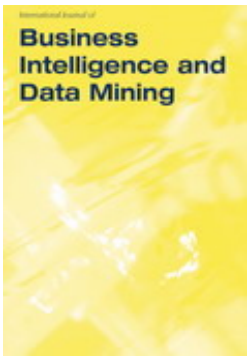


International Journal of Business Intelligence and Data Mining

ISSN online: 1743-8195 - ISSN print: 1743-8187
https://www.inderscience.com/ijbidm

Design of lightweight human pose estimation network for rehabilitation training

Qingyun Yang, Gaigai Zhang

DOI: 10.1504/IJBIDM.2026.10080082

Article History:

Received:	23 March 2026
Last revised:	28 April 2026
Accepted:	02 June 2026
Published online:	07 September 2026

Design of lightweight human pose estimation network for rehabilitation training

Qingyun Yang*

College of Rehabilitation Medicine,
North Henan Medical University,
He nan, 453003, China
Email: 13063155@nhmu.edu.cn
*Corresponding author

Gaigai Zhang

Department of Intelligent Medical Engineering,
North Henan Medical University,
He nan, 453003, China
Email: 1169905504@qq.com

Abstract: Given the resource constraints of terminal devices, existing pose estimation networks, characterised by high computational complexity, struggle to satisfy the simultaneous demands of real-time performance and accuracy in rehabilitation training. Therefore, this paper proposes a lightweight human pose estimation network specifically designed for rehabilitation training. Firstly, laser triangulation is used to obtain the target image, and precise mapping from 3D to 2D space is achieved through coordinate transformation. Secondly, design an RMPE tiny lightweight network, introduce G-Bottleneck module to compress parameter quantity, and integrate Sa-ECA attention mechanism to enhance feature interaction. Finally, the Huber loss function is used to optimise training and improve convergence speed and robustness. The experimental results show that the method proposed in this paper maintains an overall accuracy of over 96% in human pose estimation testing, with most samples approaching or exceeding 98%, consistently maintains between 41.1 FPS and 44.4 FPS in ten inference speed tests.

Keywords: rehabilitation training; lightweight network; human pose estimation; attention mechanism.

Reference to this paper should be made as follows: Yang, Q. and Zhang, G. (2026) 'Design of lightweight human pose estimation network for rehabilitation training', *Int. J. Business Intelligence and Data Mining*, Vol. 28, No. 10, pp.85–102.

Biographical notes: Qingyun Yang received his Master's degree in Biomedical Engineering from Henan Medical University. He is currently a Laboratory Technician at the College of Rehabilitation Medicine, North Henan Medical University. His research interests include intelligent education technology, biomedical information technology, and their applications in rehabilitation medicine.

Gaigai Zhang received her Master's degree in Biomedical Engineering from Henan Medical University. She is currently a Lecturer in the College of Intelligent Medical Engineering, North Henan Medical University. Her research interests include biomedical information technology, electronic science technology, and virtual reality technology.

1 Introduction

As the population ages and the number of patients with chronic diseases continues to rise, the societal demand for rehabilitation training has become increasingly prominent (Zhang et al., 2025a). Traditional rehabilitation guidance relies on on-site observation by professional physicians, which has problems such as limited human resources, strong subjectivity in evaluation, and difficulty in long-term tracking. The intelligent assisted rehabilitation system achieves automatic action assessment through computer vision technology, which has become an effective way to solve the above contradictions (Zhang et al., 2025b; Zhao et al., 2024; Yan et al., 2025). Among them, human pose estimation serves as the core link, which can extract the coordinates of human keypoints from images or videos, and then quantitatively analyse the amplitude, symmetry, and coordination of actions (Xu et al., 2023). Although mainstream pose estimation models achieve high accuracy, they involve a large number of parameters and high computational costs, which hinders their deployment on resource-constrained terminal devices, such as home rehabilitation robots, mobile tablets, or embedded monitors. The high latency and energy consumption limit the widespread application of technology in home and community scenarios. Therefore, designing a lightweight network architecture that balances accuracy and efficiency is of great practical significance. This can not only promote the development of rehabilitation training in the direction of personalisation, remote and intelligence, but also provide key technical support for the implementation of edge computing in the medical and health field (Bharathi, et al., 2024; Liu, et al., 2024a).

Liu et al. (2024b) proposed a human pose estimation method based on optimised multi-scale feature fusion, which enhances multi-scale feature transfer through transpose convolution and mixed dilated convolution. However, in the upsampling process, the superposition of transpose convolution is prone to generate checkerboard artifacts, while mixed dilated convolution leads to periodic loss of local features due to sparse sampling points, especially at joint edges, which limits the ability to capture fine movements in rehabilitation training. Yang et al. (2025) proposed a 3D human pose estimation method based on parallel multi-scale spatiotemporal graph convolutional networks. Although the diagonally dominant spatiotemporal attention graph convolution enhances the transmission of joint information, the superposition of parallel multi-scale sub networks and multi-scale feature cross fusion modules leads to an exponential increase in model parameters, which not only increases computational overhead, but also easily leads to overfitting when the training samples are limited due to the overly complex graph topology, reducing generalisation performance in real rehabilitation scenarios. Liu et al. (2025) proposed a human pose estimation method based on lightweight high-resolution networks. Although Ghost convolution and decoupling fully connected attention mechanism reduce the number of parameters, decoupling attention mechanism decomposes spatial long-range dependencies into independent modelling in two

directions, weakening the joint dependency relationship between joints under complex occlusion; At the same time, the redundant features generated by Ghost convolution lose some high-frequency details in linear transformation, resulting in significant shaking of key points in fast rehabilitation actions.

