International Journal of Computational Intelligence Systems (2023) 16:173
https://doi.org/10.1007/s44196-023-00351-1

RESEARCH ARTICLE

Recognition Method with Deep Contrastive Learning and Improved Transformer for 3D Human Motion Pose

Datian Liu[1] · Haitao Yang[2] · Zhang Lei[3]

Received: 25 April 2023 / Accepted: 18 October 2023

Abstract
Three-dimensional (3D) human pose recognition techniques based on spatial data have gained attention. However, existing models and algorithms fail to achieve desired precision. We propose a 3D human motion pose recognition method using deep contrastive learning and an improved Transformer. The improved Transformer removes noise between human motion RGB and depth images, addressing orientation correlation in 3D models. Two-dimensional (2D) pose features are extracted from de-noised RGB images using a kernel generation module in a graph convolutional network (GCN). Depth features are extracted from de-noised depth images. The 2D pose features and depth features are fused using a regression module in the GCN to obtain 3D pose recognition results. The results demonstrate that the proposed method captures RGB and depth images, achieving high recognition accuracy and fast speed. The proposed method demonstrates good accuracy in 3D human motion pose recognition.

Keywords Pose recognition · Three-dimensional human motion · Deep contrastive learning · Improved transformer · Depth image · Pose feature

1 Introduction

Human pose recognition, particularly in the context of video comprehension, is a crucial area of research in computer vision [1, 2]. The aim of human pose recognition is to enable computers to comprehend and recognize human motions in videos, make predictions and identifications accordingly, and ultimately exhibit more intelligent behavior [3]. Currently, three-dimensional (3D) human pose recognition techniques based on 3D space are gaining increasing importance. However, due to the inherent complexity of this domain, researchers face significant challenges in achieving the desired detection results, especially when applying these findings to real-world scenarios. Despite the existence of various models and algorithms [4, 5] that have enhanced detection accuracy, the expected results remain elusive. Consequently, more accurate 3D spatial positioning within 3D space has become a prominent and intricate research topic [6, 7]. To conduct 3D human pose recognition, RGB images and depth images can be captured using a Kinect camera to record the motion process of humans. These images or videos of human subjects provide the necessary data for subsequent tasks, such as predicting and recognizing human motions [8, 9]. While existing models and algorithms have yielded some improvements in detection accuracy, scholars continue to conduct further research to enhance the precision of 3D human pose recognition. Jiang et al. [10] used a Gaussian mixture model to detect foreground in human motion images, and employed the stripe flow acceleration model to extract features of human motion pose images within the foreground detection results. The extracted features were then input into a feed-forward neural network to output 3D human motion pose recognition results. The accuracy

✉ Zhang Lei
zhanglei2014@cupes.edu.cn

Datian Liu
liudatian@pe.neu.edu.cn

Haitao Yang
yanghaitao@bjut.edu.cn

[1] Physical Education Department, Northeastern University, Shenyang 110819, China

[2] Physical Education Department, Beijing University of Technology, Beijing 100124, China

[3] Institute of Physical Education and Training, Capital University of Physical Education and Sports, Beijing 100091, China

Published online: 31 October 2023

of this method for 3D human motion pose recognition is relatively high, with a recognition error of 2.5%. However, this method does not have de-noising functionality and is susceptible to noise interference, which reduces recognition accuracy. Zhao et al. [11] used a second-order average pooling feature aggregation algorithm to extract features in human motion images and extracted sequence features through a spatiotemporal attention factor-weighted feature aggregation algorithm. The two extracted features were fused in a fusion manner and input into a full-channel spatiotemporal attention factor generating network to output 3D human motion pose recognition results. The accuracy of this method for 3D human motion pose recognition is high. However, this method does not have de-noising functionality, which is susceptible to noise interference, and the feature extraction process is relatively complicated, which reduces recognition accuracy and efficiency. Deng and Wu [12] provided a detailed description of the feature data extraction process in multi-pose human motion scenes, as well as a feature selection method based on multi-information fusion. Furthermore, they proposed a pose correction algorithm based on the detection of human motion coherence, which helps address the problem of misjudgment in human motion pose. In the validation process, multiple classifiers were used, and various data, such as the number of selected features, recognition accuracy, and correction results, were examined. However, the method does not possess feature extraction capabilities and can only roughly recognize human pose. It is unable to recognize detailed information about human pose, thereby reducing the recognition accuracy of pose recognition. Ma and Yan [13] proposed a basketball action pose estimation algorithm based on multi-scale spatiotemporal correlation features. This algorithm utilizes a Transformer-based human temporal feature capture module to model spatiotemporal features at the sequence level, thereby mitigating the negative effects caused by motion blur and occlusion. Additionally, to address the complexity and variability of human shapes, a deformable convolution-based human spatial feature residual fusion module is adopted to obtain more comprehensive spatial features. However, the method is susceptible to noise, resulting in unsatisfactory recognition results. Zhang et al. [14] used hardware equipment based on a two-dimensional (2D) hetero-structure of the retina to collect human motion images, combined with a frame difference algorithm to extract human motion targets in the image, and input them into a conductance mapping neural network to output 3D human motion pose recognition results. This method can effectively collect human motion images and reduce the error of 3D human motion pose recognition. However, this method cannot collect the depth information of human motion images comprehensively, resulting in relatively poor information collection completeness, which affects pose recognition effectiveness.

