Journal of Network Intelligence ©2025 ISSN 2414-8105
Taiwan Ubiquitous Information Volume **10**, Number **1**, February **2025**

Dance Movement Pose Estimation in Complex Scenes Based on Improved High-Resolution Networks

Jia Sun*

School of Dance
Guangxi University of Arts, Nanning 530000, P. R. China
sj14472024@163.com

Lewis Song

Faculty of Engineering
University of Perugia, 06123 Perugia PG, Italy
jy4290@163.com

*Corresponding author: Jia Sun

Received March 15, 2024, revised May 31, 2024, accepted August 1, 2024.

Abstract. *Using keypoint detection techniques, such as human posture estimation, to extract keypoint information, including joint positions and postures, from dance videos can help to understand the details of dancers' movements. However, the method of determining posture by observing body movements in a manual way is time-consuming and labour-intensive, and cannot be applied to dance video sequences of complex scenes. Therefore, a pose estimation method based on an improved High Resolution Network is proposed. Firstly, Convolutional Block Attention Module (CBAM) is introduced in High Resolution Network (HRNet) for the optimisation of the pose estimation network. CBAM contains 2 independent sub-modules. CBAM expands the receptive field, thus increasing the extraction rate of human feature information from both the channel and spatial aspects of the feature network. Then, in order to solve the problem of scale sensitivity of bottom-up pose estimation network, a background information feature extraction method is proposed. The approach efficiently extracts features at many scales and combines them via attention feature fusion. This results in fused features that incorporate information from different scales, significantly improving the network's ability to handle scale variations. Through experimental validation, it can be seen that the accuracy of the pose estimation method based on background information features and CBAM on the MPII dataset and the self-constructed dance dataset is 73.5% and 79.5%, respectively, which exceeds most of the bottom-up multi-person pose estimation methods.*

Keywords: Dance movement; Pose estimation; Deep learning; Attention module; Multi-scale features

1. **Introduction.** Dance, as a form full of art and expression, has always carried the essence of human culture and the transmission of emotions [1]. However, it is not easy to accurately understand and assess the significance of dance movements as they are rich in cultural connotations, emotional expressions and technical details [2, 3]. Traditional approaches often rely on the subjective judgement of professional dance adjudicators, which is influenced by human factors and fails to provide a systematic and objective means of analysis [4]. With the rapid development of artificial intelligence technology, especially the breakthroughs in computer vision, deep learning and natural language processing, we now have more possibilities to explore the meaning of dance movements.

Through the application of computer vision techniques, we can extract rich movement information from dance videos, including movement types, key point locations, and emotional expressions [5, 6], which provides us with more dimensional data to analyse and evaluate dance works. In this context, this work will focus on the application of AI techniques to estimate dance movement gestures in complex scenes. In today's digital era, the performance and evaluation of dance art is undergoing unprecedented changes. With the rapid development of computer vision and machine learning techniques, how to accurately estimate dancers' movement gestures from dance videos in complex scenes [7, 8 ,9] has become a hot issue in the field of scientific research and technology development. In particular, High-Resolution Networks (HRNet) have shown great potential in human movement pose estimation, offering new possibilities to improve the accuracy of pose estimation by virtue of their sensitivity to details and ability to maintain high-resolution features. However, traditional HRNet still faces several challenges in dealing with dance movements in complex scenes, including background interference, lighting variations, and occlusion problems [10 ,11], all of which may affect the accuracy and reliability of pose estimation.

The research objective of this paper aims to enhance the ability of high-resolution networks to estimate dance movement poses in complex scenes by improving them. The relevant principles of human gesture estimation are first analysed, including the principles of deep learning, the mainstream methods of human gesture estimation and the principles of high-resolution networks. Then 2 targeted improvement measures are proposed, including the introduction of the attention mechanism, the extraction and fusion of background information features, etc., in order to improve the network's recognition accuracy and robustness for dance movements in complex scenes [12]. The research in this work is not only of great significance to the research and teaching of dance art, which can further promote the dissemination and development of dance art by capturing and understanding dance movements more accurately; at the same time, this research provides a new solution to the problem of gesture estimation in complex scenes at the technical level, and the results are expected to be widely used in a variety of fields, such as human-computer interaction, sports analysis, virtual reality, and so on. By deeply exploring the gesture estimation technique for dance movements based on improved high-resolution networks, this study is dedicated to promoting the application of AI technology in the intersection of art and science, and providing new perspectives and tools for researchers in related fields.