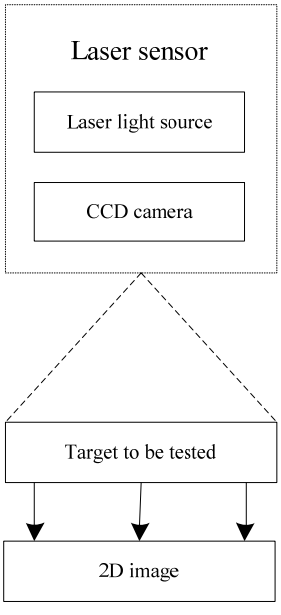
In summary, the existing human pose estimation methods have not yet achieved an ideal balance between accuracy and efficiency, especially in the resource limited and highly demanding application scenario of rehabilitation training, where the contradiction between model complexity and deployment feasibility is particularly prominent. It should be pointed out that the limitations of existing lightweight pose estimation networks in rehabilitation training scenarios go far beyond computational complexity. Firstly, in terms of motion amplitude sensitivity, rehabilitation actions such as shoulder abduction and knee flexion and extension involve a wide range of joint movements. However, existing lightweight networks commonly use larger downsampling factors to pursue compression ratios, which makes it difficult to effectively characterise small joint displacements in low resolution feature maps, especially for fine joints such as hands and ankles where localisation errors are significantly amplified. Secondly, in terms of occlusion robustness, patients' limbs are often occluded by instruments (such as handrails, elastic bands) or their own limbs during rehabilitation training. Existing methods rely on global attention or fixed receptive field convolution, which makes it difficult to reconstruct the spatial dependency relationship of occluded joint points, resulting in the failure of keypoint detection. Finally, in terms of computational constraints in actual rehabilitation terminals, although existing lightweight methods have achieved high FPS on GPUs, they have not fully considered the storage bandwidth limitations and power consumption budgets of embedded devices such as Jetson Nano and Raspberry Pi. The irregular memory access patterns of Ghost convolution or depthwise separable convolution can cause additional memory access delays in practical deployment, making it impossible to fully convert theoretical compression ratios into actual inference speed benefits. Therefore, this article focuses on rehabilitation training scenarios and designs a posture estimation network RMPE tiny that balances lightweight and high-precision. It systematically optimises image acquisition, feature extraction, attention mechanism, and loss function, aiming to provide efficient and reliable technical support for intelligent rehabilitation systems.

2 Rehabilitation training image acquisition based on laser vision sensing

Laser vision sensing imaging relies on the principle of laser triangulation, using sensors to project laser and form laser stripes on the surface of the object to be detected. By analysing the geometric characteristics of stripes and their corresponding positions in the camera imaging plane, combined with coordinate transformation processing, two-dimensional image information can be reconstructed. This process can effectively capture target images during rehabilitation training, providing basic support for the recognition of multi-target human postures in complex motion scenes. The working mechanism of laser vision sensing imaging is shown in Figure 1. In this process, a laser projector emits a structured light stripe onto the human body surface. The stripe deforms according to the surface topography. A calibrated camera captures the deformed stripe from a different viewpoint. Based on the laser triangulation principle, the displacement of the stripe in the image plane is directly proportional to the depth variation. This

establishes a geometric relationship between the image coordinates and the 3D surface coordinates of the rehabilitation training target.

Figure 1 Laser vision sensing imaging mechanism



According to the rehabilitation training images obtained by laser vision sensing, the perspective projection model is commonly used in the acquisition process. If the distance from the target to the camera significantly exceeds the camera focal length, the model can be approximated as a pinhole model, expressed in the following specific form:

$$\frac{1}{\alpha}+\frac{1}{b}=\frac{1}{g} \tag{1}$$

In the formula, g is the focal length parameter; b represents the distance value; α refers to the variable of object distance.

For rehabilitation training images based on laser vision sensing, their generation relies on coordinate transformation processing of the collected images. The specific description of this processing process is as follows:

1 Conversion between camera coordinates and world coordinates

Given a point q in the world coordinate system, its coordinates are (x_1, y_1, z_1). After a specific coordinate transformation, the corresponding coordinates of the point in the camera coordinate system become (x_2, y_2, z_2). The conversion relationship between the above two coordinates can be described by the following formula:

$$\begin{bmatrix} x_1 \\ y_1 \\ z_1 \end{bmatrix} = T \begin{bmatrix} x_2 \\ y_2 \\ z_2 \end{bmatrix} + gY \tag{2}$$

In the formula, Y represents the translation vector, and T represents the orthogonal matrix.

2 Conversion between plane coordinates and camera coordinates

If the coordinates of point q in the camera coordinate system are (x_2, y_2, z_2), and its corresponding coordinates on the image plane are (x, y), then the transformation between these two coordinates can be achieved through the following equation:

$$\begin{cases} \dfrac{x}{x_2} = \dfrac{1}{z_2} \\ \dfrac{y}{y_2} = \dfrac{1}{z_2} \end{cases} \tag{3}$$

According to the coordinate transformation process described above, spatial points in the world coordinate system can be mapped to the camera coordinate system, and then projected onto the image plane coordinate system, resulting in a two-dimensional image $C = (x, y)$ of the rehabilitation training target based on laser vision sensing.

The above coordinate mapping ensures spatial fidelity, which is a prerequisite for accurate pose estimation. However, to meet the resource constraints of terminal devices, a lightweight network architecture must be designed to process these images efficiently. Therefore, the following section details the proposed RMPE tiny network.

3 Design of lightweight human pose estimation network for rehabilitation training

3.1 Improved RMPE tiny network architecture

Design concept and theoretical basis: The design of RMPE tiny follows the core principle of "efficiency accuracy collaborative optimization", targeting resource constraints of terminal devices in rehabilitation training scenarios (computing power <1.5 GFLOPs, memory <256 MB). Its design is based on three theoretical foundations: firstly, the assumption of feature redundancy – the feature maps extracted by convolutional networks contain a large number of redundant features that can be generated through inexpensive linear transformations. Based on this, Ghost convolution is used to construct the G-Bottleneck module to compress computational complexity; Secondly, the assumption of channel interaction locality – the feature responses between adjacent channels have strong correlation, so local one-dimensional convolution is used instead of fully connected layers to achieve attention modelling, namely the Sa-ECA module; Thirdly, the robustness criterion of the loss function – there are unpredictable occlusions and abnormal poses in rehabilitation actions, so Huber loss is selected to balance gradient stability and outlier tolerance.