Despite some progress in previous research on 3D human pose recognition, there are still limitations, such as noise interference, orientation dependencies, and complex feature extraction, which reduce the accuracy and efficiency of recognition. To address these issues, we propose a 3D human pose estimation method based on deep contrastive learning and improved Transformer models, aiming to improve the accuracy and efficiency of recognition. An improved Transformer is used to de-noise human motion images. The improved Transformer can better refine the detailed information of the image, resulting in superior de-noising effect. To improve the recognition accuracy of pose recognition, deep contrastive learning is used to extract human pose features, which effectively improves the recognition accuracy and generalization performance of feature extraction by addressing the problem of difficult sample selection. To further improve the accuracy of pose recognition, a Kinect camera is used to capture RGB images of 2D human motion and depth images of human motion, effectively collecting depth information of human motion images for subsequent pose recognition. The main contributions of this paper are as follows: (1) We propose the recognition method of 3D human motion pose with deep contrastive learning and improved transformer to improve the recognition effect of 3D human motion pose. (2) Using the improved Transformer to remove the internal noise of human motion RGB and depth images. (3) Extracting 2D human motion pose features in the de-noised human motion RGB images using deep contrastive learning; extracting the depth features of human motion pose in the de-noised human motion depth images using the kernel generation module in the graph convolutional networks (GCN). (4) Based on the fusion processing of the regression module in the GCN, the 2D human motion pose features and depth features are extracted to obtain the result of 3D human motion pose recognition.

2 Methodology

2.1 Overview of the Proposed Method

The traditional Transformer model relies solely on encoding and decoding the input image sequences, which limits its ability to handle noise and local errors in the input data. This may result in the model's inability to accurately recover and remove noise, leading to inaccurate results in human motion pose estimation tasks. Self-supervised learning is a learning method that does not require manually annotated data. By designing appropriate self-supervised tasks and introducing noise or randomness during the training process, it simulates noise and variations in real-world scenarios. By

learning meaningful feature representations from unannotated data, the model can better understand noise and errors and improve its ability to handle them. Through this approach, the Transformer model can enhance its ability to handle noisy data, thereby improving the accuracy of motion estimation. Pixel un-shuffle and pixel shuffle techniques are then used to improve the model's recognition and understanding of noisy images, enhancing the accuracy and robustness of human motion image noise processing. Therefore, 3D human motion pose recognition is accomplished

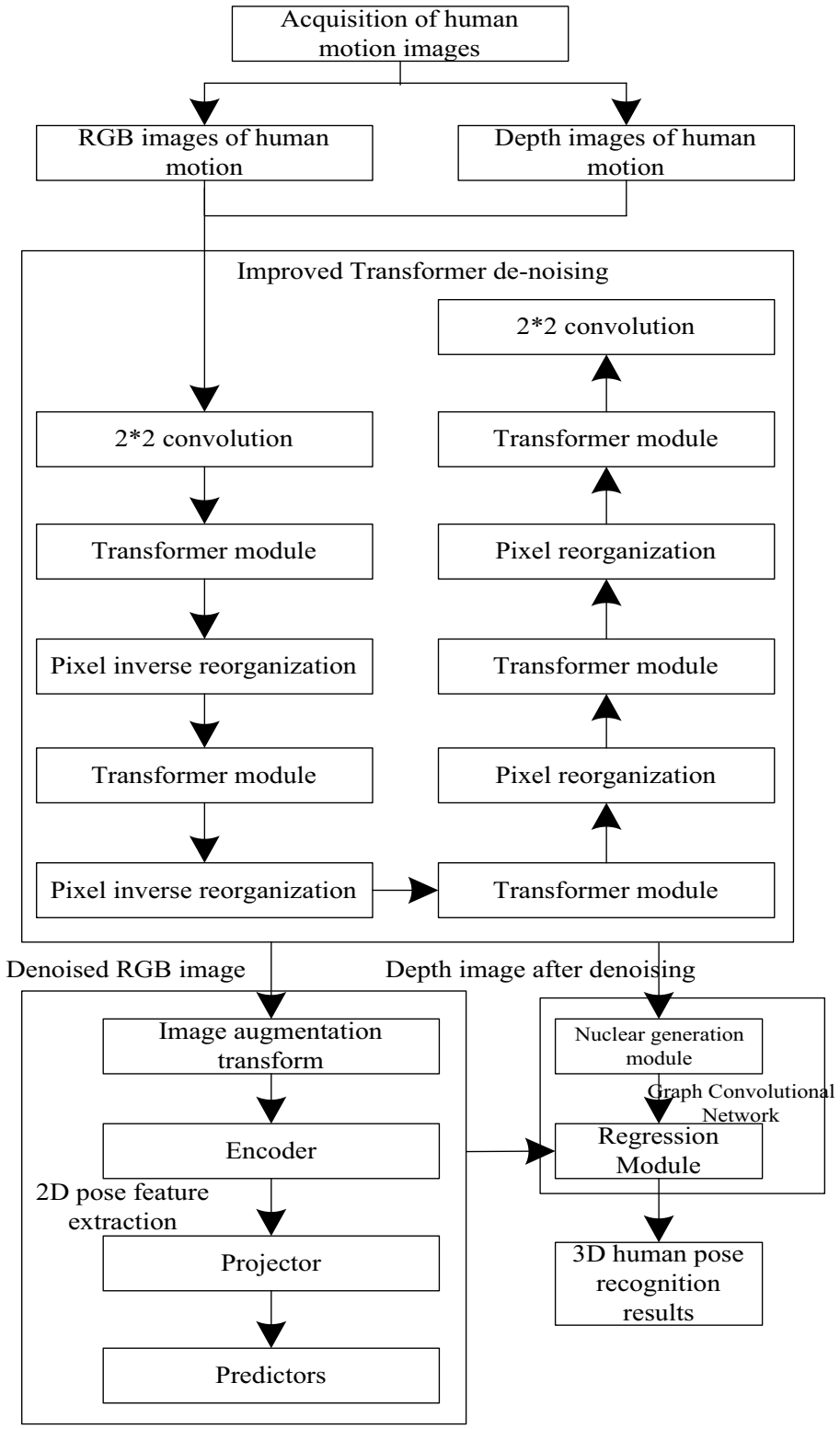


Fig. 1 The architecture of 3D human motion pose recognition

using deep contrastive learning and an improved Transformer, and the overall architecture of the method is shown in Fig. 1.