1.1. **Related Work.** Human pose estimation is one of the challenging research areas in computer vision, and with the help of deep learning and publicly available datasets, significant progress has been made in solving problems related to human pose estimation, which aims at determining the location or spatial position of key points of the human body from a given image or video, and connecting the key points in a given order to connecting them to finally obtain the human skeleton. There are two main types of human pose estimate that are based on dimensions: 2D and 3D. It is possible to further classify 2D human posture estimate as either single-person or multiple-person [13]. 2D posture estimation enhances the temporal information-based prediction of crucial points in a video system when the sample set is a video sequence. A more precise spatial representation is achieved via 3D pose estimation, which differs from just predicting the 2D locations of the body joints [14]. This method also predicts the depth information. It not only predicts the 3D position of the body, but also the detailed 3D shape of the body and body texture. Due to the complexity and high cost of 3D pose estimation, the current mainstream is still 2D pose estimation, which is the research direction of this work.

In their study, Liu et al. [15] put forward a method that combines pose regression. Given the stability and learnability of bones compared to joints, the proposed method incorporates a novel reparameterised pose. Thus, the network utilizes a collaborative linkage framework to establish the component loss functions that capture the distant interactions among bones. Luvizon et al. [16] proposed a novel regression method integrating the soft-argmax operation, which directly converts the extracted feature maps into body joint coordinates. In this architecture, it is possible to directly access the background information and aggregate it with the final prediction. Most notably, the method does not generate heat maps during the training phase. Human posture estimation can also be operated with useful information provided by manual parsing. Therefore, Zhang et al. [17] introduced an innovative network structure that utilizes human parsing knowledge for the purpose of predicting joint coordinates. The network, named "parse-induced learner", is used to subdivide substandard localisation and correct erroneous human joint classification, which can be used to solve the occlusion problem of human pose estimation. For both one- and multiple-person posture estimation, the encoder takes the input picture and converts it into a high-level, sparse representation; the adapter then learns how to tweak the position model.

Deep network-based is another architecture for human posture estimation. Petrov et al. [18] proposed to solve the human posture estimation problem by using deep neural network to capture the coordinates of key points of the whole body by direct regression through deep learning. Compared with most regression-based human pose estimation methods, deep learning is more effective. Guo et al. [19] suggested a method for successively aggregating and summing features, where ResNet is used instead of the hourglass module in the deep learning network to combine the features raised at each stage to obtain both local and global background knowledge, which is effective in solving the problems of lighting and occlusion. Yang et al. [20] made modifications to residual connections and included gated skip connections with adjustable settings for each channel. This was done to manage the data flow of each channel inside a module within a macro module. In order to reduce the post-processing of localised human joints and human enclosing frames, Li et al. [21] employed Convolutional Block Attention Module (CBAM) to increase the receptive field of the network and enhance the precision of the estimate of human poses.

1.2. **Motivation and contribution.** There are two problems with HRNet in human pose estimation tasks. Firstly, HRNet needs to detect all the key points of the human body in the input image at the same time, so the multiple scales of the human body contained in the image is one of the main reasons limiting HRNet [22]. Then, the upsampling of HRNet loses part of the feature information, and without learning parameters, the lost feature information cannot be compensated by learning [23]. Therefore, in order to solve the above problems, this paper proposes a pose estimation method based on improved high-resolution network. The main innovations and contributions of this work include:

(1) In order to solve the problem of partial feature information loss, CBAM (Convolutional Block Attention Module) is introduced in this paper.The CBAM attention mechanism can improve the model's representation ability in the channel, and at the same time emphasise the features of spatial location in the spatial dimension, so as to improve the network's representation ability in two ways.CBAM is introduced into the parallel subnets of HRNet networks with different resolutions, thus improving the extraction rate of channel and spatial feature information in the network.

(2)A feature extraction method for background information that can extract both global and local features is proposed. A combination of deep and shallow network features will be used and an attention mechanism will be utilised to enhance the extraction of global and local information. Statistical information from the global feature map (e.g., features obtained by global average pooling and global maximum pooling) will be analysed to determine which background regions are important and the feature map will be weighted accordingly to enhance the feature representation in these regions.

2. Principles related to human posture estimation.

2.1. **Deep learning.** Deep learning is a type of machine learning technique that learns features from input data by constructing multi-layer neural networks and uses these features to perform tasks such as classification, regression, and clustering. Deep learning extracts features from large amounts of data through automated learning, thus saving time and effort by avoiding the tedious feature engineering process required in traditional machine learning.