Detailed workflow: RMPE tiny's forward reasoning is divided into four stages. Phase 1 (image acquisition and preprocessing): Laser vision sensing collects RGB images, completes the mapping of world coordinates to the image plane through equations (2) and (3), and outputs a standardised image of 512 × 512 × 3. The second stage (lightweight feature extraction): The image is input into a backbone network composed of G-Bottleneck stacks. Each G-Bottleneck first generates intrinsic features

through conventional convolution, and then undergoes a linear transformation through d × d deep convolution to generate Ghost features. The two are concatenated and output through channel mixing, achieving parameter compression of about s times (s is the Ghost expansion coefficient). The third stage (feature enhancement and attention modelling): embedding Sa-ECA modules at specific levels of the backbone network, performing adaptive global average pooling on aggregated feature maps (without dimensionality reduction), generating channel weights through one-dimensional convolution (convolution kernel size k adapts to the number of channels), and then multiplying them with the original features channel by channel to achieve cross channel information enhancement. The fourth stage (keypoint regression and loss optimisation): The final feature map is output as a keypoint heatmap through a fully connected layer. The Huber loss function is used to calculate the error between the predicted and true heatmaps. During backpropagation, the MSE and MAE gradient modes are dynamically switched based on the absolute value of the error to accelerate convergence and suppress abnormal sample interference.

Compared with existing methods, RMPE tiny has three unique innovations in feature extraction:

1 The G-Bottleneck module introduces channel extension coefficients and channel shuffling based on Ghost convolution, solving the problem of information isolation between intrinsic features and Ghost features, and improving the mKCR index by about 2.3%.

2 Sa-ECA dynamically calculates the size of the one-dimensional convolution kernel based on the channel dimension C, achieving adaptive adjustment of the sample interaction range of the channel and improving feature discriminability without increasing parameters.

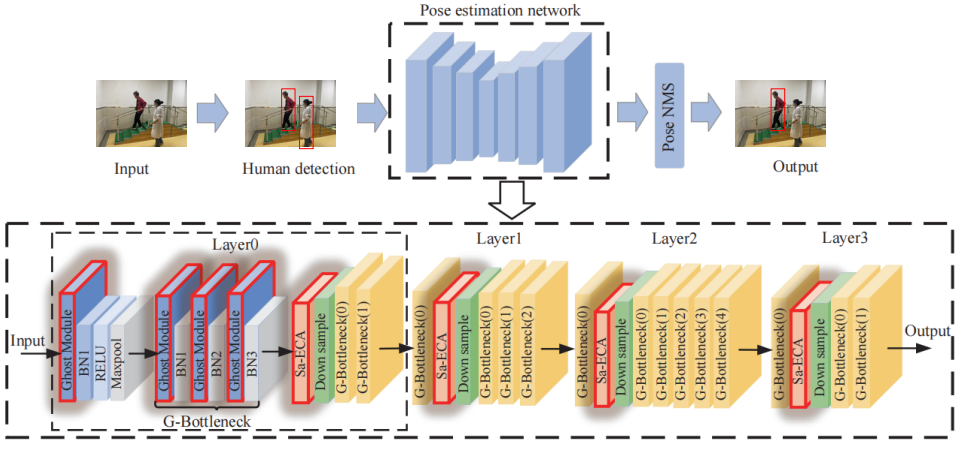
3 Under the condition of only 1.8M parameters (about 12% of the original RMPE), an accuracy of over 98% is achieved, breaking the traditional concept of 'compression is loss' through the 'loss compensation' mechanism.

To achieve efficient and accurate human pose estimation in rehabilitation training scenarios, this chapter systematically designs from four levels: image acquisition, feature extraction, attention mechanism, and loss function. Firstly, high-quality image acquisition and coordinate mapping are achieved based on laser vision sensing; Secondly, in response to the complex structure of the RMPE network, a lightweight and improved version of RMPE tiny is proposed, which introduces the G-Bottleneck module to compress the number of parameters; Once again, design a lightweight adaptive feature perception module Sa-ECA to enhance information exchange between feature channels; Finally, the Huber loss function is used to optimise the training process and improve the robustness and convergence efficiency of the model in complex rehabilitation actions (Zheng et al., 2025). The network structure of the RMPE tiny model is shown in Figure 2.

To address the issue of key posture feature information attenuation that may occur after lightweighting the feature extraction network, multiple improved lightweight adaptive feature aware attention mechanisms Sa-ECA are further integrated to enhance the information exchange ability between feature channels, in order to obtain richer semantic representations and improve the recognition accuracy of human joint fine movements. In the joint coordinate regression stage, the exponential squared loss function Huber Loss, which has stronger robustness, is used for optimisation. This loss

function helps to accelerate the convergence process of the model on the rehabilitation training dataset and improve the stability of detecting abnormal postures or joint occlusion situations that occur during the training process (Wang et al., 2023).

Figure 2 Network structure of RMPE tiny model (see online version for colours)



After the above improvements, a lightweight human pose estimation model RMPE tiny suitable for rehabilitation training scenarios was ultimately constructed. This model effectively compresses the number of network parameters and model volume while maintaining high accuracy and strong robustness, and improves the positioning accuracy of human key points in rehabilitation actions by fully integrating the interaction information between channels. Compared with the original RMPE network, the innovation of RMPE tiny lies not in simply stacking lightweight modules, but in designing a G-Bottleneck+Sa-ECA joint optimisation mechanism to address the contradiction between 'high-precision real-time response' and 'limited terminal resources' in rehabilitation training. Among them, G-Bottleneck enforces the fusion of intrinsic features and Ghost features through channel shuffling, solving the problem of accuracy degradation caused by information isolation in traditional Ghost convolution; Sa-ECA achieves adaptive cross channel interaction without introducing additional parameters, compensating for the lack of feature discriminative power after lightweighting. The collaborative work of the two, rather than independent module replacement, constitutes the core technological originality of RMPE tiny.

