

ITC 1/54 Information Technology and Control Vol. 54 / No. 1 / 2025 pp. 115-134 DOI 10.5755/j01.itc.54.1.38054	Estimation and Recognition Methods of Human Gait Pose based on Computer Vision and Transformer	
	Received 2024/07/15	Accepted after revision 2024/12/23
	HOW TO CITE: Zou, Y., Zhou, X., Cai, C., Wang, Y. (2025). Estimation and Recognition Methods of Human Gait Pose based on Computer Vision and Transformer. <i>Information Technology and Control</i> , 54(1), 115-134. https://doi.org/10.5755/j01.itc.54.1.38054	

Estimation and Recognition Methods of Human Gait Pose based on Computer Vision and Transformer

Yu Zou

School of Artificial Intelligence (School of Future Technology), Nanjing University of Information Science & Technology, Nanjing 210044, China
 Science and Technology Division, Jiangsu Open University, Nanjing 210036, China

Xianchun Zhou, Chuangxin Cai

School of Artificial Intelligence (School of Future Technology), Nanjing University of Information Science & Technology, Nanjing 210044, China

Yixuan Wang

College of Cultural Management, Communication University of China Nanjing, Nanjing 211172, China

Corresponding author: YuZou1232024@outlook.com

Human gait pose estimation and recognition, as an emerging biometric technology, have advantages such as no need for target object cooperation, difficulty in forgery, and long-distance recognition. However, compared with traditional biometric special recognition, it is more susceptible to the influence of target object's arbitrary motion. In response to the above issues, the study introduces heterogeneous transfer learning to construct a human gait pose estimation and recognition method based on computer vision and Transformer, and improves it using the perspective gradually shift training method based on this. The research results indicated that the improved human gait pose estimation and recognition model had good recognition performance in 11 perspectives with intervals of 16° from 0° to 180°, and the corresponding change curve remained stable, with an average recognition rate of over 97%. The average initial validation rate of the improved model was 65.32% higher than before, and the maximum validation rate of the improved model achieved significant improvement from different angles. In comparison with other mainstream algorithms, the improved model proposed in the study had the highest average validation rate and average accuracy, which were 98.56% and 97.51%, respectively, and the corresponding average improvement index was greater than 20%. The above results confirm the performance and reliability of the research method, providing new solutions for the problem of human gait pose estimation and recognition in complex scenarios.

KEYWORDS: Computer vision; Transformer; Human gait; Pose estimation; Infrared image.

1. Introduction

In recent years, China's information infrastructure industry and artificial intelligence field have developed rapidly, and society has put forward higher demands for personal information security and production efficiency. Traditional biometric technologies such as facial recognition, iris recognition, and fingerprint recognition are difficult to meet people's various needs in different application scenarios [15, 21, 25]. The estimation and recognition of human gait posture (ERHGP), as an emerging biometric method, has received widespread attention from many scholars. This method can perform long-distance non-contact recognition and is suitable for scenarios with frequent outbreaks of influenza or infectious diseases, without the active participation of the target object. Moreover, it has a high level of anti-counterfeiting, because the coordinated movements of different parts of the human gait posture design have the characteristic of being difficult to change [1, 22]. In addition, the cost of cameras used for acquisition is relatively low, which is suitable for large-scale deployment, and can also perform 360° full angle recognition [4]. However, the ERHGP method has the following problems: firstly, in terms of application, China is in the exploratory stage, so there are limitations in large-scale commercial applications; secondly, the target object has arbitrary motion and high computational complexity, making processing more difficult [2, 8]. Gao et al. proposed a novel skeleton-based gait recognition model to address the issue that general contour-based gait recognition methods rely on binary human contours and are easily affected by external factors. They conducted experiments on two datasets, CASIA-B and OUMVLP-Pose, and the results confirmed that this model had high recognition accuracy and significant robustness [6]. Yeo et al. designed a region-based three branch convolutional network gait recognition method to address the issue of traditional gait recognition methods being easily affected by covariate conditions, resulting in a significant decrease in accuracy. The study was validated in the CASIA-B database, and the results showed that the method exhibited good performance of 72.98% under covariate conditions [29]. Ozturk et al. aimed to achieve a reliable, easy-to-use, and high-precision gait recognition method, and constructed a gait recognition method using millimeter wave radio. The experiment

demonstrated the practicality of this method and achieved an average recognition accuracy of 79.1% [17]. Mathivanan et al. developed an advanced deep belief network algorithm based on black widow optimization, and recognized it through human gait images. The method was implemented on the MATLAB platform and compared with mainstream methods such as artificial neural networks, recurrent neural networks, and particle swarm optimization. The results verified the superiority of this method [14]. Wen et al. proposed a novel multi-perspectives recognition model based on generative adversarial networks and conducted experiments on the CASIA-B dataset. The results showed that this method achieved good recognition performance in the sequences of luggage and outerwear [26]. Rahi et al. designed a human gait recognition architecture that combines attention mechanism with multi-stream Convolutional Neural Network (CNN) to achieve high accuracy in gait re-recognition. Compared with current advanced technologies, the results showed the superiority of this method [20]. Inturi et al. [9] proposed a visual-based method for fall detection, which utilizes an AlphaPose pre-trained network to obtain a set of keypoints of an object, and processes the keypoints using a CNN layer framework. Finally, long-term dependencies were ensured through long short-term memory structures. By comparing with OpenPose network, the results showed that the research method was more accurate in detecting keypoints [9]. Ogundokun et al. developed a classification model based on deep transfer learning, which mainly classifies four main actions: sitting, bending, lying, and standing. The research results showed that the improved model had the highest testing accuracy, validation accuracy, and training accuracy, which were 94.72%, 93.79%, and 97.06%, respectively [16]. Kulikajevs et al. designed a novel deep recursive layer network model to reduce or eliminate occlusion issues related to human torso visibility in frames. Simulation results showed that the research method achieved a sitting posture recognition accuracy of 91.47% at a speed of 10fps [10]. Elavarasi et al. proposed a deep learning algorithm to observe gait patterns by obtaining image frames from real-time environments for fall detection in elderly communities. The results showed that the long short-term memory network course provided 94% accuracy and minimal false alarms [5].

Based on the above content, it can be concluded that certain achievements have been made in the performance and application of human gait pose recognition, but there are still shortcomings in accuracy. This is because visual accuracy is easily affected by factors such as backpacks from different perspectives, resulting in lower recognition accuracy. Therefore, this study combines Heterogeneous Transfer Learning (HTL) to establish an ERHGP method based on Computer Vision and Transformer (CVT). Based on this, a Perspective Gradually Shifting Training (PGST) method is designed to optimize and obtain a multi-perspectives recognition model based on improved CVT. The research aims to solve the problem of recognizing human gait pose information from different perspectives in ERHGP, improve the accuracy of human gait pose recognition in samples of different scales, and expand the application fields of human gait pose recognition. The innovation of the research mainly lies in the following two points. The first point is to establish a new method for converting the perspective of human gait pose, and combine it with HTL to design an ERHGP method based on CVT. The second point is to design the PGST method to improve the CVT model and propose a multi-perspectives recognition model based on the improved CVT.

2. Method and Materials

In response to the problem of strong randomness and complexity in identifying target movements in ERHGP, this study first selects experimental datasets

and preprocessing methods, then proposes an ERHGP method based on CVT, and finally constructs a multi-perspectives recognition model based on improved CVT.

2.1. Human Gait Pose Estimation and Recognition Dataset and Preprocessing

The dataset used in the study is from the public ERHGP dataset, and the ERHGP dataset is selected based on single and multiple angles. The Gait of Institute of Automation, Chinese Academy of Sciences (GIACAS) is selected as the data set for experiment. The data set has different perspectives and scales. Details are denoted in Table 1.

In dataset C, a single angle infrared human gait pose estimation and recognition experiment is conducted. A fixed 90° shooting angle is used with an infrared camera to capture 153 pedestrians in a single angle. Simultaneously, four gait conditions: Normal Walking (NE), Slow Walking (SW), Fast Walking (SW), and Carrying A Backpack for Walking (CBW) are collected. In dataset B, a multi-perspectives human gait pose estimation and recognition experiment is conducted, including 31 females and 93 males, respectively. In addition, the study collects images of participants under three conditions: NE, CBW, and Wearing A Coat to Walk (WCW) from 11 different viewing angles ranging from 0° to 180° with an 18° angle interval. The images are also collected indoors in the same environment to avoid the influence of outdoor natural light noise. Dataset A is a small-scale library with a total of 20 people conducting experiments. Four image sequences are collected in the directions of 0°, 45°, and 90° from the image plane, respectively. The length of

Table 1

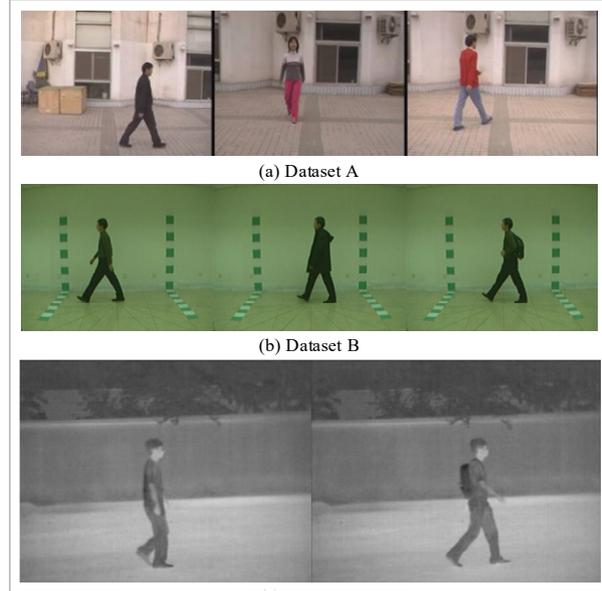
The specific content of the GIACAS dataset

Dataset	Dataset A	Dataset B	Dataset C
Environment	Outdoor	Indoor	Outdoor and Nighttime
Sample size	20	124	153
Is it an infrared image?	No	No	Yes
Number of perspectives	3	11	1
Walking conditions	The speed variation varies, and the frame rate of each sequence ranges from 37 to 127	Normal conditions, wearing a coat and carrying a package	Normal walking, fast walking, slow walking, and walking with bags
Outline data memory size/M	16	628	66

each sequence varies with the walking speed of the subjects, and a total of 13139 images are collected. Three examples of datasets are illustrated in Figure 1.