According to Fig. 1, first, the Kinect camera is used to capture RGB and depth images of human motions. Then, the collected human motion RGB and depth images are de-noised using an improved Transformer. Second, 2D human motion pose features are extracted from the de-noised human motion RGB images using deep contrastive learning. Finally, the depth features of the human motion pose are extracted from the de-noised human motion depth images using the kernel generation module in the GCN. The extracted 2D human motion pose features and depth features are fused using a regression module to obtain the recognition results of 3D human motion pose.

2.2 Human Motion Image De-noising Based on Improved Transformer

Through the Kinect camera, RGB and depth images of human motion are captured. As the Kinect contains both a color camera and a depth camera [15, 16], it can effectively capture RGB and depth images of human motions.

2.2.1 Low-Dimensional Embedding

Use a 2*2 convolutional layer to embed the collected human motion image I into a low-dimensional feature K_0.

2.2.2 Deep Feature Transformation

Use a symmetric encoder–decoder to transform K_0 into a deep feature K_d. Each level of the encoder–decoder contains several transformer modules [17, 18]. In order of resolution from largest to smallest, K_0 is sequentially input into the encoder, and pixel reassembly [19, 20] is performed on K_0 by down-sampling. The input of the decoder is the low-resolution latent feature of pixel reshuffling K_0, which is restored to the high-resolution part of K_0, and the refined feature K_r is obtained. The pixel reshuffling is performed to K_r.

2.2.3 Motion Image Generation

The residual human motion image I' is generated based on the pixel-reorganized K_r using a 2*2 convolutional layer [21, 22], expressed as

$$I' = \frac{\left(I - IK_r\right)^2}{I_{\text{all}}}. \tag{1}$$

The loss function calculates the pixel-level difference between the reconstructed image and the original image, takes the sum of squares, and computes the average. This metric is used to measure the discrepancy between the network's predicted image and the ground truth image.

The de-noised human motion image $\hat{I}$ is obtained by adding I' and I, as shown in formula

$$\hat{I} = I' + I. \tag{2}$$

According to the above process, we can obtain the de-noised RGB image $\hat{I}_\alpha$ of human motion, as well as the human motion depth image $\hat{I}_\beta$. The improved Transformer is utilized to de-noise human motion images, providing a foundation for extracting human motion pose features based on deep contrastive learning.

2.3 Human Motion Pose Feature Extraction Based on Deep Contrastive Learning

2.3.1 Contrastive Enhanced Feature Extraction

Deep contrastive learning is a self-supervised learning method used to analyze human motion pose images and extract 3D features of different poses, without the need for labels. Using deep contrastive learning, human motion pose features can be extracted from de-noised RGB images $\hat{I}_\alpha$. The deep contrastive learning method includes two neural networks, namely the online network and the target network [23, 24]. The number of encoders f_θ, projectors l_θ, and predictors b_θ in the online network is 1. The weight of the online network is θ; the number of encoders f_w and projectors l_w in the target network is 1, and the weight of the target network is w.

2.3.2 Contrastive Enhanced Viewpoint Transformation

(1) A de-noised human motion RGB image $x_{\alpha,i}$ is randomly selected within $\hat{I}_\alpha$, and the image augmentation transform [25] is expanded on $x_{\alpha,i}$ to obtain the human motion RGB images from two viewpoints with the equation

$$\begin{aligned} v_{\alpha,i} &= \varepsilon\left(x_{\alpha,i}\right), \\ v'_{\alpha,i} &= \varepsilon'\left(x_{\alpha,i}\right), \end{aligned} \tag{3}$$

where $v_{\alpha,i}$ and $v'_{\alpha,i}$ are the RGB images of human motion before and after the augmentation transformation. ε and ε' are the coefficients of the augmentation transformation of two views.

(2) Input $v_{\alpha,i}$ inside f_θ, and obtain

$$y_\theta = \lambda f_\theta\left(v_{\alpha,i}\right), \tag{4}$$

where the output of f_θ is y_θ; λ is the regularization factor.

(3) Input l_θ, and obtain

$$z_\theta = \lambda l_\theta\left(y_\theta\right), \tag{5}$$

where z_θ indicates the output of l_θ. The output of b_θ is $b_\theta\left(z_\theta\right)$.

That is, the output of the deep contrastive learning online network is $b_\theta\left(z_\theta\right)$.

(4) Input $v'_{\alpha,i}$ inside f_w, and obtain

$$y'_w = \lambda f'_w\left(v'_{\alpha,i}\right), \tag{6}$$

where the output of f_w is y'_w.