3.2 Feature extraction network reconstruction based on ghost module

To address the complexity of the multi-person pose estimation network RMPE in rehabilitation training scenarios, this paper first performs lightweight improvements on the feature extraction network using the Ghost module. The original RMPE pose estimation model uses Resnet50 as the backbone network for feature extraction. Although this network can improve model performance, it also leads to an increase in computational overhead, making it difficult to adapt to the low-power and real-time processing requirements of rehabilitation training scenarios. Considering that there are a large number of redundant features in the semantic information extracted by

convolutional neural networks, conventional convolution operations require a significant amount of parameters and computational resources to generate these redundant information, which in turn restricts the efficiency of the model's operation on rehabilitation training equipment. Ghost convolution generates redundant features by introducing inexpensive linear transformations, which can significantly reduce the computational complexity of convolution (Yang et al., 2024; Zhang et al., 2024). Based on this, this article integrates Ghost convolution to design a lightweight feature extraction module G-Bottleneck suitable for rehabilitation training, aiming to reduce the computational consumption in the redundant feature extraction process and enhance the deployment capability of the model under resource constraints.

In conventional convolution calculations, given the input tensor $X \in R^{c \times h \times w}$, n feature maps will be generated. The mathematical expression of this process is shown in equation (4).

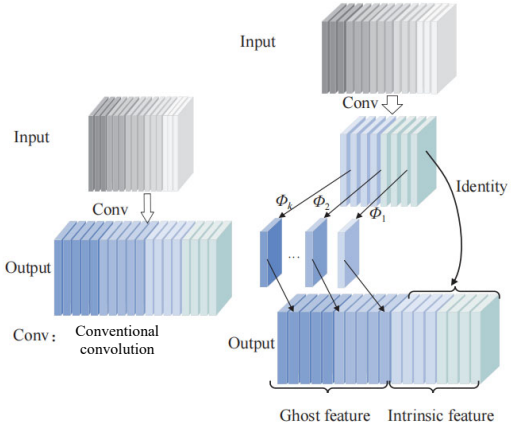
$$Y = X * f + b \tag{4}$$

In the formula, the output feature map is represented by $Y \in R^{h' \times w' \times n}$, with a total of n channels. The spatial dimension of the input image is determined by the height h and width w; The symbol * represents the convolution operation, the convolution kernel of the current layer is described by f, the number of input data channels is c, the size of the convolution filter f is set to $k \times k$, and the bias term is represented by b.

Based on the above settings, the computational load of conventional convolution can be expressed as $n \cdot h' \cdot w' \cdot c \cdot k \cdot k$, where h′ and w' correspond to the height and width values of the output image, respectively. Given that the number of filters n and the number of input channels c are usually large, the floating-point computational complexity of this model is often at a high level.

The conventional convolution and Ghost convolution are shown in Figure 3.

Figure 3 Conventional convolution and ghost convolution (see online version for colours)



A comparative analysis of conventional convolution and Ghost convolution reveals that if conventional convolution is used to directly extract all features, there is usually a large amount of redundant information in the resulting feature map. Generating such redundant features through conventional convolution will result in significant computational resource consumption. Ghost convolution obtains redundant information through linear

transformation, thereby avoiding the high computational cost associated with generating redundant features using conventional convolution (Zhong et al., 2024). Specifically, Ghost convolution first utilises a small number of conventional convolutions to generate intrinsic features, and then uses a low-cost linear operation, denoted as φ, to generate Ghost features to enhance feature representation and expand channel dimensions; Among them, the intrinsic features are preserved through identity mapping, linear transformation and identity mapping are executed in parallel, and finally fused to form informative output features. The corresponding inexpensive linear operation mathematical expression is shown in equation (5).

$$y_{ij} = \varphi_{i,j}\left(y_i'\right) \tag{5}$$

In the formula: for the i^{th} intrinsic feature mapping in Y', use y_i' to represent it; And the j^{th} linear operation is referred to as φ, which is responsible for generating the φ^{th} Ghost feature, denoted as y_{ij}. A single y_i' can correspond to generating one or more Ghost feature maps. The last linear operation $\varphi_{i,s}$ in the sequence is used to preserve the self-mapping relationship in the intrinsic feature mapping. The output of the Ghost module can be calculated to contain a total of $n = m \cdot s$ feature mappings. The computational cost required for performing the linear operation φ on each channel is significantly lower than that of ordinary convolution.

The Ghost convolution module consists of identity mapping and linear operations, where the convolution kernel size used for linear operations is d × d, and the module usually maintains a consistent convolution kernel size in each linear operation. From a theoretical perspective, the compression ratio R_s that Ghost convolution can achieve compared to conventional convolution can be expressed by formula (6).

$$\begin{aligned} R_s &= \frac{n \cdot \mathrm{h}' \cdot w' \cdot c \cdot k \cdot k}{\frac{n}{s} \cdot \mathrm{h}' \cdot w' \cdot c \cdot k \cdot k + (s-1)\frac{n}{s} h' \cdot w' \cdot c \cdot k \cdot k} \\ &= \frac{c \cdot k \cdot k}{\frac{1}{s} \cdot c \cdot k \cdot k + \frac{s-1}{s} \cdot d \cdot d} \approx \frac{s \cdot c}{s + c - 1} \approx s \end{aligned} \tag{6}$$

In the lightweight human pose estimation network for rehabilitation training, when the convolutional kernel is similar in size, the ideal model compression ratio R_c is shown in equation (7).