Figure 1

Example diagrams of three datasets



Because both dataset B and dataset C are used to obtain human contour features during walking through binarization and other operations, only unnecessary background cropping and other steps are needed to reduce the adverse effects caused by large differences in data quality. Then, all gait feature images are unified to $128 * 128$ pixels in size. Due to the fact that most previous studies only considered the instantaneous features corresponding to human gait posture, they often overlook the temporal correlation and continuity of human walking. Therefore, by introducing time domain features and dividing human gait data based on the maximum walking cycle, several gait cycle groups can be obtained. This operation not only preserves the focus of conventional human gait pose estimation and recognition on instantaneous features, but also fuses temporal correlated features in time periods, thereby enhancing the robustness of human gait recognition. The commonly used methods for calculating gait cycle include Absolute Difference Sum Algorithm (ADSA), Normalized Cross Correlation Coefficient (NCCC), and Zero Mean Normalized Cross Correlation Coefficient (ZMNC) [24].

In the ADSA, it is necessary to calculate the partial or overall similarity between the template graph and the search degree, and then search for the search graph that is most similar to the template graph to obtain the final matching result. The expression is shown in Equation (1).

$$ADSA(i, j) = \sum_{m=1}^K \sum_{t=1}^N |M(i+m-1, j+t-1) - T(m, t)|. \quad (1)$$

In Equation (1), $T(m, t)$ and M represent the template graph and search graph, respectively, while (i, j) represents the Manhattan distance $L1$ at the corresponding positions in $T(m, t)$ and M . Assuming there are two points $C(x_1, y_1)$ and $C(x_2, y_2)$, the distance $L1$ between the two points can be calculated using Equation (2).

$$L1 = |x_1 - x_2| + |y_1 - y_2|. \quad (2)$$

The principle of this method is simple, and the matching accuracy between images is also high. However, the corresponding computational load will rapidly increase with the increase of image size, and it is also very sensitive to possible noise. In the NCCC method, it is necessary to calculate the correlation between $T(m, t)$ and M , as expressed in Equation (3).

$$NCCC(x, y) = \frac{\sum_{i=1}^K \sum_{j=1}^N |M_{x,y}(i, j) - \bar{M}_{x,y}| |G(i, j) - \bar{G}|}{\sqrt{\left(\sum_{i=1}^K \sum_{j=1}^N |M_{x,y}(i, j) - \bar{M}_{x,y}|^2\right)^{\frac{1}{2}} \sqrt{\left(\sum_{i=1}^K \sum_{j=1}^N |G(i, j) - \bar{G}|^2\right)^{\frac{1}{2}}}}. \quad (3)$$

In Equation (3), \bar{G} and $\bar{M}_{x,y}$ correspond to the gray-scale mean of $T(m, t)$ and M , respectively. Due to the high sensitivity requirements for human gait posture features in infrared human gait images, and to avoid possible pattern matching errors in the above two methods, ZMNC is chosen to estimate the human gait cycle in the study, as shown in Equation (4).

$$ZMNC(x, y) = \frac{1}{n} \sum_{x,y} \frac{1}{\sigma_f \sigma_t} [f(x, y) - \mu_f] [t(x, y) - \mu_t]. \quad (4)$$

In Equation (4), $f(x, y)$ and $t(x, y)$ represent the pixel values corresponding to the original image and image template, n represents the number of pixels in the template, σ_f and σ_t , μ_f and μ_t correspond to the standard deviation and pixel mean of the original image and template image, respectively. Finally, the human gait images with temporal order are input

into the ZMNC function, and the correlation coefficient with the initial state image is calculated. By comparing and obtaining the feature repetition period, the maximum value is selected to estimate the gait cycle for subsequent experiments. In addition, the study also introduces mathematical morphology, which uses certain structural elements to measure and extract corresponding shapes in images. This not only simplifies image data and removes unimportant structures, but also preserves the basic shape characteristics of the image. The study mainly used the open operation method in morphological image processing to break narrow necks, eliminate fine protrusions, and make the contours of the image smoother.

2.2. Human Gait Pose Estimation and Recognition Method Based on CVT

In the past, convolutional modules (Conv Blocks) and other methods are used for image recognition in human gait pose estimation and recognition. The human gait feature maps are processed as a whole, which ignore the differences in human gait pose during dynamic walking, resulting in many useful human gait features being discarded, leading to problems such as easy saturation of recognition accuracy [19, 7, 3]. Therefore, based on CVT, this study proposes a method that can grid segment the human gait contour map, use independent feature subspaces to fit feature deconstruction, and finally refine the extraction of human gait features from spatiotemporal dimensions to improve the accuracy of human gait recognition. The images in the training set are optimized gait maps with only one target, so adding attention mechanisms and other methods cannot improve the recognition rate under occlusion conditions. In this case, increasing the number of convolutional layers for optimization is the most effective method, but it can also lead to low training efficiency. Therefore, additional lightweight methods were used in the study. Due to the reduction in the number of minimum data units used in pediatrics caused by the division of human gait cycle groups in the above section, a sliding window human gait cycle division method is used to expand the number of minimum data units. The specific process is as follows. The capacity of the human gait cycle is set to T , and the human gait characteristics up to the $N - T + 1$ th time can be used as the starting time of the human gait characteristics at each moment in the

time period. This not only ensures the spatiotemporal characteristics of the human gait, but also maximizes the approach to complex environmental applications, making the experimental results more reliable. Considering the fitting features of human gait recognition models in small sample datasets, a shallow dual path residual network is designed based on the Resnet network for comparison with subsequent CVT human gait pose recognition models. The schematic diagram of the Dual Path Convolutional Neural Network (DPCNN) model for human gait pose recognition is shown in Figure 2.

In Figure 2, firstly, it is necessary to parallelize the shallow convolutional block groups of Block1 and Conv1 Block, and at the same time, process the strengthened human gait features through a homogenization fusion module. Next, it needs to input it into Block2-4 to fit high-level human walking posture features. Then, it will input the binary adaptive average pooling layer to adjust the data format. Finally, it will output the results through the fully connected layer. For human gait pose recognition, the CVT method can distinguish the feature fitting methods of different positions, and to achieve a feature processing mode close to the Transformer, small blocks are converted into one-dimensional spatial tensors of equal size. This can achieve the effect of fitting human gait poses at different positions through independent feature weight space within a certain range. However, the above operations can cause spatial disorder. Therefore, this study uses position encoding to add position embedding values to the spatial tensor, and then adds a value that identifies the position at a fixed position of the spatial tensor to calibrate the relative position of the small block. There are currently two methods for adding positional embeddings, and the first expression is shown in Equation (5).

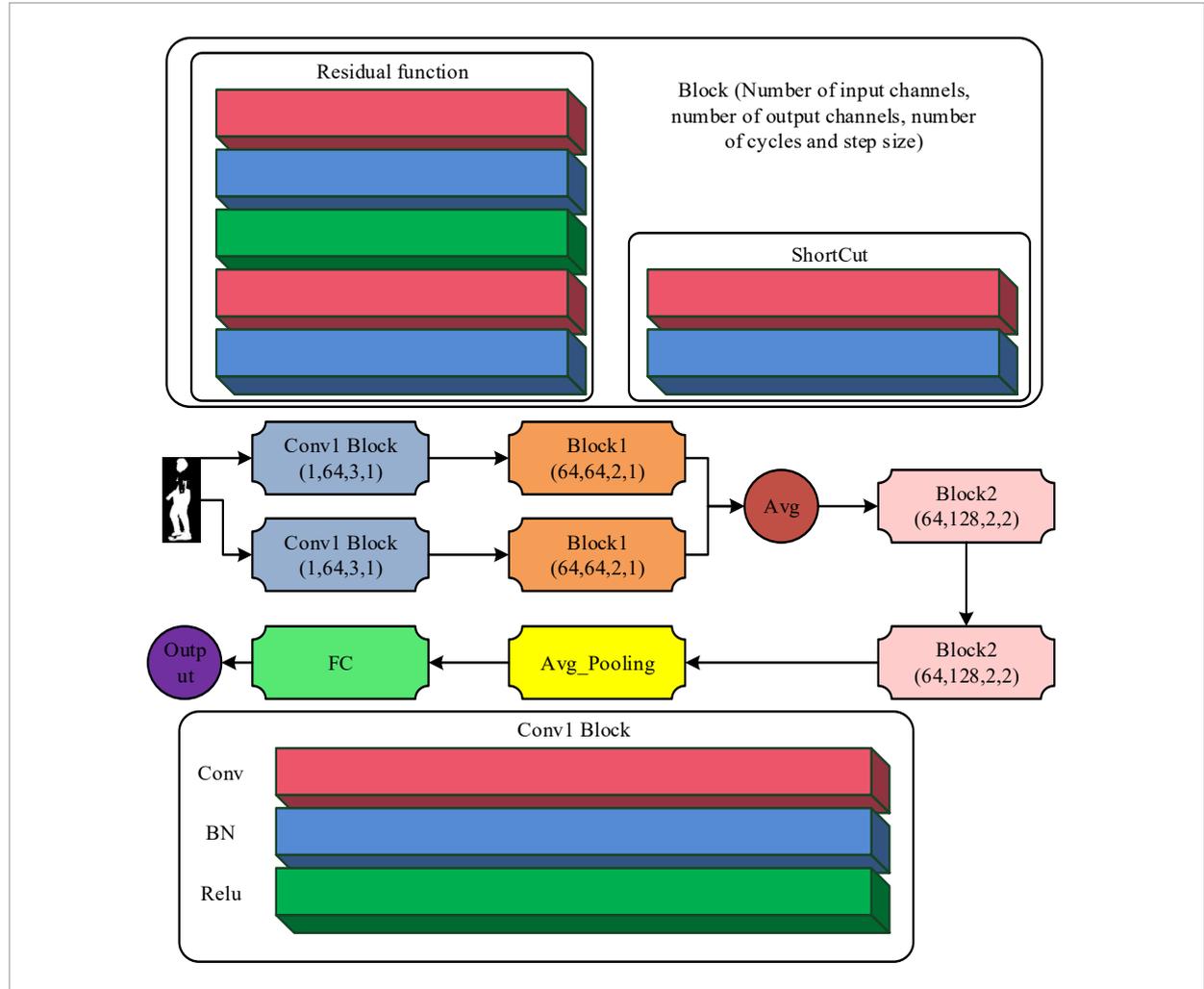
$$P_{(ps,2i)} = \sin \left(\frac{ps}{10000^{\frac{2i}{d_i}}} \right), i = 0, \dots, \frac{d_i}{2}, d_i = 512. \quad (5)$$

