints to establish the long-range relationship and importance between different frames. Therefore, a temporal self-attention model is proposed to study the temporal motion relationship of each joint. For any joint node $v \in V$, there is a temporal sequence of vectors $s_i \in S = \{s_1, \ldots, s_T\}$, a query $q \in R^{d_q}$, a key $k \in R^{d_k}$, and a value vector $v \in R^{d_v}$. All the above sequences can be adjusted by a learnable linear transformation. For node $v_m$, the importance of the link between frame $i$ and frame $j$ can be evaluated by the dot product $\lambda_{ij}^m = q_i^m \cdot k_j^m \ \forall i, j = 1, \ldots, T$. To obtain the final attention embedding of a node, the value vector $v_j$ of all other nodes is first multiplied by the corresponding evaluation score $\lambda_{ij}$, which is subsequently scaled by the softmax function and the weighted sum is calculated to obtain the attention embedding $z_i^m \in R^{C'}$ of each node, where $C'$ is the number of out channels. This attention embedding is denoted as:

$$z_i^m = \sum_j \text{softmax}_j \left( \frac{\lambda_{ij}^m}{\sqrt{d_k}} \right) v_j^m, \tag{3}$$

where $d_k$ is the dimension of the key sequence, dividing the evaluation score by $\sqrt{d_k}$ is to increase the gradient stability. To achieve better performance, a multi-head attention mechanism is often used, which applies multiple attention and different sets of training parameters to obtain multiple attention embeddings and combines them to obtain the final result.

*4) Model Implementation:* Finally, as shown in Fig. 2, a network for human activity recognition is constructed, which consists of 9 multi-scale spatio-temporal convolution blocks (MSSTs), each containing a multi-scale spatial graph convolutional network SGCN and an attention-convolution network ATCN. For the SGCN, it is achieved by the above spatial method. For the ATCN, the input data firstly is transformed from $\tilde{X} \in R^{B \times C \times T \times V}$ to $\tilde{X} \in R^{BV \times T \times C}$, where B is the batch size of the input, moreover batch size $B$ and the number of joints $V$ are fused into one dimension. And then, it enters the Self-Attention block to obtain the attention embedding. Then, use a 2D convolutional network to perform feature aggregation on each identical node at time $\tau$. In addition, nodes on the body usually perform movements in groups; however, a joint may appear in more than one part while performing an action, and these nodes should receive more attention. Therefore, on the MSST, a learnable spatio-temporal attention mask $M$ is added for making a critical evaluation of each joint, which acts on the multi-scale adjacency matrix $\hat{A}$ through the Hadamard product, then the adjacency matrix is transformed into $\hat{A} \otimes M$, where $\otimes$ is Hadamard product. Through this temporal graph convolution, more temporal location information can be captured. For the first three MSSTs blocks, each block has 64 output channels, the next three layers have 128 output channels, and the last three layers have 256 output channels and temporal frame $\tau = 9$. For each SGCN and ATCN module, we first process the input data with batch normalization and ReLU activation function and add residual links after each MSST module to reduce the problem of gradient vanishing caused by too deep layers. We also add dropout layers with a dropout rate of 0.5 after each MSST module to prevent overfitting. Finally, we use global average pooling and fully connected layers to obtain a tensor with dimensions corresponding to the number of categories to obtain classification results.

Fig. 4. Developed dual-arm robot for human-robot collaboration.

## IV. Experimental Validation

All experiments in this letter were performed on an ubuntu 20.04 computer equipped with an Intel I9-12900 k processor (4.9 GHz), 128 GB RAM and two NVIDIA GeForce 3090 (24 G RAM). The neural network is based on Pytorch deep learning framework [29] using SGD optimizer with cross-entropy as loss function. The model was trained for 50 epochs using a batch size of 32, an initial learning rate of 0.003, and divided by 10 at rounds 20, 30, and 40. A bi-manual robot is established to demonstrate the feasibility of the proposed approach. Each robotic arm are equipped with a 5-degrees of freedom arm and a robotic hand with 6 degrees of freedom. The detailed structural design and the detailed kinematic parameters of the anthropomorphic dual-arm robot, as shown in Fig. 4. Furthermore, the anthropomorphic robot arm is equipped with a dexterous hand in order to perform functions such as object transfer and humanoid hand movements. Hence, the anthropomorphic dual-arm robots are capable of collaborative operation to demonstrate the function of the proposed approach.