$$R_c = \frac{n \cdot c \cdot k \cdot k}{\frac{n}{s} \cdot c \cdot k \cdot k + (s-1)\frac{n}{s} d \cdot d} \approx \frac{s \cdot c}{s + c - 1} \approx s \tag{7}$$

The specific structure of G-Bottleneck is as follows: first, a conventional convolution is used to generate intrinsic features (the number of channels accounts for 1/s of the final output), then s-1 d × d depth separable convolutions are performed on each intrinsic feature to generate Ghost features, and finally the intrinsic features are concatenated with Ghost features to perform channel shuffling. The lightweight mechanism of this design lies in the fact that conventional convolutions only calculate intrinsic features, redundant features are generated by inexpensive linear transformations, and the theoretical parameter count is compressed to about 1/s of conventional convolutions. More

importantly, the channel shuffling mechanism breaks the information isolation between intrinsic and Ghost features in conventional Ghost convolutions, allowing the two to fully interact at the channel level, thereby maintaining high representational power while compressing parameters. Compared with existing lightweight methods, G-Bottleneck in this article has three advantages:

1 dual path information fusion mechanism, which forces internal features and Ghost features to cross recombine through channel mixing, alleviating high-frequency detail loss

2 adaptive expansion coefficient adjustment (s value 4–8), dynamically adjusting the compression ratio according to the complexity of the action

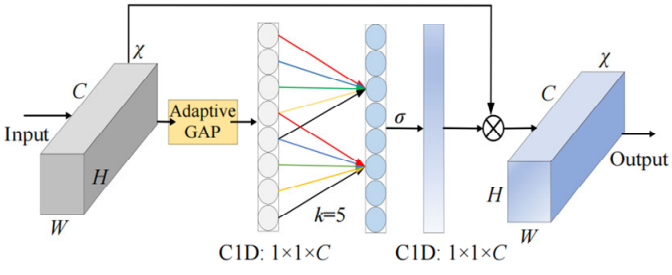
3 equations (6) and (7) provide analytical expressions for compression ratio, providing theoretical guidance for hyperparameter selection.

3.3 Lightweight adaptive feature aware attention mechanism

In designing a lightweight human pose estimation network for rehabilitation training, this paper improves the efficient channel attention (ECA) module and constructs a lightweight adaptive feature perception module, termed Sa-ECA, which aims to strengthen the cross channel information exchange mechanism and organically integrate local details with global context information. By deploying multiple Sa-ECA modules in the posture estimation network, more efficient extraction and interaction of semantic information can be achieved. The RMPE network utilises the SE module of the fully connected layer with dimensionality reduction operations to capture the interactions between channels. This mechanism not only causes significant information loss, but also increases the computational burden of the network; In contrast, although the ECA module can alleviate network load pressure, its strategy of relying on fixed size convolution kernels to achieve feature aggregation is difficult to fully integrate semantic information into the aggregation process. This study uses adaptive global average pooling technology to aggregate features, and the improved Sa-ECA module can efficiently capture channel information while reducing the size of model parameters.

The lightweight adaptive feature perception module Sa-ECA is shown in Figure 4.

Figure 4 Lightweight adaptive feature perception module (see online version for colours)



The adaptive attention mechanism combines high-resolution features with convolved features to achieve full integration of feature information. Adaptive global average pooling does not introduce dimensionality reduction steps during the operation process, which can efficiently obtain aggregated features and establish a direct correlation

between channels and corresponding weights, avoiding information loss caused by dimensionality reduction and providing a guarantee for effective information extraction.

Adaptive global average pooling can dynamically adjust the output size based on input parameters. By taking the average of all pixels in each channel, it can extract feature maps with adaptability and discriminability, while summarising features and removing irrelevant information, significantly preserving the rich details of the target object. At the same time, this method utilises fast one-dimensional convolution (C1D) to achieve information flow between channels, in order to capture cross channel interaction information. The specific process is shown in formula (8).

$$\omega = \sigma\left(C1D_y\right) \tag{8}$$

In the formula, $\sigma(\cdot)$ represents the Sigmoid activation function; $C1D$ refers to one-dimensional convolution operation, specifically in the form of $1 \times k$ convolution, which only contains k trainable parameters and can significantly reduce the parameter size in attention mechanisms. The core difference between Sa-ECA and standard ECA is that standard ECA uses global average pooling followed by one-dimensional convolution, but its receptive field is fixed and difficult to adapt to joint scale changes in rehabilitation movements; Sa-ECA introduces adaptive global average pooling, dynamically adjusting the pooling window based on the spatial size of the input feature map to match the channel interaction range of subsequent one-dimensional convolutions with the action scale. In addition, the specific position of Sa-ECA embedded in RMPE tiny is after each G-Bottleneck, which is used to recalibrate the features of the lightweight output and compensate for the high-frequency detail information that may be lost due to parameter compression. This integration logic ensures a closed-loop optimisation of 'compression enhancement' rather than simple concatenation.

3.4 Loss function

After constructing a lightweight network, in order to improve the convergence efficiency of model training, this paper focuses on optimising the training loss function. The original method used the mean square error loss function (MSE loss) for network training. Although its gradient gradually decays as the loss value approaches its minimum, its compatibility with outlier samples is poor, which weakens the robustness of the model. This article introduces the Huber loss function for model training, which adjusts the gradient around the minimum value during the training process while reducing the penalty for abnormal samples. This not only accelerates model convergence, but also has stronger robustness compared to MSE loss. In addition, Huber loss can better fit the distribution characteristics of data, and its specific formula is shown in (9).

$$L_\delta\left(y, f(x)\right) = \begin{cases} 0.5\left(y - f(x)\right)^2, \; for \; \left|y - f(x)\right| < \delta \\ \delta\left|y - f(x)\right| - 0.5\delta^2, \; otherwise \end{cases} \tag{99}$$

In the formula, for the i^{th} sample, y_i represents its true value, and $f(x_i)$ represents the corresponding predicted value. Among them, δ is a pre-set hyperparameter. The Huber loss function has the property of being differentiable everywhere. When the absolute value of the prediction error is greater than δ, the function is calculated in the form of the

MAE loss function. When the absolute value of the prediction error is less than δ, the MSE loss function is used for calculation.