In Equation (5), ps represents the position sequence number of the token globally. The second expression is shown in equation (6).

$$P_{(ps,2i+1)} = \cos \left(\frac{ps}{10000^{\frac{2i}{d_i}}} \right), i = 0, \dots, \frac{d_i}{2}, d_i = 512. \quad (6)$$

Figure 2

Structural diagram of DPCNN model



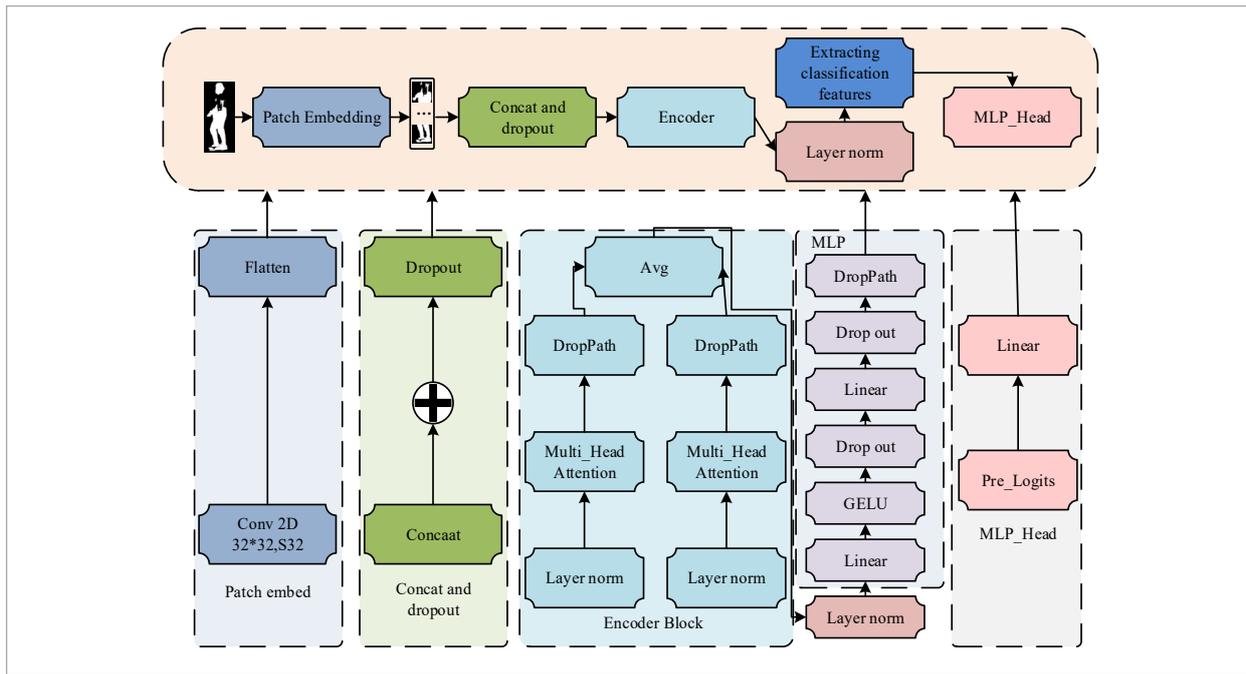
According to the size of the segmentation blocks, the above position embedding methods have good applicability. The CVT model with the above key points has certain methodological advantages in human gait recognition, but it has the problem of poor data fitting performance in small sample datasets. This is due to the independent subspace feature processing method constructed by the CVT method, which slows down its convergence speed in small sample data and enhances the dependence on model data enhancement and regularization [23, 11, 13]. Therefore, the study introduces the HTL method to pre-train large non-human gait datasets, and then transfers the training weight parameters to the dataset used in the study. In addition, to ad-

dress the limitations of the CVT method, a symmetrical dual attention mechanism human gait model based on CVT is studied and designed, as shown in Figure 3.

In Figure 3, it is necessary to first extend the image segmentation method based on CVT to the segmentation of human gait contour maps. The independent feature subspace is used to fit the independent human gait feature weights corresponding to the segmentation blocks of human gait contour maps at different times. Then, a dual channel multi-head attention module is established to enhance the feature processing ability. Finally, the human gait pose cycle group is input to provide more diverse factors for extracting human gait features.

Figure 3

A symmetrical dual attention mechanism human gait model based on CVT



2.3. Construction of a Multi-Perspectives Recognition Model Based on Improved CVT

At present, human gait pose recognition generally revolves around the adverse effects of extreme weather on recognition performance, as well as the more complex issue of human gait pose from multi-perspectives. Compared with biological features such as iris and face, human gait pose has non-coordination and difficulty in camouflage, among which non-coordination is the advantage of CVT method [27, 30, 18]. For the multi-perspectives problem for human gait pose recognition, an improved CVT model is constructed based on multi-perspectives human gait pose feature tensor transformation. By transforming viewpoint features into high-dimensional human gait pose features, the relationship between high-dimensional features between viewpoints can be searched, and concise viewpoint feature transformation is completed. In response to the problem of loss and low accuracy of human gait pose contour features caused by perspective deviation in multi-perspectives scenarios, this study uses a Siamese neural network as the architecture to calculate the correlation between multi-perspectives human

gait pose features, and uses it as the basis for view feature transformation to perform transformation enhancement processing. Based on the features of human gait pose contour from multi-perspectives, a dual channel Siamese module of CVT and convolution is studied and established. The specific structural diagram is shown in Figure 4.

In Figure 4, the Siamese network module is composed of two feature extraction networks for extracting multi-perspectives human gait pose information. In different modules, high-dimensional local features of human gait pose contours are extracted through convolutional channels, and global and local high-dimensional information is obtained through Mobile CVT (MCVT) channels. Finally, the above information is homogenized and fused for use in the current view of high-dimensional human gait pose feature tensors. In addition, the view set waiting for transformation is extracted using the MCVT module to extract the view feature tensor, which is based on the MCVT model, combined with Transformer and Conv Block, and replaces the local processing of convolution with deeper global processing. The internal details of the MCVT module are shown in Figure 5.

Figure 4
Structural diagram of improving the CVT model

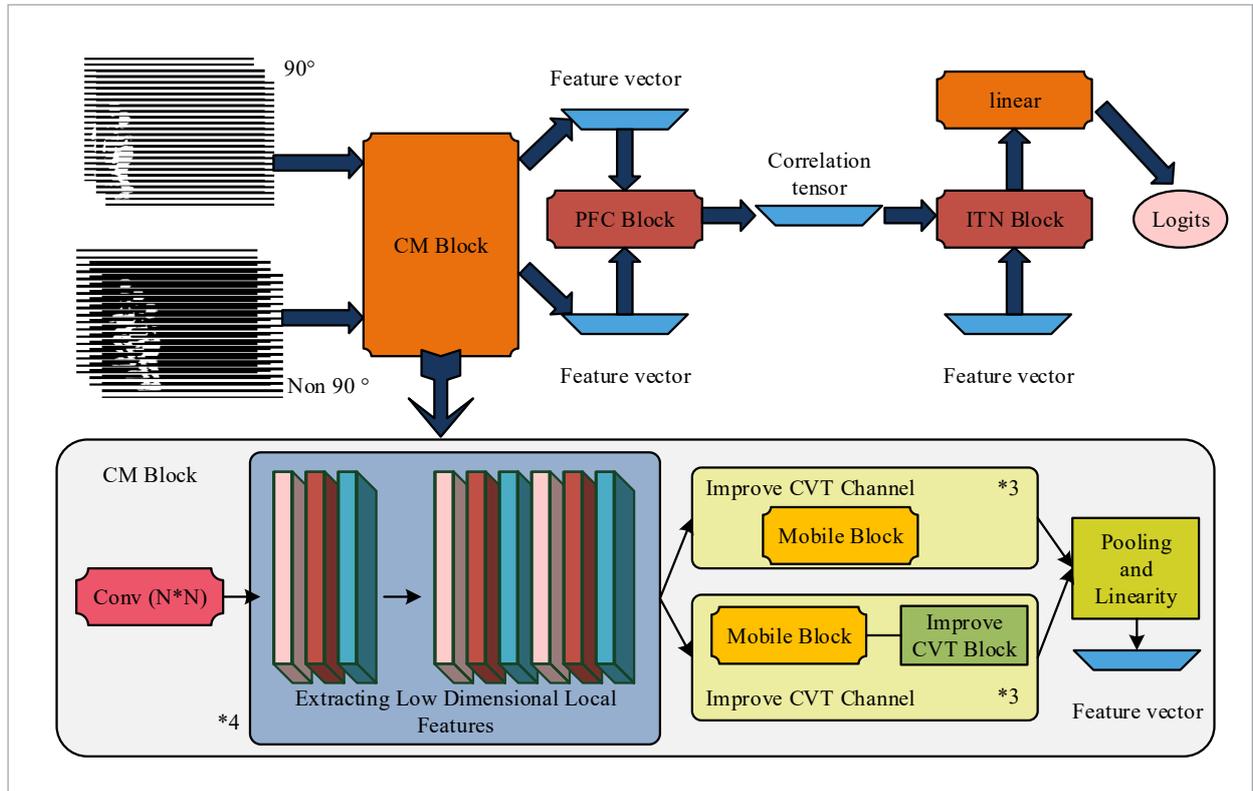
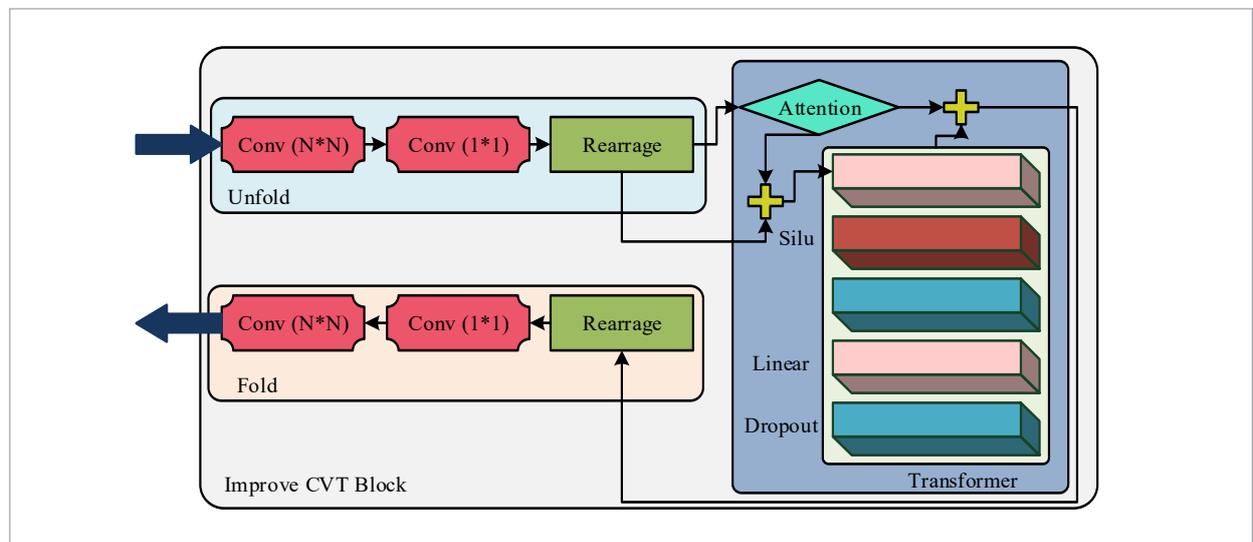