### A. Comparison With the State-of-The-Art

The proposed multi-scale spatio-temporal graph convolutional network is evaluated against existing state-of-the-art models for human recognition, using the human skeleton dataset collected in this research letter. The goal is to assess the performance of the system in human-robot collaboration tasks.

TABLE I
COMPARISON OF ACTION RECOGNITION ACCURACY FOR ALL HUMAN ACTION CATEGORIES

| Action Category | Do nothing | Hold high box | Hold for box | Hold flat box | Carry box | Pick up box | Push box | Pull box | All |
|---|---|---|---|---|---|---|---|---|---|
| ST-GCN [10] | 100% | 96.25% | 86.25% | 98.75% | 88.75% | 100% | 41.25% | 86.25% | 87.18% |
| 2S-AGCN [11] | 100% | 90% | 79.17% | 91.67% | 89.17% | 91.67% | 37.5% | 60.83% | 81.25% |
| NAS-GCN [28] | 100% | 100% | 96.67% | 98.33% | 74.17% | 94.17% | 25% | 82.5% | 83.85% |
| MS-G3D [12] | 100% | 100% | 99.17% | 100% | 91.67% | 100% | 81.67% | 79.17% | 93.96% |
| ST-TR [13] | 91.25% | 72.5% | 95% | 76.25% | 95% | 27.5% | 40% | 58.75% | 69.53% |
| MS-ST(Ours) | 100% | 100% | 100% | 100% | 92.5% | 95% | 70.6% | 95.2% | 94.16% |

Firstly, the recognition of various actions by different networks is analyzed. Table I shows the recognition accuracies for the 8 actions and the whole action dataset. The first model is ST-GCN, which has a good accuracy of 87.18% although it is the earliest skeleton-based human recognition network to appear. The main problem is that the network cannot distinguish between pushing and pulling boxes, which is due to the fact that ST-GCN uses only a 2D convolution for feature extraction in the time dimension, which is not enough to extract useful information in the time dimension. 2SAGCN introduces an encoder-decoder structure and extends the existing skeleton graph to improve the network accuracy. NAN-GCN solves the problem of capturing higher order graph node relationships with the help of automatic network search (NAS) methods, but the accuracy rates of both were not high enough, 81.25% and 83.85% respectively, which may be due to the fact that the network uses the idea of dual streams, capturing both joints and bones as inputs to extract more information, but the dataset in this letter is only about node information not contain the bone information. The MS-G3D model uses the multi-order adjacency key point graph of the structure and performs multi-scale aggregation to obtain more order of features and finally obtains higher accuracy, 93.96%, but its network is too large to be suitable for real-time human-machine collaboration tasks. The ST-TR targets the motion of joint parts and uses a space-time transformation network to improve the network accuracy. A spatio-temporal transformation network is used, based on the Transformer-based model, and an attention mechanism in time and space is proposed to extract features for efficient feature extraction. However, the network has the lowest accuracy of 69.53%, mainly because ST-TR addresses the problem of efficiently encoding the underlying information under the 3D skeleton, especially extracting effective information from the joint motion patterns and their correlations. However, the dataset in this letter is a 2D skeleton and all motions do not contain evasive joint motions, leading to its poor performance. The last one is the MS-ST proposed in this letter, and it can be seen from the experimental results that the proposed model achieves 100% accuracy in all the movements except the three types of predictions of pulling the box, picking up the box and carrying the box, and has an overall accuracy of 94.16%, which is better than the other models in terms of performance. Fig. 5 demonstrates the confusion matrix of the prediction results of the proposed method, and it can be seen that the prediction errors still mainly occur between pull and push. In order to minimize the error, the above method was trained and tested multiple times in the experiment.

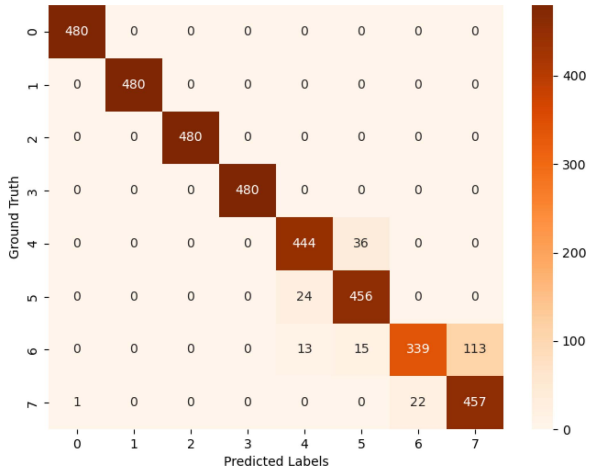


Fig. 5. Confusion matrix of test results for the proposed method.