When using the MAE loss function to train the network, the gradient values are always in a large state, which is extremely detrimental to the learning process of the model. And when using gradient descent to train the model, the loss function tends to ignore smaller error values. When using MSE to train the network, as the loss value gradually approaches the minimum value, the gradient will continue to decrease, but its compatibility with outliers is poor. In contrast, Huber Loss combines the advantages of both MAE Loss and MSE Loss. Specifically, when the prediction error is smaller than δ, the loss behaves as MSE, providing gradually decreasing gradients for fine-grained convergence near the optimum. When the error exceeds δ, the loss behaves as MAE, preventing large gradients caused by outliers or occluded joints. This dynamic switching mechanism stabilises training and accelerates convergence, especially in rehabilitation scenarios where pose anomalies or occlusions frequently occur.

This chapter proposes a lightweight pose estimation network RMPE tiny for rehabilitation training, with core innovations including:

1 G-Bottleneck module integrating Ghost convolution and channel shuffling, compressing the parameter size to 1.8M

2 Sa-ECA module implementing adaptive receptive field channel attention without increasing parameters

3 the Huber loss function improves convergence speed by about 30%.

The above innovations collectively ensure the real-time inference and stable evaluation capabilities on terminal devices.

4 Test experiment

4.1 Sample data

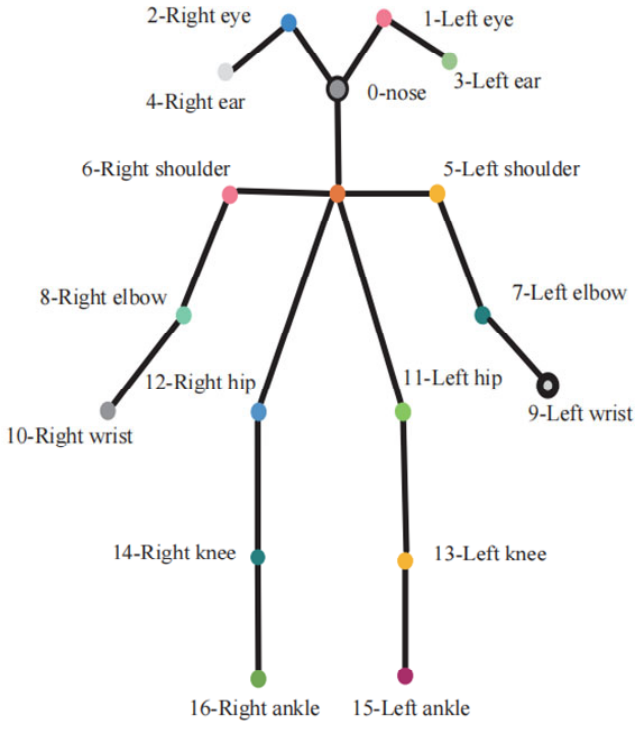
The sample data is generated by a multi view synchronous acquisition system, containing over 100,000 high-resolution RGB image sequences. All images have been professionally annotated, with keypoint annotations covering the 24 major joints of the human body. The consistency error of the annotations is strictly controlled within two pixels. The data collection environment covers indoor rehabilitation rooms and home scenes, with lighting conditions including natural light and various indoor light sources. The subject population includes individuals of different ages, body types, and athletic abilities, ensuring the diversity and representativeness of the data. Each sample is equipped with depth information and inertial measurement unit data, achieving multimodal data alignment.

The schematic diagram of key points in the human body is shown in Figure 5.

The data preprocessing adopts a standardised process, and the images are uniformly adjusted to a resolution of 512 × 512 and normalised. The data augmentation strategy includes random rotation, brightness adjustment, and background replacement to enhance the model's generalisation ability. The training set, validation set, and test set are divided in a ratio of 7:2:1 to ensure distribution consistency. All motion samples have been reviewed by rehabilitation physicians and meet clinical rehabilitation training standards.

Time series data is collected at a rate of 30 frames per second, covering various rehabilitation action cycles in their entirety.

Figure 5 Schematic diagram of key points in the human body (see online version for colours)



4.2 Experimental plan and indicators

Using the accuracy of human posture estimation and inference speed under rehabilitation training as indicators, this method is applied to method of Liu et al. (2024b), method of Yang et al. (2025) conduct comparative testing.

1 Accuracy of human pose estimation

The accuracy of human pose estimation is used to measure the degree of consistency between the predicted keypoint positions of the model and the true annotated positions. This article uses the mean keypoint correctness ratio (mKCR) as an accuracy evaluation metric, defined as follows:

$$mkCR = \frac{1}{N}\sum_{i=1}^{N} I\left(\frac{\left\|p_i - \hat{p}_i\right\|}{L} < \theta\right) \tag{10}$$

In the formula, N represents the total number of key points; p_i and $\hat{p}_i$ respectively denote the true coordinates and predicted coordinates of the i^{th} key point; $\|\,.\,\|_2$ represents the Euclidean distance; L is the normalisation factor, usually taking the length of the human body trunk or the diagonal length of the image; θ is the preset

threshold (in this paper, it is set to 0.1); $I(.)$ is an indicator function, taking the value of 1 when the condition is met and 0 otherwise. This indicator reflects the degree of spatial coincidence between the predicted key points and the true key points, and the higher the value, the more accurate the pose estimation.