(5) Output l_w, and obtain

$$z'_w = \lambda l'_w\left(y'_w\right), \tag{7}$$

(6) The normalized treatment of $b_\theta\left(z_\theta\right)$ and z'_w with the following equation:

$$\overline{b}_\theta\left(z_\theta\right) = \frac{b_\theta\left(z_\theta\right)}{\left\|b_\theta\left(z_\theta\right)\right\|_2}, \tag{8}$$

$$\overline{z'}_w = \frac{z'_w}{\left\|z'_w\right\|_2}. \tag{9}$$

As a result, the normalized results of $b_\theta\left(z_\theta\right)$ and z'_w are $\overline{b}_\theta\left(z_\theta\right)$ and $\overline{z'}_w$. The mean square error (MSE) between $\overline{b}_\theta\left(z_\theta\right)$ and $\overline{z'}_w$ is calculated as follows:

$$e_{\theta,w} = \left\|\overline{b}_\theta\left(z_\theta\right) - \overline{z'}_w\right\|_2^2. \tag{10}$$

2.3.3 Total Error Optimization and Pose Feature Extraction

Input $v'_{\alpha,i}$ in the online network, input $v_{\alpha,i}$ in the target network, the MSE at this time is $e'_{\theta,w}$. The minimum total error function is used as the objective to output the human motion pose feature extraction results with the following equation:

$$J_{\theta,w} = \min\left(e_{\theta,w} + e'_{\theta,w}\right). \tag{11}$$

According to $J_{\theta,w}$ optimization θ, the optimization process of θ is as follows:

$$\theta \leftarrow optimizer\left(\nabla_\theta J_{\theta,w}, \eta\right), \tag{12}$$

where *optimizer* is the optimizer; η is the learning rate. ∇_θ is the optimizer variable.

That is, the optimization process of w is

$$w \leftarrow \theta - h\theta, \tag{13}$$

where h is the decay rate. w is the human motion pose feature extraction for deep contrastive learning.

2.4 Implementation of 3D Human Motion Pose Recognition

Based on the removal of human motion image noise using the improved Transformer method, the kernel generation module within the GCN [21] in deep contrastive learning is used to extract the depth features D_β of the human motion pose within the $b_\theta\left(z_\theta\right)$ and $\hat{I}_\beta$ of the 2D pose feature extraction of contrastive learning using the regression module to fuse the depth and obtain the 3D human motion pose recognition results.

2.4.1 High-Dimensional Feature Extraction

The dimensionality of $\hat{I}_\beta$ is compressed and input into the multilayer perceptron of the kernel generation module to obtain the height feature u and the width feature s of the human motion depth image, which represent the probability of the edge distribution of the human joint points in the ranks of the human motion depth image, respectively.

2.4.2 Location Distribution Probability Calculation

Calculate the probability of the location distribution of each joint on $\hat{I}_\beta$ with the following equation:

$$P(u,s) = \varphi\left(p_u \times p_s\right), \tag{14}$$

where in the human motion depth image $\hat{I}_\beta$, and the edge distribution probabilities correspond to u and s are p_u and p_s. U are mapping functions.

2.4.3 Depth Feature Extraction and Regression

The depth feature Q_β of the human motion pose is extracted by spatial filtering kernel, convolution $\hat{I}_\beta$, with the following equation:

$$Q_\beta = \sum_{u=1}^{N}\sum_{s=1}^{M} P(u,s)\rho(u,s), \tag{15}$$

where the number of u and s is N and M.ρ is the spatial filtering kernel.

In the regression module of the GCN, input $b_\theta\left(z_\theta\right)$ and Q_β to obtain the 3D human motion pose recognition results with the following equation:

$$C = \sigma\left(\left[\sum (A + I_\eta)\right]^{-\frac{1}{2}}(A + I_\eta)\left[\sum (A + I_\eta)\right]^{-\frac{1}{2}}(b_\theta(z_\theta), Q_\beta)\omega\right), \tag{16}$$

where σ is the activation function; ω is the weight; A is the adjacency matrix of the human motion image; I_η is a unit array of order η.

3 Experimental Analysis and Results

3.1 Experimental Environment and Datasets

The server configuration used in the experiment for motion pose classification consisted of an Intel(R) Xeon(R) Silver 4114 CPU, 64GB of memory, and a NVIDIA Quadro RTX A4000 16G. We used a 3D human motion pose recognition method based on the Transformer model, which was implemented using a Kinect camera. We also used an improved Transformer to eliminate internal noise from the RGB and depth images of human motion.

In this experiment, we used the HiEve dataset, the MPII Human Pose dataset and our dataset as our objects of study. We applied our method to perform real-time multi-person pose tracking on videos from both datasets to verify the effectiveness of our 3D human motion pose recognition method. (1) The HiEve dataset consists of 32 video sequences collected from YouTube that include abnormal scenes. Most of these videos are over 900 frames long, with a total length of 33 min and 18 s. The dataset is divided into 19 training sets and 13 testing sets. (2) The MPII Human Pose dataset is a dataset used to evaluate human pose recognition. It covers 410 human activities, and each image is labeled with an activity tag. (3) To collect data for our experiments, we installed Kinect cameras in front of and behind a fitness center and used our method to capture RGB and depth images of human motion in the fitness center. The resolution of the RGB images is 960 × 540, while the depth images are 512 × 424. The sampling rate of both datasets is consistent at 40 Hz. The RGB images and depth images of human motion are acquired using the proposed method, where the results of the RGB images and depth images of human motion within the dataset are shown in Fig. 2.

(a) RGB image (b) depth image

Fig. 2 RGB images of human motion and depth image acquisition results

According to Fig. 2, this method can effectively capture RGB images and depth images of human motion in the process of human motion, and the clarity of image acquisition is better.

The proposed method is trained on a total of 10,000 video sequences from the HiEve dataset and the MPII Human Pose dataset. These datasets contain diverse video target objects, providing a certain level of generality. The deep contrastive learning method is employed and the network is trained for 1200 iterations. The initial learning rate is set to 0.001 and is decayed and stabilized at each epoch. The weight decay factor is set to 0.0001, the batch size is set to 64, the number of layers is 4, and the number of hidden units is set to 256. In two datasets, 80% of the images were selected as the training set and 20% of the images were selected as the test set. By training on different datasets, the model can learn a wider range of pose variations and motion features in various environments, thereby enhancing its generalization capability. Such training process enables the model to better adapt to unseen testing data and effectively handle pose recognition tasks in new scenarios and different backgrounds. This enables the model to perform better in real-world 3D human motion pose recognition tasks.