Figure 5
Display diagram of internal details of MCVT module



In Figure 5, Conv Block has two types: $N * N$ and pointwise convolution, consisting of one convolutional layer and one batch normalization layer each. At the same time, the activation functions of different modules use Sigmoid Linear Unit (Silu), as expressed in Equation (7).

$$Silu(x) = x \cdot Sigmoid(x). \tag{7}$$

In Equation (7), x is the input matrix and $Sigmoid(\square)$ is the Sigmoid function. The pooling layer uses global average pooling for calculation, as shown in Equation (8).

$$PL(x_w) = Avg(x_w). \tag{8}$$

In Equation (8), x_w represents the calculation area of the pooling layer, and $Avg(\square)$ represents the average value calculation function. To make the model more lightweight, some convolutional attributes are added to the Transformer module, while retaining the attributes that can be processed globally. A large global receptive field is divided into different image blocks through non overlapping methods, and the relationship between image blocks is encoded using Transformer. In addition, research proposes a Perspective

Feature Conversion (PFC) module and an Inverse Transformation (ITN) module, with specific structural diagrams shown in Figure 6.

In Figure 6, the PFC module is used to calculate the relationship between the perspective feature tensors obtained through the Siamese network, while treating the corresponding relationship tensor as the perspective conversion factor. The calculation is shown in Equation (9).

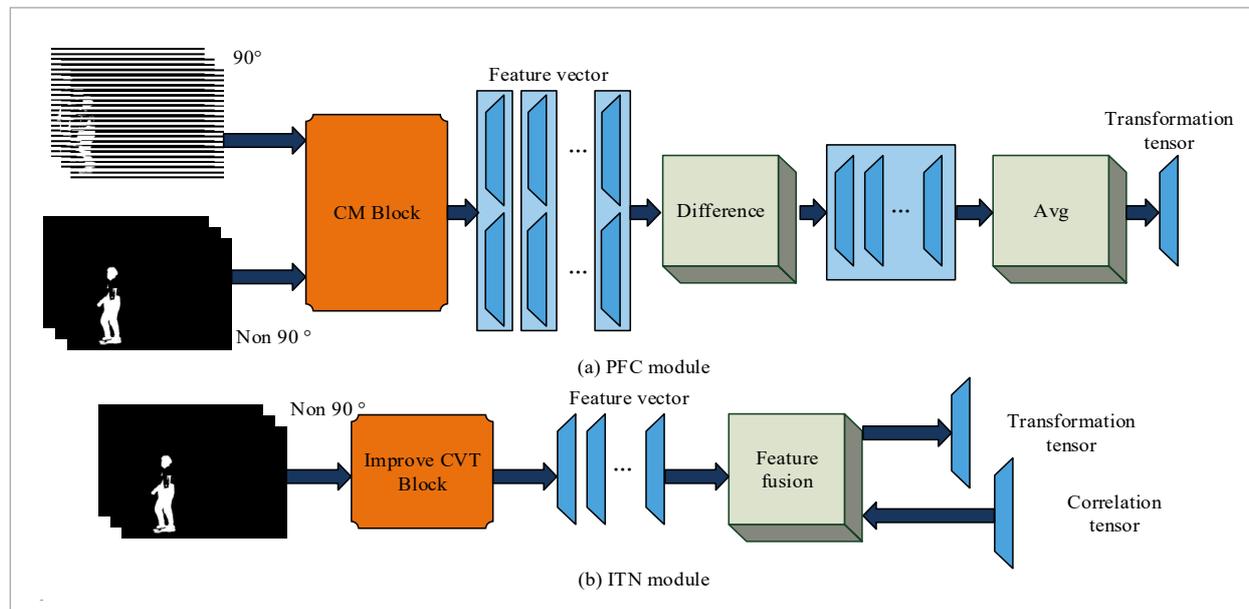
$$PFC(a,b) = \frac{\sum_{i=1}^{N'}(a_i, b_i)}{N'}. \tag{9}$$

In Equation (9), a and b represent the tensors of human gait pose feature from two perspectives, and N' represents the scale of the target perspective set that meets the conversion criteria. The ITN module is used for converting two high-dimensional perspective features, as expressed in Equation (10).

$$ITN(a,b) = a + PFC(a,b). \tag{10}$$

Finally, a more scientifically appropriate PGST is designed for the multi-perspectives recognition model based on improved CVT. To better calculate the difference in human gait pose features between

Figure 6
Structural diagram of PFC and ITN modules



90° and other perspectives, the study uses the same pre-training weights in the perspective feature relationship calculation module for calculation. Since the purpose of this module is not classification, the study will remove the pre-training weights of the last two layers of the extracted network module. In the training of the classification module, the study uses a more accurate 90° weight as the starting point, and then trains towards 0° and 180°, respectively. At the same time, the cross entropy loss function is used to calculate the loss, as shown in Equation (11) [12].

$$LOSS(OP, CS) = \omega_{CS} \left[-OP_{CS} + \log \left(\sum_j e^{OP_j} \right) \right]. \quad (11)$$

In Equation (11), CS and OP represent the actual labels and predicted results of the samples, OP_{CS} represents the corresponding elements of CS in OP , and ω_{CS} represents the weight parameters. Based on the above content, a multi-perspectives recognition model based on improved CVT can be obtained.

3. Results and Discussion

To verify the effectiveness and feasibility of the proposed method, the study first obtained estimation and comparison results of different human gait pose through ZMNC. Secondly, the performance of the ERHGP method based on CVT was analyzed. Then, the effectiveness and feasibility of the multi view ERHGP method based on improved CVT were analyzed. Finally, the results obtained from the research method were discussed.

3.1. Prediction Results of Different Human Gait Pose Based on Preprocessing

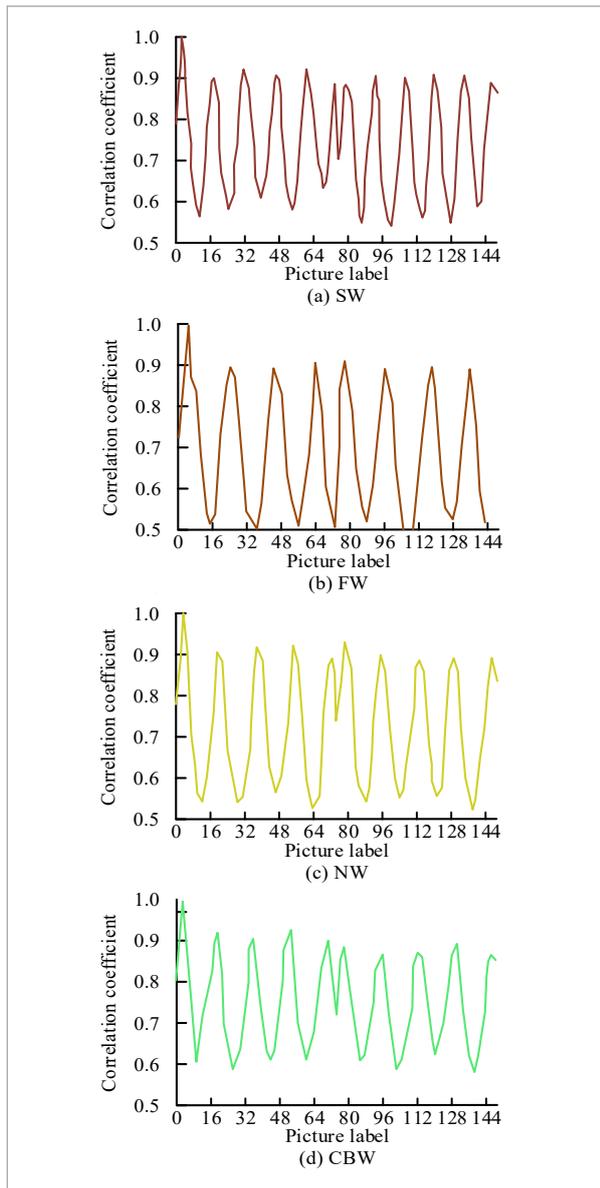
Firstly, the correlation coefficients of SW, FW, NW, and CBW human gait poses were calculated using ZMNC preprocessing operations, and the estimated comparison results were obtained. At the same time, the study also introduced simple step data for performance testing, such as cross step, stride, left step, and right step. Moreover, to better explore the application effect of research methods, the trained model was applied on both the MPII dataset and the COCO dataset. The former consists of approximately 25000 images, annotated with 16 joint point infor-

mation of human targets, and includes data for single frame single person, single frame multi person, and multi person poses. The latter is a dataset used for image recognition, with 80 object categories and different scene types. To more scientifically validate the performance of research methods, the study also compared the current mainstream methods, namely Gait ViT-based gait recognition method using visual transformers, Gait CNN ViT-based multimodal gait recognition using CNNs combined with visual transformers, and Gait AViT based on automatic gait analysis using visual transformers. The experimental mathematical equipment selected was a computer with an operating system of Windows 10 and 16GB of memory, and the software selected was Python 3.8 to build the model. The hyperparameters in the experiment are set as follows: the number of encoders and kenaf was 6, the number of attention heads was 10, the optimizer was Adam, and the number of node queries was 150. The model parameters are set as follows: the initialization learning rate and weight decay of the feature extraction network were set to 1^{-5} and 1^{-4} , respectively, the number of iterations was set to 200, and the number of data in each batch was 16. In addition, the study evaluated the accuracy, precision, and mean accuracy of commonly used methods. In the COCO dataset, the Object Keypoint Similarity (OKS) metric was used to measure the accuracy of the research method, while in the MPII dataset, the Percentage of Correct Keypoints (PCK) was used to evaluate the detection results, which actually calculated the percentage of predicted values within the normalized distance. The results are shown in Figure 7.