Deep learning is growing rapidly and the main reason for this is that deep Convolutional Neural Networks (CNNs) can be applied to a wide range of problems and show good performance. The advent of deep learning has simplified the problems, converted some of the manual labour into automation and increased the efficiency of machine learning. Machine learning techniques have shifted from shallow learning to deep learning. The most important part of deep learning is the convolutional neural network. The core idea of CNN is to first extract image features through the convolutional layer, then reduce the feature dimensions through the pooling layer, and finally use the fully-connected layer to complete the classification, regression and other tasks, as shown in Figure 1.

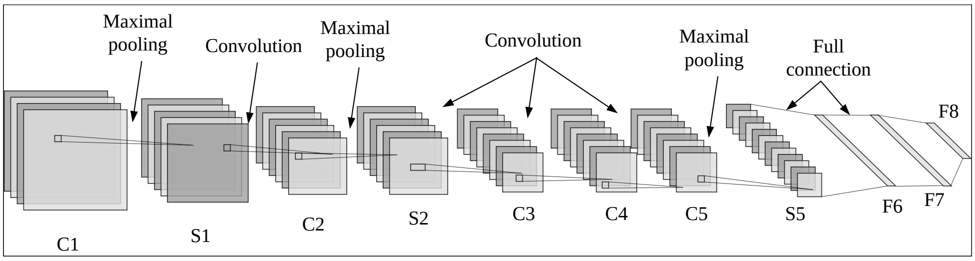

Figure 1. CNN structure

2.2. **Methods for human pose estimation.** Human pose estimation is the process of recovering N joint points from an input RGB image or video. The human skeleton is obtained by detecting the sample human joint points and determining the coordinates of the human joint points on the sample by connecting the human joint point coordinates in a certain order. In recent years, the effectiveness of human posture estimation based on the detection of human skeleton joint points has continued to improve, and its application to the monitoring of dance students' movement postures can provide them with appropriate posture suggestions, thus helping them to improve their learning outcomes.

Currently, traditional-based methods and deep learning-based methods are the two most typical methods for human posture estimation. Traditional methods for human pose estimation are mainly based on an image structure model. The model divides a sample into multiple parts, showing the spatial constraints between them. Traditional methods are time efficient, the downside is that when extracting features you need to manually set up HOG and SHIFT features, which limits the use of the algorithm. However, deep learning techniques do not require manual extraction of features but make predictions based on convolutional features. Currently, there are two approaches for human posture estimation in terms of deep learning, which are top-down approach, and bottom-up approach.

(1) Top-down approach

A top-down approach to human pose estimation is one in which the overall position of the human body is first detected and then individual joint points are localised and the pose estimated. This approach involves a human body detector and a single person pose estimator.

(2) Bottom-up approach.

The bottom-up approach for human pose estimation refers to the process of first extracting all the key points from the image and then achieving pose estimation by combining these key points into meaningful poses. The steps of the bottom-up approach are key point detection, key point screening, and pose combination.

2.3. **High Resolution Networks.** Several dominant convolutional neural network models in the field of human posture estimation, namely: hourglass networks, cascaded pyramid networks, and High Resolution Network (HRNet) [**?**]. HRNet uses higher resolution input images, which can provide more detailed information during the inference process, and thus more accurate human posture estimation. Compared to low-resolution networks, high-resolution networks have better ability to capture subtle changes in body joints and parts. Therefore, HRNet is chosen as the research object in this paper, as shown in Figure 2.