3.2 Evaluating Indicator

(1) Feature extraction point distribution: the human is extracted at each joint point to generate images. By comparing the point distribution, the feature extraction accuracy of different methods can be viewed directly from the visual aspect.

(2) Motion pose recognition: Generate human motion pose recognition images in the 3D coordinate map, and objectively display the accuracy of each joint point and its connection line.

(3) Recognition speed analysis: the running time of motion pose recognition algorithm under the influence of different noises. It can analyze the system in real time according to its operation in different scenarios of recognition. The calculation formula is

$$v = f\left[s/(t - r)\right], \tag{17}$$

where v is the recognition speed, f is the sweep time. t is the unit time. r is the sweep time in case of noise.

(4) Recognition accuracy: human motion pose recognition accuracy. The calculation formula is

$$y = kx + b, \tag{18}$$

where y denotes the recognition accuracy; k is the recognition accuracy when the intensity of iterative attack varies; x represents the iterative attack intensity; b denotes the recognition accuracy when the iterative attack strength $x = 0$.

3.3 Results and Discussion

The pose recognition of deep learning uses convolutional networks to extract key parts of the human skeleton. This can be divided into two ways: one is based on skeleton pose recognition, which classifies based on the position relationship of the key parts of the human skeleton, and the other is through video image recognition, which recognizes the human motion process based on skeletal features (2D and 3D). In which, human pose recognition technology based on skeletal features analyzes the motion process composed of skeletal features (2D and 3D) in the human motion process to achieve rapid and accurate human motion. Therefore, according to the proposed method, the human motion pose features are extracted from the RGB images of human motion captured in Fig. 2. The OpenPose pose estimation algorithm is utilized to directly extract the coordinates of human body key-points from the images, the skeletal points are depicted, and the feature results are displayed as the graph shown in Fig. 3. The size of Fig. 3 is 257×228, with a bit depth of 32.

As can be seen in Fig. 3, the proposed method can effectively reflect the human motion pose features, and the extracted features cover most parts of the human, which can provide more comprehensive data support for the subsequent 3D human motion pose recognition.

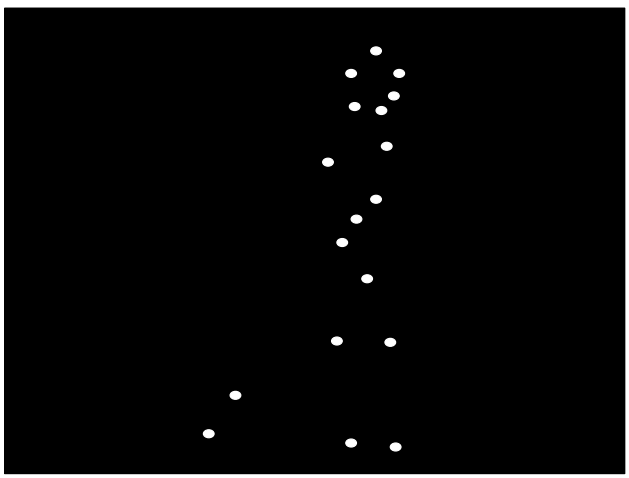


Fig. 3 Human motion pose feature extraction results

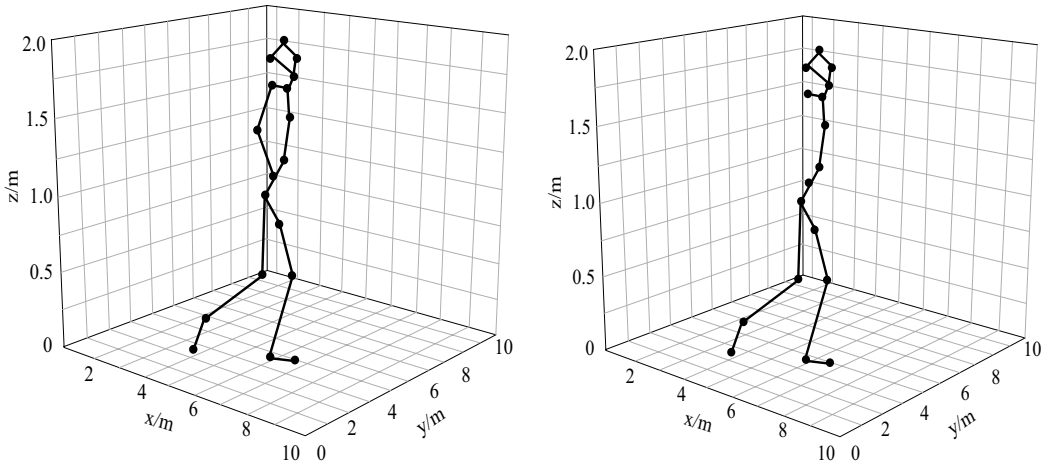
The comparison methods of ABDSF [10], PACAC [11], HMARM [12], MSCFL [13], A2DRH [14], and the proposed method were used to recognize the 3D human motion pose of Fig. 3 using these six methods, and the recognition results are shown in Fig. 4.