Figures 7(a)-(d) correspond to the correlation coefficient variation curves of SW, FW, NW, and CBW poses, respectively. The correlation coefficients of SW, FW, NW and CBW mainly fluctuated in the range of 0.54-0.92, 0.50-0.90, 0.51-0.93, and 0.57-0.93, respectively. Among them, the adjacent peaks represented a single legged human gait pose cycle, while there were differences in the gait pose cycles of different bipedal human postures. The corresponding periods for SW, FW, NW, and CBW were 27, 22, 24, and 23, respectively. Based on the above results, the cycle capacity corresponding to different human gait poses can be determined.

Figure 7

Comparison results of different human gait estimations based on preprocessed ZMNC operations



3.2. Performance Analysis of ERHGP Method Based on CVT

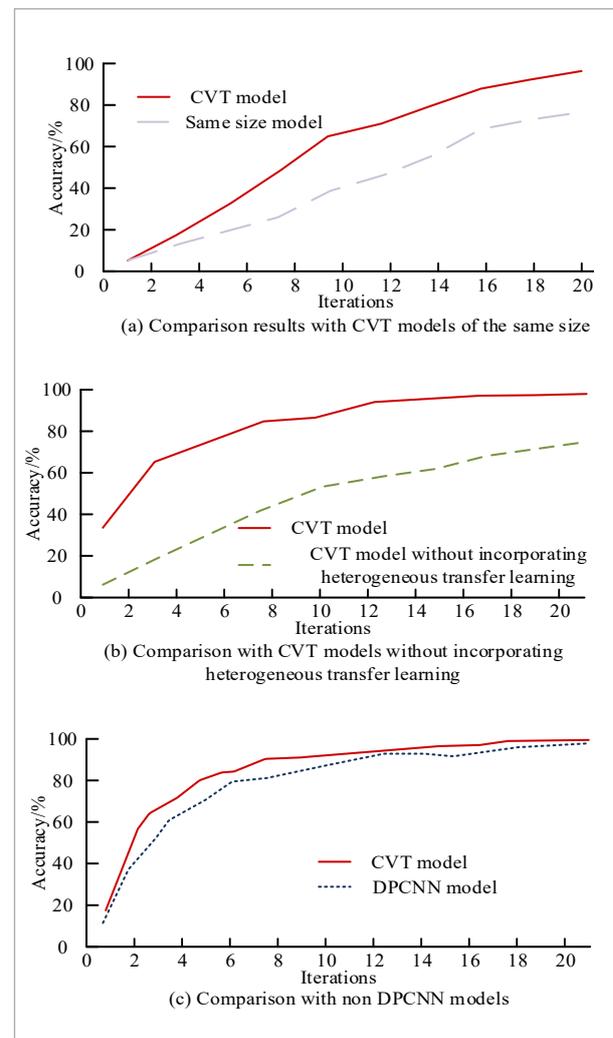
To test the performance of the CVT-based ERHGP method proposed in the study, experiments were conducted on a computer with Windows 10, 8GB memory and CPU 15-10400, and a model was built using Python 3.8. The dataset C was used for experiments,

with a total of 100346 infrared images, and was divided into training and validation sets in an 8:2 ratio. The experimental parameters were set as follows: learning rate, batch size, number of iterations, and embedding size were set to 0.001, 14, 20, and 32, respectively. The study compared the classification accuracy of CVT models and CVT basic models of the same size, and conducted 10 experiments with the average value as the final result. In addition, the study also conducted comparative experiments on CVT models without HTL and DPCNN models, as shown in Figure 8.

Figures 8(a)-(c) show the comparison results between the CVT model and the same size CVT basic

Figure 8

Comparison results of accuracy of different models



model, the CVT model without HTL, and the DPCNN model, respectively. Figure 8 shows that both the CVT model and the same size CVT basic model had a fast convergence speed before the 10th iteration. After this, the convergence acceleration of the two models continued to slow down, and the final average classification accuracy corresponded to 95.5% and 79.2%, respectively. The classification accuracy of the CVT model using HTL was significantly better than that of the CVT model without HTL, indicating that adding HTL can effectively improve the recognition performance of the CVT model. In the 8th iteration alone, the recognition accuracy of the CVT model exceeded 90%. In the 12th iteration, the recognition accuracy of the DPCNN model reached saturation, while the recognition accuracy of the CVT model continued to steadily increase, reaching 98.6% in the 20th iteration. The above results indicated that the CVT model proposed in the study performs better in terms of data fitting speed, stability, and accuracy.

3.3. Result Analysis of Multi-Perspectives ERHGP Method Based on Improved CVT

To verify the effectiveness and feasibility of the multi-perspectives ERHGP method based on improved CVT proposed in the study, 90° was set as the baseline perspective, and the human gait pose data from the other 10 perspectives were converted based on this perspective. In addition, due to the fact that the human gait pose data obtained in practical applications may not have the entire cycle of gait pose, and the corresponding features have strong randomness. Therefore, the study randomly dispersed the human gait pose groups to make the experimental results more in line with the complex scenarios of

practical applications. To evaluate the effectiveness of the improved CVT model and PGST method, experiments were conducted at angles, namely 55° and 125° , where there was significant feature loss due to perspective shift. The structure is shown in Figure 9.

Figures 9(a)-(b) show the training loss variation curves of the multi-perspectives ERHGP method based on improved CVT under 55° and 125° view angles, respectively. From Figure 9, it can be observed that the improved CVT-based multi-perspectives recognition model trained by the PGST method only needed 526 and 275 iterations to reach a stable state under 55° and 125° perspectives. Compared with the improved CVT model without PGST method training, the loss jump situation of this method was significantly improved, and the rate of loss reduction was significantly increased. The study was based on dataset B and trained on five angles with angles less than 90° . The results are shown in Figure 10.

Figures 10(a)-(e) show the comparison of training effects of the improved CVT method before and after the PGST method from 0° , 18° , 36° , 54° , and 72° perspectives, respectively. From Figure 10, under only 54° and 72° perspectives, the recognition rate of the improved CVT basic model showed a continuous increasing trend with the increase of iteration times. Under 0° , 18° , and 36° perspectives, the model exhibited significant fluctuations to varying degrees. The improved CVT model trained by the PGST method showed stable recognition rates in all angles below 90° , with an average recognition rate exceeding 98%. The study trained five angles with angles greater than 90° , as shown in Figure 11.

Figure 9

Training loss variation curve of multi-perspectives ERHGP method based on improved CVT from different perspectives

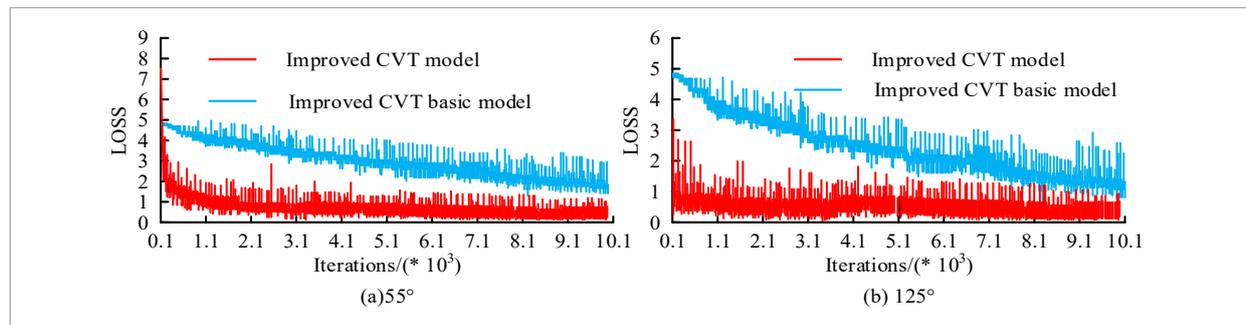


Figure 10

Comparison of the training effect of the improved CVT method before and after the gradual training of the viewing Angle under a viewing Angle of less than 90°