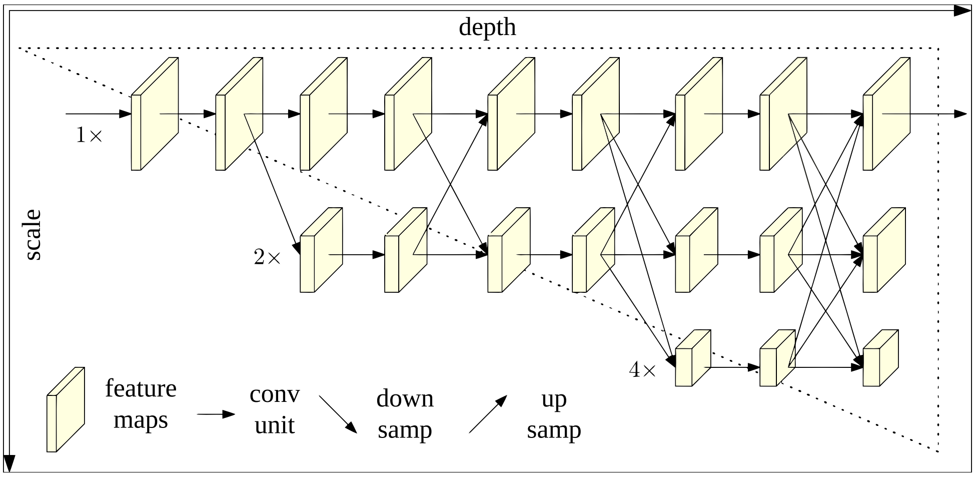

Figure 2. The structure of HRNet

HRNet starts with a Stem structure consisting of two convolutional layers with a step size of 2 to rapidly downsample the input image by a factor of 4, and input the downsampled results into the main part. Incorporating multi-scale feature fusion units into each step of HRNet enhances the network's resilience, allowing different resolution subnetworks to interact with each other. HRNet performs multi-resolution feature fusion at each stage of the network, so that more semantic information can be obtained while maintaining high-resolution features, improving the accuracy of pose estimation. HRNet uses multiple parallel branches, using a method called "layer-by-layer". HRNet uses multiple parallel branches and employs a strategy called "layer-by-layer downsampling" [**?**, **?**], where features are downsampled to different degrees in different branches to gradually reduce the resolution of the feature map while maintaining the high quality of the feature map.

3. **Dance movement pose estimation based on multi-scale attention feature fusion HRNet.**

3.1. **Overall network structure design.** Since the prediction speed of the top-down method in natural scenes changes depending on how many individuals and depends heavily on the accuracy of the human detector, the bottom-up method is gradually becoming the mainstream method. Therefore, in this paper, the HRNet human posture estimation model will be used to obtain the key points of the dance students for posture estimation, the area where the dance student is located in an environment such as a classroom is only a small part of the whole scene and the background is more complex. High-resolution networks typically have deeper network structures that can better learn and encode the spatial information in the input image. This allows the network to better understand the relative positions and relationships between body parts and to better address issues such as occlusion and complex backgrounds in pose estimation.

However, there is room for continued improvement in the accuracy of HRNet. There are two problems with HRNet in the human pose estimation task. Firstly, HRNet needs to detect all the key points of the human body in the input image at the same time, so the multiple scales of the human body included in the image is one of the main reasons limiting HRNet [?]. Then, the up-sampling of HRNet loses part of the feature information and there is no learning parameter, and the lost feature information cannot be compensated by learning. To improve the network's scale invariance, this paper proposes a feature extraction method for background information that can extract global and local features. Secondly, in order to solve the problem of partial feature information loss, CBAM (Convolutional Block Attention Module) is introduced in this paper. The CBAM is used in the concatenated sub-networks of HRNet network with different resolutions, so as to improve the extraction rate of channel and spatial feature information in the network.

The overall structure of the improved HRNet is shown in Figure 3. The main body of the network consists of a total of four stages: stage1, stage2, stage3, and stage4.

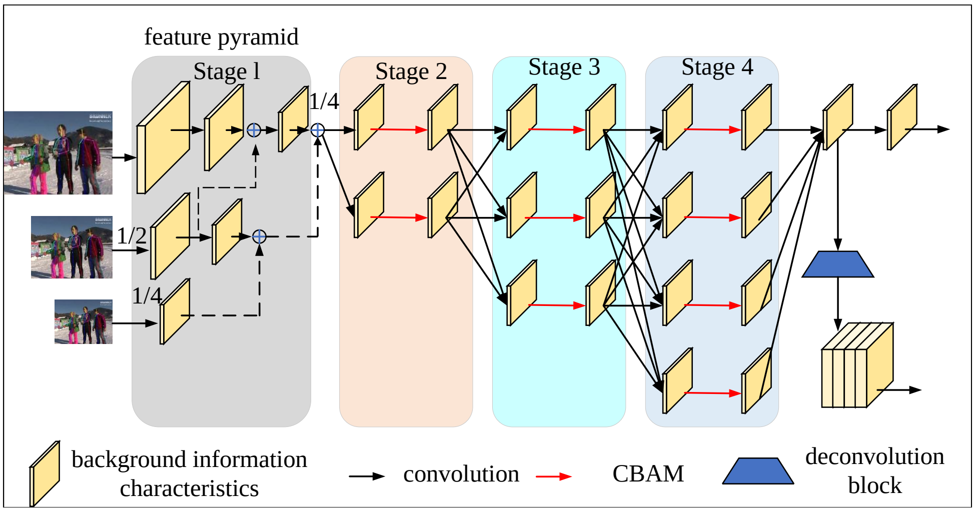

Figure 3. The overall structure of improved HRNet

In stage1, since HRNet uses a Stem structure for fast downsampling, this crude approach loses important detail location information and ignores the significance of low-level characteristics. In order to resolve this issue, this paper introduces a feature pyramid module with three scales in stage1 of HRNet to replace the Stem module. This design can start processing multi-scale information at an early stage of the model, which helps to improve the model's ability to perceive targets of different sizes, especially when performing human pose estimation in complex scenes.