According to Fig. 4a, the distribution of the 3D human joints recognized by proposed method is consistent with the distribution in Fig. 2. By comparing Fig. 4b–e, and Fig. 4a, it can be seen that methods of ABDSF [10], PACAC [11], and HMARM [12] show partial limb disappearance phenomena during human recognition. Figure 4b does not recognize the pose of the arms, and Fig. 4c does not recognize the poses of both feet and the right calf. Figure 4d does not recognize the poses of both feet and the right calf either. According to Fig. 4e, the recognition results of MSCFL [13] method have problems with the lack of curvature in the upper body, the fracture of the lower body, and poor coherence. According to Fig. 4f, the A2DRH [14] method recognized images missing the head frame, the pose of the left thigh, and the pose of the foot, which is significantly different from the recognition results of the proposed method. In summary, based on the above analysis, when recognizing the human motion pose in Fig. 2, the accuracy of the proposed method is higher than the other five methods, and the deviation in the 3D direction of our method is smaller than that of the other five methods. Therefore, the proposed method can accurately depict most of the poses of the target task, that is, it can accurately complete the 3D human motion pose recognition task with high accuracy.

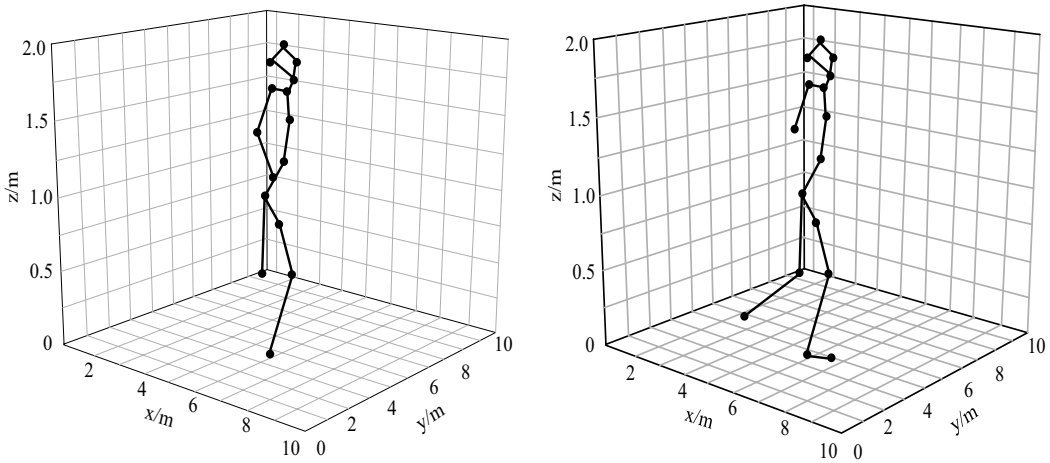
Using frame rate to measure the speed of 3D human motion pose recognition by the proposed method, the average frame rate of 3D human motion pose recognition by proposed method and five other methods for all images in both datasets at different noise levels is analyzed, and the analysis results are shown in Table 1. The frame rate threshold is 15 frame/s.

According to Table 1, as the noise level increases, the recognition speed of the six methods becomes slower, and the proposed method has the fastest recognition speed. When the noise level is 40, the recognition speed of the proposed method for the two datasets is 9.1 frame/s and 9.2 frame/s, respectively, which is faster than the recognition speed of methods ABDSF [10], PACAC [11], HMARM [12], MSCFL [13], and A2DRH [14]. In summary, under the influence of different noise levels, the proposed method has a higher recognition speed than the compared methods. Therefore, the proposed method has a high recognition speed for human motion pose and can be applied to human motion pose recognition.

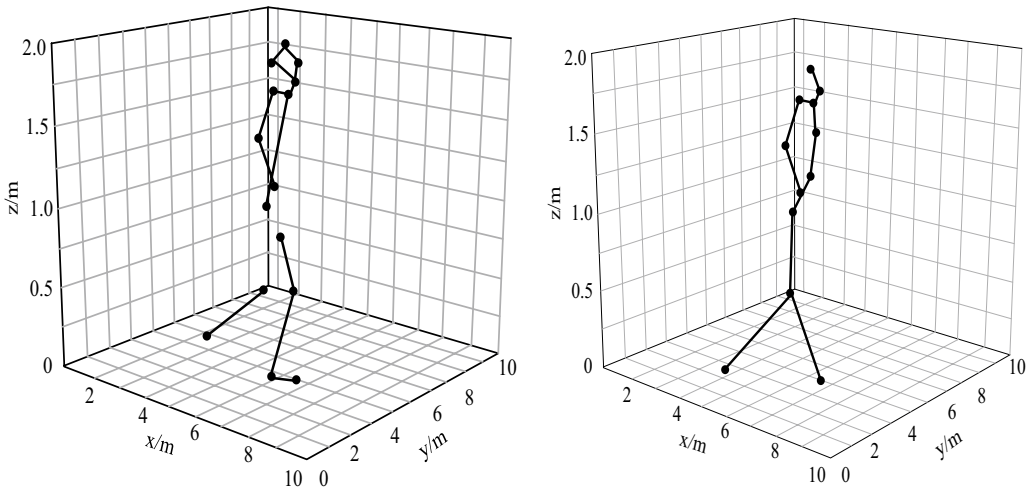
The recognition accuracy of the above six methods for recognizing 3D human motion pose when different iterations of attack intensity are analyzed, and the analysis results are shown in Fig. 5. It can be seen that as the iterative



(a) The recognition results of the proposed method (b) The recognition results of ABDSF[10] method



(c) The recognition results of PACAC[11] method (d) The recognition results of HMARM [12] method



(e) The recognition results of MSCFL [13] method (f) The recognition results of A2DRH [14] method

Fig. 4 Comparison results of 3D human motion pose recognition

Table 1 Comparison results of 3D human motion pose recognition speed analysis (frame/s)