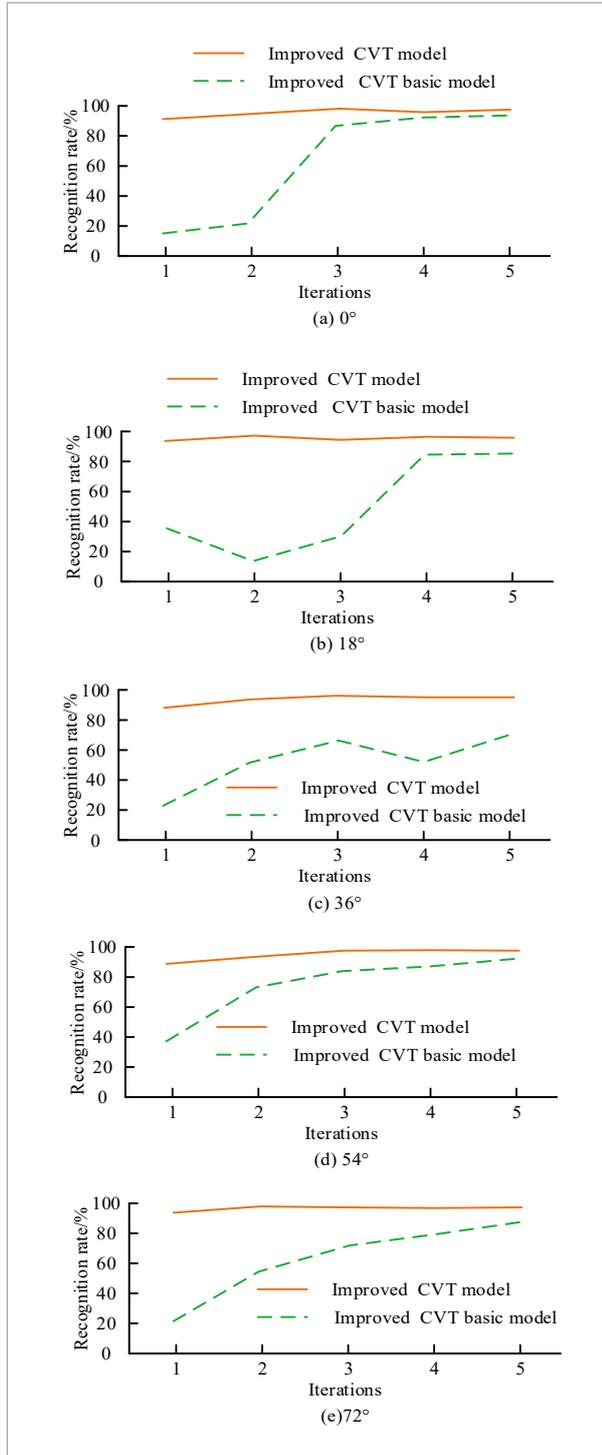
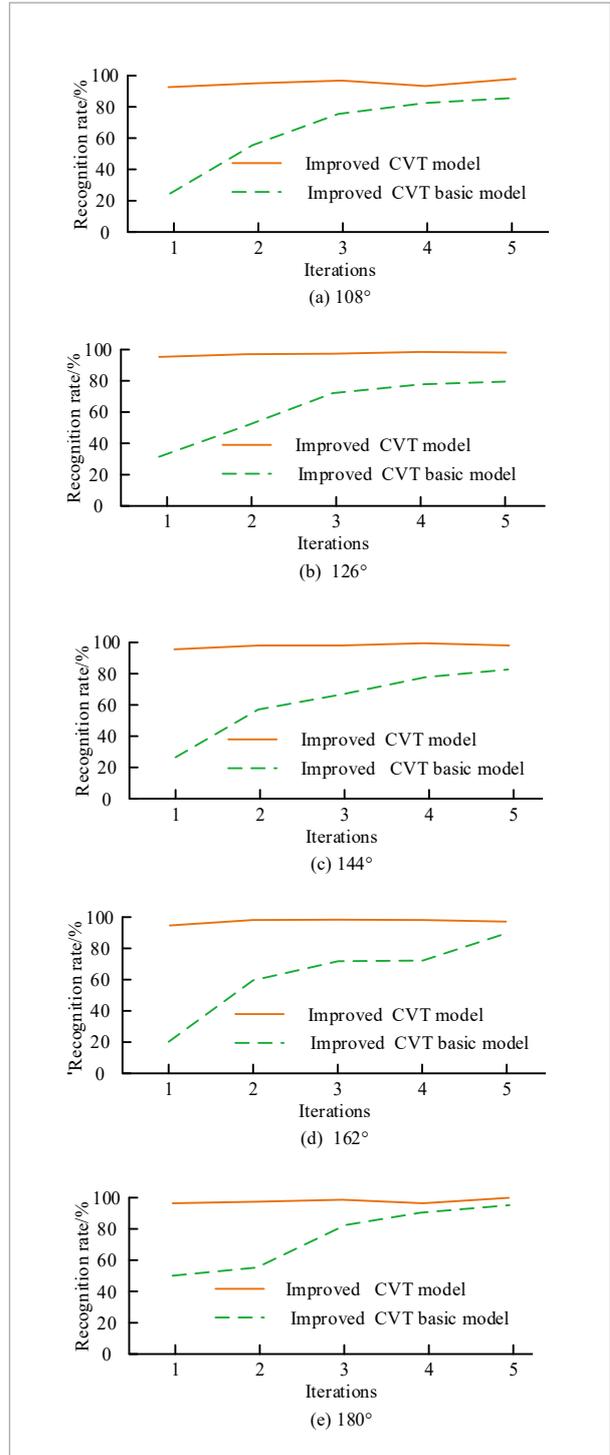


Figure 11

Comparison of the training effect of the improved CVT method before and after the gradual training of the angle of view at an angle greater than 90°



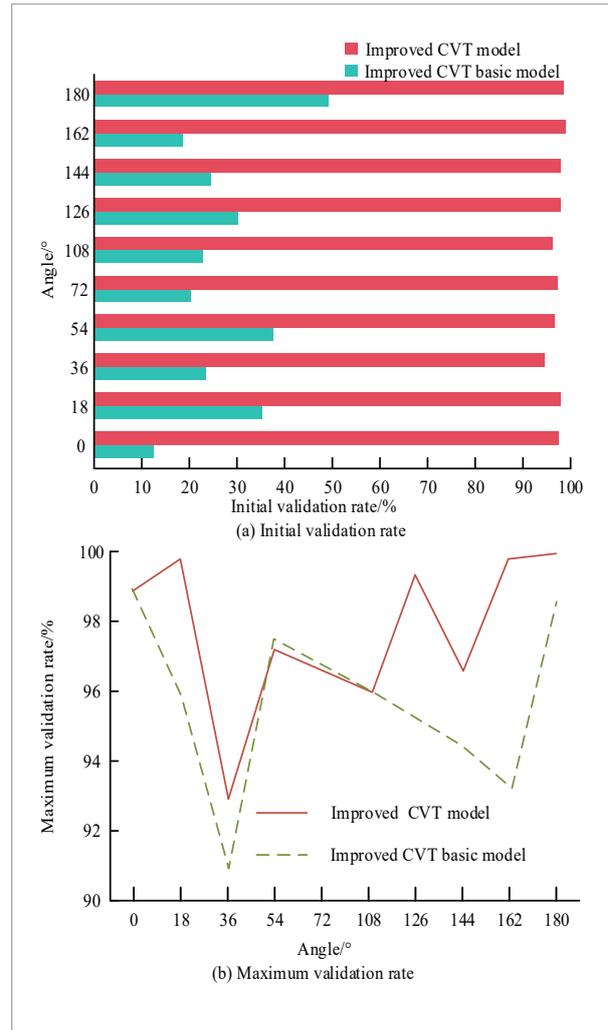
Figures 11(a)-(e) correspond to the training results of the improved CVT method before and after the PGST method from 108° , 126° , 144° , 162° , and 180° perspectives, respectively. From Figure 11, the change curve of the improved CVT basic model showed a continuous growth trend under five different perspectives greater than 90° , while the improved CVT model after PGST algorithm only showed slight fluctuations in recognition rate under two perspectives of 108° and 180° , with corresponding average recognition rates exceeding 97%. To more intuitively evaluate the improvement and stability effects of the PGST method on the model, a set of ablation experiments were conducted, and a comparison experiment was conducted between the initial and maximum validation rates of the improved CVT model. The results are shown in Figure 12.

Figures 12(a)-(b) show the comparison of the initial and maximum validation rate results of the improved CVT-based multi-perspectives recognition model with and without the PGST method, respectively. Figure 12 shows that the improved CVT model trained with PGST achieved high recognition rates in all 10 perspectives except for the 90° perspective, with an average initial recognition rate of 97.65%, which was 65.32% higher than the improved CVT initial model. The maximum validation rate of the improved CVT model trained by PGST was improved to a certain extent from different angles, indicating that the model has better robustness. To further validate the feasibility of the proposed model, the study compared it with current mainstream recognition models, namely Continuous Density Hidden Markov (CHM), Bessel Curve Fitting (BCF), CNN, and Deep Convolutional Constrained Boltzmann Machine (DCCBM). The corresponding average validation rate results and the reliability test results of the PGST method from a 36° perspective are shown in Figure 13.

Figures 13(a)-(b) show the average validation rate results of different recognition models from multi-perspectives and the reliability test results of the PGST method from a 36° perspective, respectively. The average validation rate of the improved CVT model reached 98.56%, while the average validation rates of the CHM, BCF, CNN, and DCCBM models were 72.51%, 73.45%, 74.62%, and 78.65%, respectively. In addition, the overall fluctuation of the change curve of the improved CVT basic model was significant, while the improved CVT model could achieve a stable rec-

Figure 12

Initial and maximum validation rate results of a multi-perspectives recognition model based on improved CVT with and without PGST method

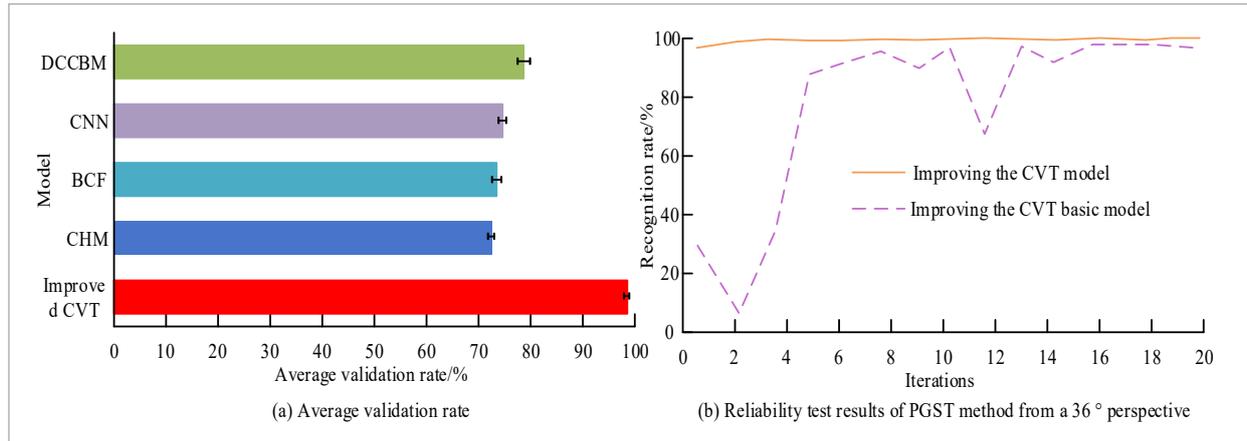


ognition accuracy only in the third iteration, with an average recognition accuracy of 99.13%. In dataset B, different recognition models were used to compare accuracy from different angles, and the results are shown in Table 2.