## B. Ablation Study

This section validates the effectiveness of the proposed multi-scale temporal convolution and temporal self-attention mechanisms. To verify the improvement of the two modules on the accuracy of human activity recognition separately, ablation experiments are set up to compare the network performance of the baseline network, the baseline network with the temporal attention mechanism module and the multiscale graph convolution module of different scales, respectively. The baseline network is obtained by st-gcn modification, which does not have multi-scale convolution operation, i.e., scale $k = 1$.

*Multi-scale aggregation:* First, the effectiveness of the multi-scale spatio-temporal convolution module is verified. The contribution of multi-scale key point aggregation to accuracy is verified by varying the number of scales k at multiple scales. Plot 1 to plot 4 in Fig. 6 show the accuracy curves for different scales k. Compared with that at $k = 1$, the accuracy of the model significantly improves to about 89.64% at $k = 2$ and continues to improve with increasing scale, with accuracy up to about 94.16% at scale $k = 8$. The gain of increasing the scale on the loss can also be seen in the loss curve in the 5th plot in Fig. 6, and the gain is not obvious when $k > 5$. Table II shows the model accuracy of multi-scale modules of different scales, with 94.16% accuracy when both have the time-attention mechanism module and scale $k = 8$.

*Temporal Self-Attention Mechanism:* For the temporal self-attention mechanism module used for validation, with the addition of the multiscale spatio-temporal convolution module, it mainly extends the information aggregation of key points of the

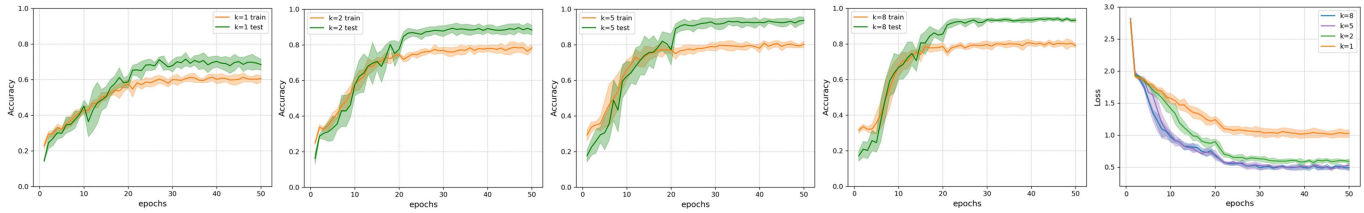


Fig. 6. Accuracy curves and loss curves for different multi-scale skeleton aggregation scales. The first four plots show the accuracy curves and standard deviations for scales 1, 2, 5 and 8, respectively, where orange is the training curve and green is the test curve, and each curve is averaged over six training sessions. The fifth plot shows the loss curves for the network training at different scales.

TABLE II
ABLATION EXPERIMENTS

| Methods | TAM | MSM | Acc |
|---|---|---|---|
| MS-ST | | | 67.71% |
| MS-ST | ✓ | | 70.11% |
| MS-ST | ✓ | ✓(k=2) | 89.64% |
| MS-ST | ✓ | ✓(k=5) | 93.69% |
| MS-ST | ✓ | ✓(k=8) | 94.16% |

skeleton in space and improves in accuracy, but the accuracy does not improve for actions that are similar and different only in the temporal direction (e.g., pushing a box and pulling a box, which are not different in space but opposite in time (actions)). By comparing these actions, it can be seen that the temporal self-attention mechanism module extracts the temporal relationships on the skeleton sequence well and improves the accuracy significantly. In the dataset provided in this letter, each category accounted for 12.5% of the total, whereas in the network without the addition of this module, the actions of pushing and pulling box were not well recognized and resulted in a lower overall accuracy. Table II shows the effect of the temporal attention mechanism on the model, and it can be seen that without the addition of TAM, the accuracy of the model is only 67.71%, while when TAM is added, the accuracy of the model is: 70.11%, where the contribution of the temporal attention mechanism module to the accuracy is mainly due to distinguishing out the two actions of push and pull. Thus, the temporal attention mechanism has significant applications in this model, especially for actions that exhibit spatial similarity and temporal opposition.