Noise level	5	10	15	20	25	30	35	40
Proposed method								
HiEve dataset	4.8	6.1	6.7	6.8	7.2	7.9	8.6	9.2
The MPII Human Pose dataset	4.6	5.9	6.7	7.2	7.8	8.2	8.5	9.1
ABDSF [10] method								
HiEve dataset	4.8	6.2	6.8	6.9	7.3	8.0	8.8	9.5
The MPII Human Pose dataset	5.1	6.1	6.9	7.3	7.9	8.4	8.6	9.4
PACAC [11] method								
HiEve dataset	4.9	6.0	7.0	7.0	7.5	8.1	8.7	9.4
The MPII Human Pose dataset	5.1	6.2	6.9	7.2	8.0	8.6	8.6	9.2
HMARM [12] method								
HiEve dataset	5.2	6.4	6.7	6.9	7.6	8.4	8.9	9.3
The MPII Human Pose dataset	4.9	6.2	6.9	7.4	7.9	8.5	8.8	9.4
MSCFL [13] method								
HiEve dataset	5.1	6.3	6.8	7.0	7.5	8.2	8.8	9.5
The MPII Human Pose dataset	4.9	6.0	7.0	7.4	8.0	8.3	8.7	9.4
A2DRH [14] method								
HiEve dataset	4.8	6.2	6.8	7.0	7.4	8.3	8.7	9.3
The MPII Human Pose dataset	5.0	6.4	6.7	7.3	8.2	8.4	8.6	9.4

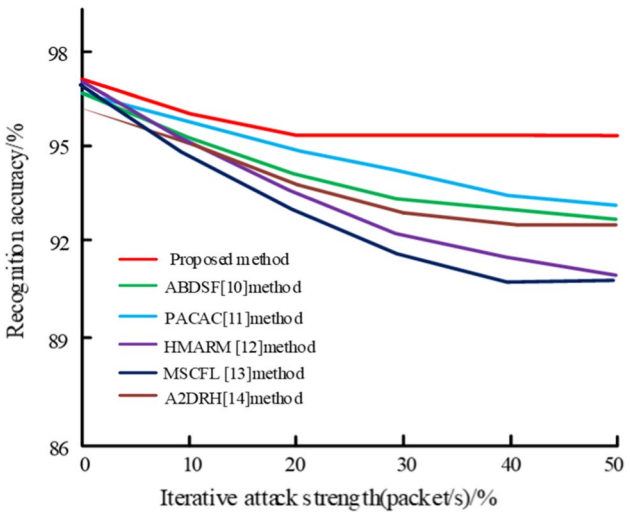


Fig. 5 Comparison results of recognition accuracy of 3D human motion pose

attack strength increases, the accuracy of all six methods decreases, with the smallest decrease observed for the proposed method. When the iterative attack strength reaches 20, the recognition accuracy of the proposed method stabilizes at around 95%. In contrast, the recognition accuracies of the methods in ABDSF [10], PACAC [11], and A2DRH [14] steadily decrease before stabilizing at around 92.5%. Overall, the proposed method consistently achieves higher recognition accuracy than the other five methods under different iterative attack strengths. Therefore, the proposed method demonstrates superior accuracy in 3D human motion pose recognition under different iterative attack strengths, highlighting its effectiveness in achieving high accuracy pose recognition.

Based on the aforementioned experimental results, the proposed method effectively captures RGB and depth images of human motion, with high image clarity. This indicates that the extracted features cover most parts of the human body and provide comprehensive data support for 3D human motion pose recognition. The proposed method demonstrates strong generalization capability in handling different human poses and motion types. Based on the results of pose recognition, the proposed method accurately describes most poses and achieves high precision in 3D human motion pose recognition tasks. This indicates that the method maintains good recognition performance even in challenging scenarios involving deformation, noise, and occlusion, demonstrating a certain level of robustness. Additionally, the accuracy analysis results of 3D human motion pose recognition under different iterations show high accuracy, indicating the method's robustness against targeted attacks.

4 Conclusions

Deep learning and convolutional neural networks have made significant progress in the field of computer vision. This paper proposes a method for 3D human motion pose recognition based on deep contrastive learning and an improved Transformer. By utilizing the kernel generation module within a GCN, this method extracts depth features of human motion poses, achieving accurate recognition of 3D human

Table 2 Abbreviations of description symbols/terminology of this paper

Symbols/nomenclature	Description	Abbreviation
I'	Residual human motion image	–
$\hat{I}_{\alpha}$	De-noised RGB image	–
$\hat{I}_{\beta}$	De-noised depth image	–
f_{θ}	Encoders of online network	–
l_{θ}	Projectors of online network	–
b_{θ}	Predictors	–
f_{w}	Encoders of target network	–
l_{w}	Projectors of target network	–
θ	Weight of the online network	–
$v_{\alpha,i}$	RGB images of human motion before augmentation transformation	–
$v'_{\alpha,i}$	RGB images of human motion after augmentation transformation	–
λ	Regularization factor	–
$e'_{\theta,w}$	MSE at this time	–
η	Learning rate	–
∇_{θ}	Optimizer variable	–
ρ	Spatial filtering kernel	–
σ	Activation function	–
ω	Weight	–
A	Adjacency matrix of the human motion image	–
v	Recognition speed	–
f	Sweep time	–
t	Unit time	–
Graph convolutional networks	–	GCN
Two-dimensional	–	2D
Three-dimensional	–	3D
The mean square error		MSE

motion poses. Experimental results demonstrate that the proposed approach effectively captures RGB and depth images of human motion and performs well in extracting 2D human motion pose features. Moreover, this method exhibits fast speed and high accuracy in 3D human motion pose recognition, as evidenced by the experimental results under different iteration attack intensities. However, challenges still exist in multi-person pose estimation, such as occlusion, interference between individuals, and variability in poses, which affect the accuracy of human motion pose recognition. Therefore, in future research, it is necessary to explore more accurate and robust methods for constructing human detectors to address the challenges posed by occlusion and multi-person pose estimation.