According to Table 2, among the 11 non-cross perspective offset perspectives, the average accuracy of the CHM, BCF, CNN, DCCBM, and improved CVT models were 74.18%, 74.13%, 73.47%, 75.83%, and 97.51%, respectively. Compared with other recognition models, the average improvement index of the improved

Figure 13

The average validation rate of different recognition models from multi-perspectives and the reliability test results of PGST method from a 36° perspective

**Table 2**

The accuracy results of different recognition models in dataset B at different angles

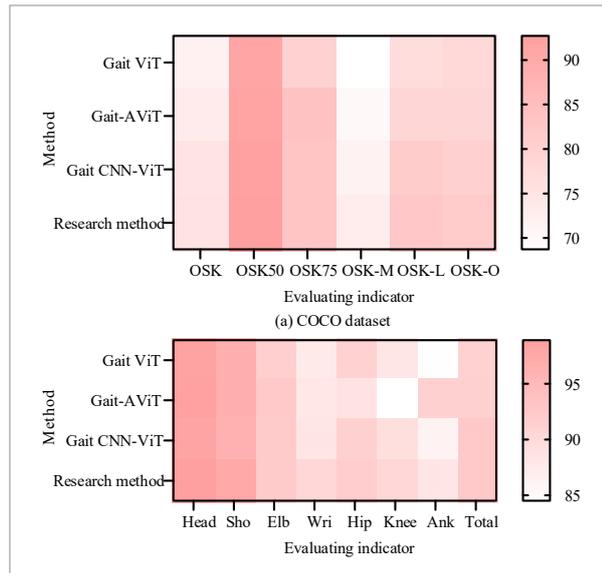
Angle/°	Recognition model accuracy/%				
	Improved CVT	CHM	BCF	CNN	DCCBM
0	98.04	75.21	73.47	75.23	73.26
18	97.24	96.72	76.46	73.12	74.21
36	95.28	72.62	72.35	73.12	77.95
54	98.07	70.21	83.46	72.62	79.26
72	97.73	74.23	76.53	71.56	75.62
90	97.52	69.19	71.82	76.12	75.45
108	97.86	69.72	70.26	74.26	75.958
126	97.26	73.69	74.03	70.86	77.26
144	96.95	71.27	72.32	74.68	73.92
162	97.83	69.92	74.36	73.35	76.25
180	98.79	73.25	70.42	73.26	74.95

CVT model was greater than 20%. In summary, the research proposed a multi-perspectives recognition model based on improved CVT, which can balance high-precision and human gait pose angle recognition extension, and enhance the possibility of human gait pose estimation and recognition in multi-perspectives application scenarios. To further analyze the robustness of the research method, different gait recognition methods were tested on the COCO dataset and MPII dataset, and the results are shown in Figure 14.

Figures 14(a)-(b) correspond to the OSK results of gait recognition algorithms on the COCO dataset and MPII dataset, respectively. Among them, OSK50 and OSK75 were the prediction accuracies of OSK under the conditions of 0.5 and 0.75, respectively. OSK-M and OSK-L were the prediction accuracies of medium-sized and large-sized targets, respectively. OSK-O was the average recall rate of the model prediction results. From Figure 14(a), the OSK result of the research method was 86.2%, and it showed the

Figure 14

Robustness results of different gait recognition methods based on COCO dataset and MPII dataset



best performance among other indicators. From Figure 14(b), both the research method and Gait CNN-ViT had convolutional modules that can correct erroneous results through feedback prediction variation. The Gait-AViT method had a better effect on ankle joint detection, with a rate of 84.6%. Overall, the improvement effect of the research method was better, with an increase of about 2.2% compared to the mainstream Gait ViT method. Finally, to quantify the accuracy statistical results of the research method on different datasets, paired sample t-tests were used for analysis, and the results are shown in Table 3.

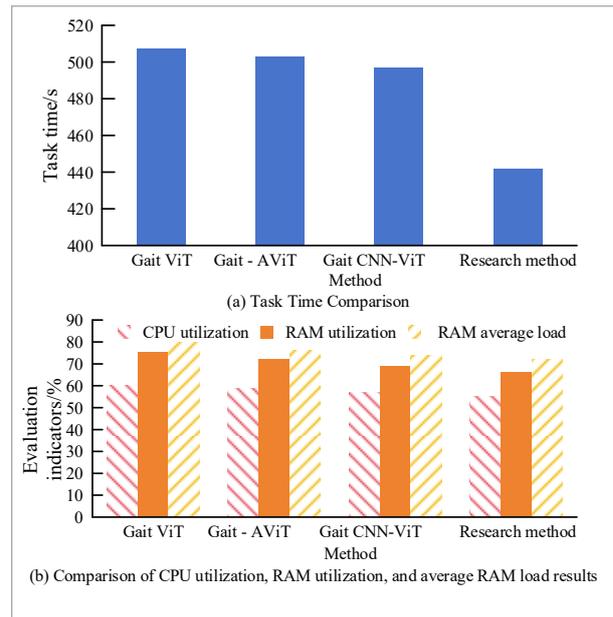
According to Table 3, there was no statistically significant difference in the recognition accuracy results of the research method between the two datasets, indicating that the research method is suitable for gait recognition of different scales and types and has excellent application effects. To analyze the computational efficiency of various gait recognition algorithms, the study

introduced task completion time, Central Processing Unit (CPU) utilization, Random Access Memory (RAM) utilization, and RAM average load for evaluation. Each algorithm underwent 20 simulation experiments to ensure fairness of the experiments. The comparison of computational efficiency results of different gait recognition algorithms is shown in Figure 15.

Figure 15(a) shows a comparison of task completion times for different gait recognition algorithms. It can be seen that compared to Gait CNN ViT, the research method reduces the average task completion time by 12%, greatly improving the computational efficiency of task completion. Figure 15(b) shows the CPU utilization, RAM utilization, and average RAM load results of different gait recognition algorithms. It can be observed that the research method corresponds to CPU utilization, RAM utilization, and average RAM

Figure 15

Comparison of computational efficiency results of different gait recognition algorithms

**Table 3**

Paired sample testing results of recognition accuracy of research methods on different datasets

Paired samples	Pairing difference			t-value	Significance (dual tailed)
	Average value	Standard deviation	Mean standard error		
COCO-MPII	0.034411	0.11249	0.03389	1.00700	0.41200

load of 55%, 66%, and 72%, respectively, while Gait ViT's various indicators correspond to 60%, 75%, and 80%. The above results confirm that the research method enables full utilization of resources and effectively reduces the average RAM load.

3.4. Discussion

As an emerging method in the field of biometric technology, human gait pose estimation and recognition are gradually entering industries such as transportation, security, and industry for related applications due to their advantages of not requiring hard coordination, supporting long-distance recognition, and strong environmental adaptability. It has a very broad development prospect. However, due to the dynamic nature of the ERHGP method, its recognition process is more complex compared to traditional static biometric recognition. At present, the application of this method in China is still in the exploratory stage. Therefore, the study first selected publicly available datasets for introduction and proposed a ZMNC preprocessing method for estimating the gait attitude period. Then, the HLT method was introduced, and a human gait pose periodic group containing continuous time features was considered, and an ERHGP method based on CVT was designed. Finally, based on this, the PGST method and feature tensor transformation were optimized to obtain a multi-perspectives recognition model based on improved CVT.

The study first calculated the correlation coefficients between the four poses of SW, FW, NW, and CBW, and obtained the corresponding change curves to determine the cycle capacities corresponding to different human gait poses, which were 27, 22, 24, and 23, respectively. Secondly, the study compared the performance of the ERHGP method based on CVT with the same size CVT basic model, the CVT model without using HLT method, and the DPCNN model. The results are as follows: firstly, the CVT model and the same size CVT basic model both had a fast convergence rate before the 10th iteration. Afterwards, the convergence acceleration of the two models continued to slow down, and the final average classification accuracy corresponded to 95.5% and 79.2%, respectively. The second is that the classification accuracy of the CVT model using the HLT method was significantly improved, indicating that the HLT method can effectively improve the network's accuracy and convergence speed. The third

issue is that the DPCNN model experienced oversaturation during the 12th iteration, while the CVT model continued to grow, achieving a recognition accuracy of 98.6% in the 20th iteration. The above results confirm the effectiveness and feasibility of the CVT model, and its performance is significantly better than traditional methods, providing a new solution for the application of small sample datasets.

Finally, experiments were conducted to improve the performance, reliability, and scientificity of the CVT model, and the following results were obtained. The first was that in the other 10 perspectives except for 90°, the improved CVT model only showed slight fluctuations in recognition rate in 0°, 18°, 36°, 108°, and 180° perspectives, but the average recognition rate in each perspective exceeded 97%. The recognition rate variation curves of the CVT model before improvement showed significant fluctuations in various perspectives, indicating that the model's performance was unstable. The second was the comparison of the initial and maximum validation rates before and after the improvement of the CVT model. However, the improved CVT model could still achieve high recognition rates in various perspectives, with an average initial recognition rate of 97.65%, which is 65.32% higher than before the improvement. This means that the improved CVT model has stronger robustness. The third was the average validation rate of different recognition models from multiple perspectives and the reliability test results of the PGST method from a 36° perspective. The highest average validation rate of the improved CVT model was 98.56%, while the average validation rates of the CHM, BCF, CNN, and DCCBM models were 72.51%, 73.45%, 74.62%, and 78.65%, respectively. The improved CVT model could achieve a stationary recognition accuracy of 99.13% only in the third iteration, while the variation of the improved CVT basic model was significant. The fourth was the internal recognition accuracy results of multiple offset perspectives without cross view. The average accuracy of CHM, BCF, CNN, DCCBM, and improved CVT models were 74.18%, 74.13%, 73.47%, 75.83%, and 97.51%, respectively. Based on the above results, it can be concluded that the proposed method has good accuracy, robustness, and convergence speed, expanding the possibility of efficient recognition of human gait pose from multi-perspectives in complex application scenarios.

4. Conclusions

In response to the high difficulty and strong randomness of human gait pose estimation and recognition, a series of preprocessing operations were first carried out to obtain the correlation feature changes of asynchronous gait posture estimation. Then, the HTL method was introduced to design the ERHGP method based on CVT. Finally, the PGST method was used to optimize and obtain a multi-perspectives recognition model based on improved CVT. The experimental results showed that the classification accuracy of the CVT model using HTL was significantly better than that of the CVT model without HTL, indicating that incorporating HTL can effectively improve the recognition performance of the CVT model. The recognition accuracy of the CVT model exceeded 90% in the 8th iteration alone, and reached 98.6% in the 20th iteration, while the recognition accuracy of the DPCNN model reached saturation in the 12th iteration. In comparison experiments with current mainstream algorithms, the average validation rate and accuracy of the improved CVT model reached 98.56% and 97.51%, respectively, while the average validation rate and accuracy of the CHM, BCF, CNN, and DCCBM models corresponded to 72.51% and 74.18%, 73.45% and 74.13%, 74.62% and 73.47%, 78.65% and 75.83%, respectively. The above results are due to the integration of the twin neural network idea into the research method and the improvement achieved through the PGST method, which effectively obtains information from different perspectives and greatly improves the

recognition accuracy and robustness of the model application in various environments. In summary, the research method innovatively designs a multi-scale aggregation module to ensure the acquisition of more features while effectively analyzing and reducing the existing information redundancy problem. However, there are still limitations in the research, and in practical applications, research methods may still encounter missed detection problems. Therefore, targeted network repair missed detection targets can be designed in further research in the future.