### C. Human-Robot Cooperation Based on Human Activity Recognition

To verify the effectiveness and correctness of multi-scale spatio-temporal graph neural networks in human activity recognition, we applied the algorithm to a robotic system and conducted human-robot collaboration experiments. The experiments simulate a scenario in which a robot and a human collaborate to carry an object in an environment such as a warehouse, in order to verify whether the collaborative robot can accurately perceive the movements of its human partner and make correct collaborative responses. We used an RGB camera for real-time image acquisition, and then transmitted the images to OpenPose

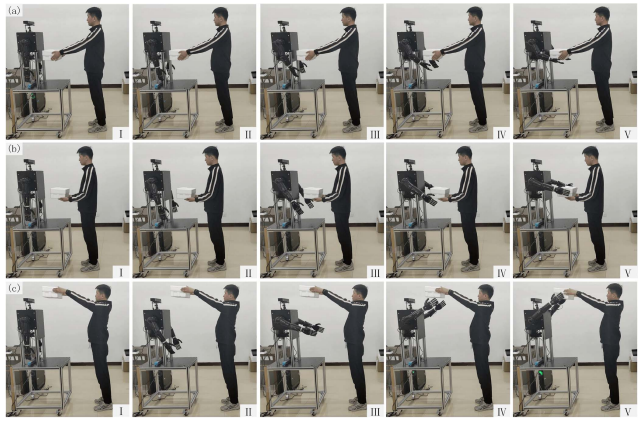


Fig. 7. Human-robot interaction experiment in which the operator passes the box to the robot in different ways and the robot recognizes the operator's action and makes decisions on how to receive the box.

to extract human skeleton sequences. Then, our human action recognition model predicts human actions based on this skeleton data. The task of the experiment is object transfer between the operator and the robot. In the task, the operator passes the box in different poses, and the robot, after recognizing the human's actions, uses our special interpolation-based trajectory planning algorithm to achieve the collaborative behavior. We define a grasping target position for each action. Once the camera detects a human action, the system uses an interpolation-based approach to dynamically plan the robot's trajectory. The human-robot collaboration process is shown in Fig. 7.

The figure shows the passing of the box in three passes, namely: (a) The operator passes the box forward by holding the sides of the box with both hands. (b) The operator holds the box from underneath and passes it. (c) The operator passes the box with his arms raised. The robot then makes a pair of corresponding interaction actions. The experiments show that the method can correctly perform the human-robot collaboration task. The video of human-robot collaboration experiments are available at https://youtu.be/ryBdcYr0Aog.

## V. CONCLUSION

This study presents a multi-scale skeleton-based human activity recognition method for human-robot collaboration tasks. Considering practical application scenarios, a skeleton sequence

dataset for human activity recognition is collected, which contains skeleton sequences of eight interactive actions. To accomplish accurate recognition of human activity, a multi-scale graph convolutional neural network is utilized to extract multi-order neighborhood point features of skeletal key points, and a temporal attention mechanism is included to increase feature extraction of temporal graphs. In this study, the proposed algorithm was evaluated, and the outcomes indicate that the proposed multi-scale graph convolutional neural network attains a recognition accuracy of 94.16% on this dataset, surpassing other analogous approaches. Finally, this letter deploys the recognition algorithm to a robotic system and demonstrates the implementation of a human-robot collaboration task in a real-world environment through a human recognition model. Future works will adopt adaptive nonliner control solutions [30] to achieve stable control of complex robotic systems.

## REFERENCES


[1] L. Rapetti et al., "A control approach for human-robot ergonomic payload lifting," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2023, pp. 7504–7510.

[2] A. Poulose et al., "HIT HAR: Human image threshing machine for human activity recognition using deep learning models," *Comput. Intell. Neurosci.*, vol. 2022, 2022, Art. no. 1808990.

[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.

[4] W. Yang, C. Paxton, M. Cakmak, and D. Fox, "Human grasp classification for reactive human-to-robot handovers," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 11123–11130.

[5] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017.