Appendix

Abbreviation of this paper is shown in Table 2.

Author Contributions Conceptualization and methodology: DL. Data curation and validation: HY. Writing—review and editing: ZL.

Funding The authors received no funding for this study.

Data Availability The data used to support the findings of this study are included in the article.

Declarations

Conflict of Interest The authors declare that they have no competing interests.

References

1. Kim, S.H., Jeong, S., Park, S., Chang, J.: Camera motion agnostic method for estimating 3D human poses. Sensors **22**(20), 7975 (2022)
2. Xia, Z., Xing, J., Li, X.: Gesture tracking and recognition algorithm for dynamic human motion using multimodal deep learning. Secur. Commun. Netw. **2022**, 1–11 (2022)
3. Tan, T., Hus, J., Liu, S., Huang, Y., Gochoo, M.: Using direct acyclic graphs to enhance skeleton-based action recognition with a linear-map convolution neural network. Sensors **21**(9), 3112 (2021)
4. Wang, D., Li, T., Deng, P., et al.: A generalized deep learning algorithm based on NMF for multi-view clustering. IEEE Trans. Big Data **9**(1), 328–340 (2023)
5. Xu, W., Zhu, Z.: Estimation for Human Motion Posture and health using improved deep learning and Nano biosensor. Int. J. Comput. Intell. Syst. **16**(55), 1–10 (2023)
6. Hnatiuc, M., Geman, O., Avram, A., Gupta, D., Shankar, K.: Human signature identification using IoT technology and gait recognition. Electronics **10**(7), 852 (2021)
7. Shimada, S., Golyanik, V., Xu, W., Pérez, P., Theobalt, C.: Neural monocular 3d human motion capture with physical awareness. ACM Trans. Graph. **40**(4), 1–15 (2021)
8. Wu, B., Wang, C., Huang, W., Huang, D., Peng, H.: Recognition of student classroom behaviors based on moving target detection. Traitement du Signal **38**(1), 215–220 (2021)
9. Miao, A., Liu, F.: Application of human motion recognition technology in extreme learning machine. Int. J. Adv. Rob. Syst. **18**(1), 4–18 (2021)
10. Jiang, J., Wang, X.Y., Gao, M., Pan, J., Zhao, C., Wang, J.: Abnormal behavior detection using streak flow acceleration. Appl. Intell. Intell. **52**(9), 10632–10649 (2022)
11. Zhao, R., Lang, C., Li, Z., Liang, L., Wei, L., Feng, S., et al.: Pedestrian attribute recognition based on attribute correlation. Multimedia Syst. **28**(3), 1069–1081 (2022)
12. Deng, P., Wu, M.: Human motion attitude recognition method based on machine learning. J. Chin. Inertial Technol. **30**(1), 37–43 (2022)
13. Ma, X., Yan, Y.: Multiscale spatio-temporal correlation feature learning for human pose estimation. J. South-Central Univ. Nationalities (Natural Science Edition) **42**(1), 95–102 (2023)
14. Zhang, Z., Wang, S., Liu, C., Xie, R., Hu, W., Zhou, P.: All-in-one two-dimensional retinomorphic hardware device for motion detection and recognition. Nat. Nanotechnol.Nanotechnol. **17**, 27–32 (2022)
15. Wang, H., Huang, D., Wang, Y.: Gridnet: efficiently learning deep hierarchical representation for 3d point cloud understanding. Front. Comput. Sci. **16**(1), 161301 (2022)
16. Ohri, K., Kumar, M.: Review on self-supervised image recognition using deep neural networks. Knowl.-Based Syst..-Based Syst. **224**(8), 107090 (2021)
17. Wei, C., Xu, Y., Jiang, X., et al.: Automatic segmentation algorithm of dermoscopy image based on transformer and convolutional neural network. J. Comput.-Aid. Design Comput. Graph. **34**(12), 1877–1886 (2022)
18. Gao, F., Ji, S., Guo, J., et al.: A multi-stage transformer network for image dehazing based on contrastive learning. J. Xi'an Jiaotong Univ.Jiaotong Univ. **57**(1), 195–210 (2023)
19. Chen, N., Zhang, Y., Wu, J., Zhang, H., Chamola, V., Albuquerque, V.: Brain–computer interface-based target recognition system using transfer learning: a deep learning approach. Comput. Intell.. Intell. **38**(1), 139–155 (2022)
20. Duan, X., Huang, J.: Deep-learning-based 3d cellular force reconstruction directly from volumetric images. Biophys. J.. J. **121**(11), 2180–2192 (2022)
21. Cao, Y., Qiu, Q.: Two-channel dilated convolution attentional image denoising network. Appl. Res. Comput. **40**(5), 1548–1552 (2023)
22. Vilar, C., Krug, S., Thrnberg, B.: Processing chain for 3d histogram of gradients based real-time object recognition. Int. J. Adv. Rob. Syst. **18**(1), 76–486 (2021)
23. Hao, Y., Liang, W., Yang, L., He, J., Wu, J.: Methods of image recognition of overhead power line insulators and ice types based on deep weakly-supervised and transfer learning. IET Gener. Transm. Distrib.Gener. Transm. Distrib. **16**(11), 2140–2153 (2022)
24. Chen, Y., Kuang, C.: Pedestrian re-identification based on CNN and TransFormer multi-scale learning. J. Electron. Inf. Technol. **45**(6), 2256–2263 (2023)
25. Su, Y., Liu, C.: Three-dimensional human reconstruction model based on high-resolution net and graph convolutional network. J. Comput. Appl. **43**(2), 583–588 (2023)