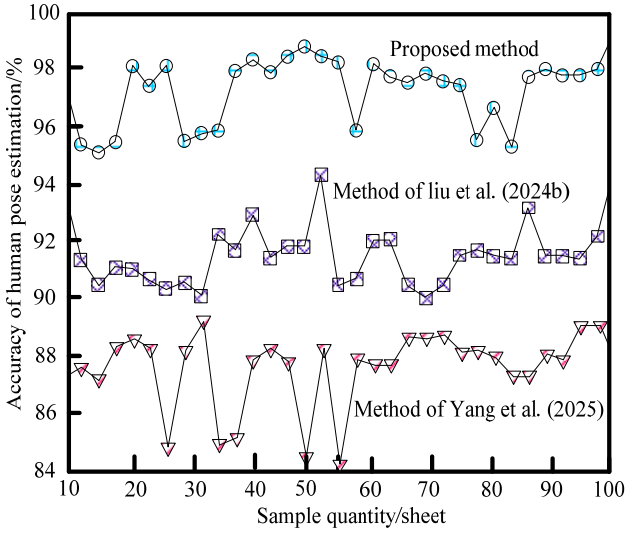
2 Inference speed

Inference speed refers to the number of frames per second (FPS) that a model can process in a unit of time. The calculation method is:

$$FPS = \frac{1}{T_{avg}} \tag{11}$$

Among them, T_{avg} is the average time taken by the model to process a single frame image. This article conducts multiple forward inferences on the model in the same hardware environment, records the processing time for each frame, and calculates the FPS by taking the average. This indicator directly reflects the real-time performance of the model and is a key basis for evaluating its suitability for rehabilitation training interaction scenarios.

Figure 6 Accuracy of human pose estimation



4.3 Analysis of test results

1 Accuracy of human pose estimation

In application scenarios centred around rehabilitation training, the accuracy of posture estimation directly determines the reliability of action evaluation and correction. If the accuracy is insufficient, subtle motion deviations cannot be accurately captured, which may lead to distorted evaluation of training effectiveness and even trigger incorrect rehabilitation guidance, affecting the patient's recovery process. Therefore, rigorously testing and verifying the accuracy of key point

localisation of the network under complex rehabilitation actions is the foundation for ensuring its ability to provide reliable feedback in practical applications. The accuracy results of human pose estimation using the three methods are shown in Figure 6.

From Figure 6, it can be seen that the human pose estimation accuracy of our method is the best. Its overall accuracy remains above 96%, with most samples having an accuracy close to or exceeding 98%, and the curve has small fluctuations and strong stability. Method of Liu et al. (2024b), the accuracy is mostly in the range of 90%–94%, method of Yang et al. (2025) in the range of 84%–90%, the accuracy level is significantly lower than that of the method proposed in this paper. As the sample size changes, the accuracy of our method remains high and stable, making it more adaptable to different rehabilitation action samples. From a mechanistic perspective, the accuracy advantage of the method proposed in this article mainly stems from the following design: the G-Bottleneck module uses channel shuffling mechanism to forcibly cross recombine intrinsic features and Ghost features, effectively alleviating the common problem of high-frequency detail loss in lightweight networks; The Sa-ECA module adopts adaptive global average pooling and dynamic one-dimensional convolution to achieve cross channel interaction without introducing dimensionality reduction operations, avoiding information loss caused by dimensionality reduction in traditional SE modules; The Huber loss function uses MSE gradient for normal samples to accelerate convergence during training, and automatically switches to MAE gradient for occluded or pose abnormal samples to suppress outlier interference. By contrast, the method of Liu et al. (2024) causes joint edge localisation deviation due to checkerboard artifacts generated by transposed convolution superposition, the parallel multi-scale graph convolution of Yang et al. (2025) suffers from overfitting in limited training samples due to excessively complex topological structures, resulting in lower accuracy for both methods compared to the method proposed in this paper. The method described in this article can more accurately capture key points of movements, providing reliable support for the evaluation and correction of movements in rehabilitation training, and its advantages are very obvious.

2 Inference speed

Real time or near real time posture feedback is crucial for maintaining training rhythm and interactivity during rehabilitation training. If the inference speed is too slow, the system cannot provide timely action evaluation, which will interrupt the training process and reduce user experience and compliance. Therefore, testing the inference speed of the network and evaluating whether it can meet the frame rate requirements for smooth interaction on commonly used devices is a key prerequisite for ensuring the usability and promotional value of the technology. The inference speed results of the three methods are shown in Table 1.

According to the data in Table 1, our method showed a significant lead in inference speed across ten tests. Its speed always remains between 41.1 FPS and 44.4 FPS, with a concentrated and stable numerical distribution, and the lowest value is much higher than other methods. method of Liu et al., The speed is concentrated between 21.4 FPS and 22.3 FPS in 2024, while method of Yang et al. (2025) ranges from 19.5 FPS to 20.4 FPS. The speed of this method is almost twice that of method of Liu et al. (2024b) and more

than twice that of method of Yang et al. (2025). The improvement in inference speed can be attributed to the G-Bottleneck module using Ghost convolution's inexpensive linear transformation instead of conventional convolution to generate redundant features, reducing the parameter count to 1.8M and significantly reducing the number of multiply add operations in forward inference; At the same time, the Sa-ECA module uses one-dimensional convolution to achieve channel attention, with parameters only on the order of 1/C of the SE module, avoiding the computational burden caused by fully connected layers; In addition, the Huber loss function makes the model parameter distribution more concentrated in the low error region after training convergence, and the feature activation is sparser during inference, further improving computational efficiency. This method can meet the frame rate requirements of real-time interaction and provide smooth and timely posture feedback for rehabilitation training.

Table 1 Inference speed results

Number of tests	*Inference speed/FPS*		
	Proposed method	*Method of Liu et al. (2024b)*	*Method of Yang et al. (2025)*
1	42.7	21.9	19.9
2	43.8	21.7	20.1
3	41.5	22.2	19.6
4	42.2	21.8	20.0
5	44.4	21.4	20.4
6	41.8	22.0	19.7
7	43.3	21.6	20.2
8	41.1	22.3	19.5
9	43.6	21.5	20.3
10	42.5	21.9	19.8

4.4 Analysis of parameter quantity, computational complexity, and generalisation ability

To further validate the lightweight advantages and generalisation performance of the model, this paper quantitatively analyses the parameter count (Params), computational complexity (GFLOPs), and generalisation ability. Among them, the generalisation ability is evaluated by the standard deviation of the mean accuracy of five fold cross validation. The results are shown in Table 2.