[6] K. Zinchenko, C.-Y. Wu, and K.-T. Song, "A study on speech recognition control for a surgical robot," *IEEE Trans. Ind. Informat.*, vol. 13, no. 2, pp. 607–615, Apr. 2017.

[7] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3200–3225, Mar. 2023.

[8] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1110–1118.

[9] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3D skeletons as points in a lie group," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 588–595.

[10] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 7444–7452.

[11] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12026–12035.

[12] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 143–152.

[13] C. Plizzari, M. Cannici, and M. Matteucci, "Skeleton-based action recognition via spatial and temporal transformer networks," *Comput. Vis. Image Understanding*, vol. 208, 2021, Art. no. 103219.

[14] Z. Gao et al., "Focal and global spatial-temporal transformer for skeleton-based action recognition," in *Proc. Asian Conf. Comput. Vis.*, 2022, pp. 382–398.

[15] J. Zhang, H. Liu, Q. Chang, L. Wang, and R. X. Gao, "Recurrent neural network for motion trajectory prediction in human-robot collaborative assembly," *CIRP Ann.*, vol. 69, no. 1, pp. 9–12, 2020.

[16] C. Messeri, G. Masotti, A. M. Zanchettin, and P. Rocco, "Human-robot collaboration: Optimizing stress and productivity based on game theory," *IEEE Robot. Automat. Lett.*, vol. 6, no. 4, pp. 8061–8068, Oct. 2021.

[17] Z. Wang et al., "Vision-based calibration of dual RCM-based robot arms in human-robot collaborative minimally invasive surgery," *IEEE Robot. Automat. Lett.*, vol. 3, no. 2, pp. 672–679, Apr. 2018.

[18] H. Su, C. Yang, G. Ferrigno, and E. De Momi, "Improved human–robot collaborative control of redundant robot for teleoperated minimally invasive surgery," *IEEE Robot. Automat. Lett.*, vol. 4, no. 2, pp. 1447–1453, Apr. 2019.

[19] J. Laplaza, F. Moreno-Noguer, and A. Sanfeliu, "Context and intention aware 3D human body motion prediction using an attention deep learning model in handover tasks," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2022, pp. 4743–4748.

[20] W. Qi, S. E. Ovur, Z. Li, A. Marzullo, and R. Song, "Multi-sensor guided hand gesture recognition for a teleoperated robot using a recurrent neural network," *IEEE Robot. Automat. Lett.*, vol. 6, no. 3, pp. 6039–6045, Jul. 2021.

[21] A. Ghadirzadeh, X. Chen, W. Yin, Z. Yi, M. Björkman, and D. Kragic, "Human-centered collaborative robots with deep reinforcement learning," *IEEE Robot. Automat. Lett.*, vol. 6, no. 2, pp. 566–571, Apr. 2021.

[22] Y. Wang, Z. Liu, J. Xu, and W. Yan, "Heterogeneous network representation learning approach for ethereum identity identification," *IEEE Trans. Comput. Social Syst.*, vol. 10, no. 3, pp. 890–899, Jun. 2023.

[23] L. van der Spaa, M. Gienger, T. Bates, and J. Kober, "Predicting and optimizing ergonomics in physical human-robot cooperation tasks," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 1799–1805.

[24] H. Su, W. Qi, C. Yang, J. Sandoval, G. Ferrigno, and E. De Momi, "Deep neural network approach in robot tool dynamics identification for bilateral teleoperation," *IEEE Robot. Automat. Lett.*, vol. 5, no. 2, pp. 2943–2949, Apr. 2020.

[25] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7291–7299.

[26] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6299–6308.

[27] M. Niepert, M. Ahmed, and K. Kutzkov, "Learning convolutional neural networks for graphs," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 2014–2023.

[28] W. Peng, X. Hong, H. Chen, and G. Zhao, "Learning graph convolutional network for skeleton-based human action recognition by neural searching," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 2669–2676.

[29] A. Paszke et al., "Automatic differentiation in Pytorch," in *Proc. NeurIPS Autodiff Workshop*, 2017, pp. 1–4.

[30] J. Zhao and Y. Lv, "Output-feedback robust tracking control of uncertain systems via adaptive learning," *Int. J. Control, Automat. Syst.*, vol. 21, no. 4, pp. 1108–1118, 2023.