Funding

The research is supported by 2023 Jiangsu Province Industry-University-Research Cooperation Project BY20230641 System Development of Behavior Detection of Construction Personnel Based on Human Posture Estimation; 2021 Project of Jiangsu Province Education Science "14th Five-Year Plan" C-c/2021/03/30 Research on Supervision Model of Open Education Based on Machine Learning in the Perspective of Artificial Intelligence; General Project of Philosophy and Social Science Research in Jiangsu Higher Education Institutions 2022SJYB0852 Research on Early Warning Model of Distance Education Based on Deep Learning in the Context of Artificial Intelligence; Jiangsu Province Higher Education Institutions Basic Science (Natural Science) Research Project 20KJB630009 "Constructing Evaluation Indicator System of Scientific and Technological Achievements Based on Policy Analysis Perspective and Empirical Research".

References

- Albarran-Diego, C., Poyales, F., Lopez-Artero, E., Garzon, N., Garcia-Montero, M. Interoctular Biometric Parameters Comparison Measured with Swept-Source Technology. *International Ophthalmology*, 2022, 42(1), 239-251. <https://doi.org/10.1007/s10792-021-02020-8>
- Bakc, H., Santopietro, M., Guest, R. Activity-Based Electrocardiogram Biometric Verification Using Wearable Devices. *IET Biometrics*, 2022, 12(1), 38-51. <https://doi.org/10.1049/bme2.12105>
- Bhosle, K., Musande, V. Evaluation of Deep Learning CNN Model for Recognition of Devanagari Digit. *Artificial Intelligence Applications*, 2023, 1, 2, 114-118. <https://doi.org/10.47852/bonviewAIA3202441>
- Dowling, P., Brewster, C., Holland, P. HRM and the Smart and Dark Side of Technology. *Asia Pacific Journal of Human Resources*, 2022, 60(1), 62-78. <https://doi.org/10.1111/1744-7941.12319>
- Elavarasi, S. A., N. S., P. S., R. N. Gait Posture Recognition and Detection Using Deep Learning Algorithms. *International Conference on Computing and Data Science (ICCDs)*, Chennai, India, April 2024, 1-6. <https://doi.org/10.1109/ICCDs60734.2024.10560412>
- Gao, S., Yun, J., Zhao, Y., Liu, L. Gait-D: Skeleton-Based Gait Feature Decomposition for Gait Recognition. *IET Computer Vision*, 2022, 16(2), 111-125. <https://doi.org/10.1049/cvi2.12070>

7. Hasanvand, M., Nooshyar, M., Moharamkhani, E., Seylari, A. Machine Learning Methodology for Identifying Vehicles Using Image Processing. *Artificial Intelligence Applications*, 2023, 1(3), 170-178. <https://doi.org/10.47852/bonviewAIA3202833>
8. Hemamalini, S., Sharma, O. P. Hybrid Structures-Based Face Recognition Method Using Artificial Neural Network. *International Journal of Advanced Research in Information Technology and Computing*, 2022, 13(1/2), 12-23. <https://doi.org/10.5958/0975-8089.2022.00002.1>
9. Inturi, A. R., Manikandan, V. M., Garrapally, V. A Novel Vision-Based Fall Detection Scheme Using Keypoints of Human Skeleton with Long Short-Term Memory Network. *Arabian Journal for Science and Engineering*, 2023, 48(2), 1143-1155. <https://doi.org/10.1007/s13369-022-06684-x>
10. Kulikajevas, A., Maskeliūnas, R., Damasevicius, R. Detection of Sitting Posture Using Hierarchical Image Composition and Deep Learning. *PeerJ Computer Science*, 2021, 7, e442. <https://doi.org/10.7717/peerj-cs.442>
11. Kurnia, J. C., Chaedir, B., Wijayanta, A. T. Convective Heat Transfer Enhancement of Laminar Herschel-Bulkley Non-Newtonian Fluid in Straight and Helical Heat Exchangers with Twisted Tape Inserts. *Industrial & Engineering Chemistry Research*, 2022, 61(1), 814-844. <https://doi.org/10.1021/acs.iecr.1c02234>
12. Li, K., Wang, B., Tian, Y., Qi, Z. Fast and Accurate Road Crack Detection Based on Adaptive Cost-Sensitive Loss Function. *IEEE Transactions on Cybernetics*, 2023, 53(2), 1051-1062. <https://doi.org/10.1109/TCYB.2021.3103885>
13. Liu, J., Yu, S., Zhao, X., Sun, X., Meng, Q., Liu, S., Xu, Y., Lv, C., Li, J. Resolution Enhancement of Tongue Tactile Image Based on Deconvolution Neural Network. *Journal of Texture Studies*, 2022, 54(4), 456-469. <https://doi.org/10.1111/jtxs.12778>
14. Mathivanan, B., Perumal, P. Gait Recognition Analysis for Human Identification Analysis-A Hybrid Deep Learning Process. *Wireless Personal Communications*, 2022, 126(1), 555-579. <https://doi.org/10.1007/s11277-022-09758-z>
15. Miller, K. W. Facial Recognition Technology: Navigating the Ethical Challenges. *Computer*, 2023, 56(1), 76-81. <https://doi.org/10.1109/MC.2022.3206840>
16. Ogundokun, R. O., Maskeliūnas, R., Misra, S., Damasevicius, R. A Novel Deep Transfer Learning Approach Based on Depth-Wise Separable CNN for Human Posture Detection. *Information*, 2022, 13(11), 520. <https://doi.org/10.3390/info13110520>
17. Ozturk, M. Z., Wu, C., Wang, B., Liu, K. J. R. GaitCube: Deep Data Cube Learning for Human Recognition with Millimeter-Wave Radio. *IEEE Internet of Things Journal*, 2022, 9(1), 546-557. <https://doi.org/10.1109/JIOT.2021.3083934>
18. Parrany, A. M., Mirzaei, M. A New Image Processing Strategy for Surface Crack Identification in Building Structures Under Non-Uniform Illumination. *IET Image Processing*, 2022, 16(2), 407-415. <https://doi.org/10.1049/ipr2.12357>
19. Purohit, J., Dave, R. Leveraging Deep Learning Techniques to Obtain Efficacious Segmentation Results. *Advances in Artificial Intelligence and Engineering Sciences*, 2023, 1(1), 11-26. <https://doi.org/10.47852/bonviewAAES32021220>
20. Rahi, B. H., Qi, M. Novel Architecture for Human Re-Identification with a Two-Stream Neural Network and Attention Mechanism. *Computing and Informatics*, 2022, 41(4), 905-930. https://doi.org/10.31577/cai_2022_4_905
21. Reinders, S., Guan, Y., Ommen, D., Newman, J. Source-Anchored, Trace-Anchored, and General Match Score-Based Likelihood Ratios for Camera Device Identification. *Egyptian Journal of Forensic Sciences*, 2022, 67(3), 975-988. <https://doi.org/10.1111/1556-4029.14991>
22. Sun, X., Hu, T., Ma, L., Jin, W. The Encryption and Decryption Technology with Chaotic Iris and Compressed Sensing Based on Computer-Generated Holography. *Journal of Optics-UK*, 2022, 51(1), 124-132. <https://doi.org/10.1007/s12596-021-00750-7>
23. Wang, H., Song, Y., Huo, L., Chen, L., He, Q. Multiscale Object Detection Based on Channel and Data Enhancement at Construction Sites. *Multimedia Systems*, 2023, 29(1), 49-58. <https://doi.org/10.1007/s00530-022-00983-x>
24. Wei, S., Liu, R., Xu, P., Long, Y., Yang, C., Ying, J. A Sparse Optimisation Method Based on Cross-Correlation Function for Chirp Signals. *Electronics Letters*, 2022, 58(3), 127-129. <https://doi.org/10.1049/ell2.12367>
25. Wei, Y., Zeng, A., Zhang, X., Huang, H. RAG-Net: ResNet-50 Attention Gate Network for Accurate Iris Segmentation. *IET Image Processing*, 2022, 16(11), 3057-3066. <https://doi.org/10.1049/ipr2.12538>
26. Wen, J., Shen, Y., Yang, J. Multi-View Gait Recognition Based on Generative Adversarial Network. *Neural Processing Letters*, 2022, 54(3), 1855-1877. <https://doi.org/10.1007/s11063-021-10709-1>

27. Xiang, Z., Wu, H., Zhou, D. Metallic Debossed Characters Industrial Online Non-Segmentation Identification Based on Improved Multi-Scale Image Fusion Enhancement and Deep Neural Network. *IET Image Processing*, 2022, 16(3), 852-868. <https://doi.org/10.1049/ipr2.12391>
28. Yao, R., Shao, W., Jiang, X., Yu, T. Wind Speed Retrieval from Chinese Gaofen-3 Synthetic Aperture Radar Using an Analytical Approach in the Nearshore Waters of China's Seas. *International Journal of Remote Sensing*, 2022, 43(7/8), 3028-3048. <https://doi.org/10.1093/gji/ggab534>
29. Yeo, S. S., Rho, S., Kim, H., Safdar, J., Zia, U., Durrani, M. Y. A Triplet-Branch Convolutional Neural Network for Part-Based Gait Recognition. *Transactions on Signal Processing*, 2023, 47(11), 2027-2047. <https://doi.org/10.32604/csse.2023.040327>
30. Zhang, Z., Li, Y., Duan, J., Duan, Y., Guo, Y., Cao, Y., and Rehtanz, C. A Non-Intrusive Load State Identification Method Considering Non-Local Spatiotemporal Feature. *IET Generation, Transmission & Distribution*, 2022, 16(4), 792-803. <https://doi.org/10.1049/gtd2.12330>



This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 (CC BY 4.0) License (<http://creativecommons.org/licenses/by/4.0/>).