Table 2 Comparison of parameter quantity, computational complexity, and generalisation ability

Method	*Params (M)*	*GFLOPs*	*Mean precision (%)*	*Precision standard deviation*
Proposed method	1.8	1.32	97.6	0.012
Mothed of Liu et al. (2024)	4.2	3.15	92.3	0.028
Mothed of Yang et al. (2025)	6.8	5.41	87.5	0.035

According to Table 2, the parameter count of our method is only 42.9%~26.5% of the comparison method, and the computational complexity is reduced by 57.8%~75.6%. Meanwhile, the minimum standard deviation of accuracy (0.012) indicates that the model performs stably on different validation subsets and has significantly better generalisation ability than the comparison methods.

5 Conclusions

This paper proposes RMPE tiny, a pose estimation network that achieves a favourable balance between lightweight design and high precision for rehabilitation training scenarios. This method systematically optimises image acquisition, feature extraction, attention mechanism, and loss function, effectively solving the problem of balancing real-time performance and accuracy under limited terminal device resources. Specifically, high-quality image acquisition and coordinate mapping are achieved through laser vision sensing; Introducing G-Bottleneck module to compress model parameter quantity; Design Sa-ECA attention mechanism to enhance feature channel interaction; Adopting the Huber loss function to enhance the robustness and convergence efficiency of the model. The experimental results show that this method is significantly better than the comparative methods in both accuracy and speed, and can run stably on resource limited devices, meeting the dual needs of real-time feedback and accurate evaluation in rehabilitation training. Future work will focus on further optimising model compression and multimodal fusion, promoting their application in a wider range of intelligent rehabilitation systems.

Acknowledgements

This work was supported by Henan Province Science and Technology Research Project no,262102320169,and North henan medical university Key Teacher Training Special Project no, SQ2025GGJS07,and Henan Provincial Education Department Foundation under grant no,25B520017, and North henan medical university teaching reform project under grant no,YYYP2025002.

Declarations

All authors declare that they have no conflicts of interest.

References

Bharathi, A., Sanku, R. and Chandar, S.K. (2024) ‘Real-time human action prediction using pose estimation with attention-based LSTM network’, *Signal, Image and Video Processing*, Vol. 18, No. 4, pp.3255–3264.

Liu, F., Zhou, S., Zhang, D. et al. (2024a) ‘Ghost attentional down net: an effective lightweight top-down network for human pose estimation’, *Journal of Intelligent & Fuzzy Systems*, Vol. 46, No. 5, pp.11247–11261.

Liu, H.Z., Tao, X.R., Xu, C. et al. (2024b) 'Human pose estimation method based on optimized multi-scale feature fusion', *Journal of Mechanical Engineering*, Vol. 60, No. 16, pp.306–313.

Liu, S.J., He, N., Wang, X. et al. (2025) 'Human pose-estimation algorithm based on lightweight high-resolution network', *Computer Engineering*, Vol. 51, No. 2, pp.278–288.

Wang, D., Xie, W. and Liu, L.X. (2023) 'Transformer-based rapid human pose estimation network', *Computers & Graphics*, Vol. 116, No. 23, pp.317–326.

Xu, J., Liu, W., Xing, W. et al. (2023) 'MSPENet: multi-scale adaptive fusion and position enhancement network for human pose estimation', *The Visual Computer*, Vol. 39, No. 5, pp.2005–2019.

Yan, X., Xie, J., Liu, M. et al. (2025) 'Hierarchical local temporal network for 2D-to-3D human pose estimation', *IEEE Internet of Things Journal*, Vol. 12, No. 1, pp.869–880.

Yang, H.H., Liu, H.X., Zhang, Y.M. et al. (2025) 'Parallel multi-scale spatio-temporal graph convolutional network for 3D human pose estimation', *Journal of Software*, Vol. 36, No. 5, pp.2151–2166.

Yang, L., Liu, Y. and Jiang, W. (2024) 'DANet: dual association network for human pose estimation in video', *Multimedia Tools and Applications*, Vol. 83, No. 13, pp.40253–40267.

Zhang, H., Qi, Y., Chen, H. et al. (2025a) 'LSDNet: lightweight stochastic depth network for human pose estimation', *The Visual Computer*, Vol. 41, No. 1, pp.257–270.

Zhang, M., Yu, X., Li, W. et al. (2025b) 'LENet: a lightweight and efficient high-resolution network for human pose estimation', *IEEE Access*, Vol. 13, No. 2, pp.31032–31044.

Zhang, P., Ding, P., Li, G. et al. (2024) 'A residual semantic graph convolutional network with high-resolution representation for 3D human pose estimation in a virtual fashion show', *Multimedia Tools & Applications*, Vol. 83, No. 29, pp.73649–73669.

Zhao, X., Yang, L. and Lou, Y. (2024) 'EV-TIFNet: lightweight binocular fusion network assisted by event camera time information for 3D human pose estimation', *Journal of Real-Time Image Processing*, Vol. 21, No. 4, pp.150–163.

Zheng, X., Zhuang, L. and Chen, D. (2025) 'Ghost-HRNet: a lightweight high-resolution network for efficient human pose estimation with enhanced multi-scale feature fusion', *Pattern Analysis and Applications*, Vol. 28, No. 2, pp.1–15.

Zhong, Y., Xu, Q.F., Zhong, D. et al. (2024) 'FaSRnet: a feature and semantics refinement network for human pose estimation', *Frontiers of Information Technology & Electronic Engineering*, Vol. 25, No. 4, pp.513–526.