Second, the bottleneck residual blocks in the 1/8, 1/16, and 1/32 resolution subnetworks are replaced using the background information feature module in stage2, stage3, and stage4. Retaining the residual blocks in the 1/4 resolution subnetwork gives the network the ability to extract global and local features. Improve the network's ability to maintain consistent performance across different scales by effectively combining multi-scale information utilizing the attention method. The loss function for the HRNet is shown below:

$$L^h = \frac{1}{N}\sum_{i=1}^{N}\left\|H_i^h - H_i^r\right\|_2^2 \tag{1}$$

where H_i^h and H_i^r are the labelled heat maps at two different scales for the i-th keypoint, respectively.

Assuming that x denotes the 2-dimensional coordinates of the heat map and x_i denotes the true coordinates of the i-th key point.

$$H_i(x) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{\|x - x_i\|_2^2}{2\sigma^2}\right) \tag{2}$$

3.2. **Convolutional module attention.** The CBAM attention mechanism contains two separate sub-modules: CAM (Channel Attention Module) and SAM (Spatial Attention Module). CAM helps the model to identify and reinforce those feature channels that are most critical for the recognition task at hand (e.g., action-posture estimation). SAM learns the weights of spatial locations by examining the feature responses of different channels at each location. In this way, the model is better able to localise to the specific location of a critical action or gesture, which is particularly important for handling complex action-gesture estimation tasks. Therefore, the CBAM attention mechanism can filter unimportant information in the channel to improve the model's representation ability, and spatially learn features at different locations to improve the model's sensory field.

The overall structure of the CBAM attention mechanism is shown in Figure 4. By combining these two attention mechanisms, CBAM enables the model to focus more efficiently and accurately on important features and regions when dealing with complex tasks such as action pose estimation, improving overall performance.

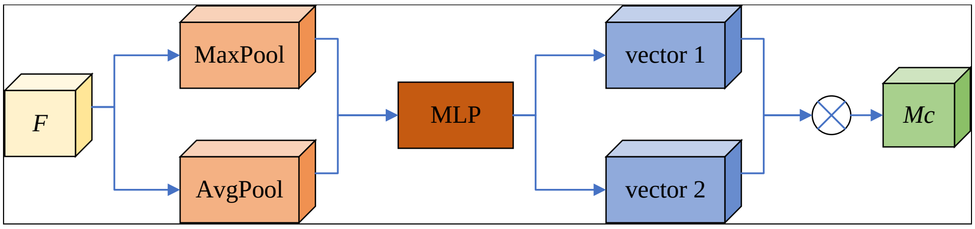

Figure 4. Flowchart of CBAM's attention mechanism

After that, the features that come out of the MLP are multiplied element-wise [?, ?], and then they are activated using a sigmoid function to create the last channel attention feature, M_c. The input features needed by the channel attention module are generated by applying an element-wise multiplication operation on M_c and the input feature map F. The channel attention module is computed as follows:

$$\begin{aligned} M_c(F) &= \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\ &= \sigma\left(W_1\left(W_0\left(F^c_{avg}\right)\right) + W_1\left(W_0\left(F^c_{\max}\right)\right)\right) \end{aligned} \tag{3}$$

These steps are used to calculate the spatial attention module M_s.

$$\begin{aligned} M_s^{(F)} &= \sigma\left(f^{7\times 7}([AvgPool(F); MaxPool(F)])\right) \\ &= \sigma\left(f^{7\times 7}\left(\left[F^s_{avg}; F_{\max^s}\right]\right)\right) \end{aligned} \tag{4}$$

3.3. **Background information feature extraction and fusion.** To improve the network's scale invariance and make it adaptable to human keypoints at different scales, background information characteristics are included. HRNet's high-resolution and low-resolution network branches are utilised to capture features at different scales. The high-resolution branch captures fine-grained local features, while the low-resolution branch captures macroscopic global features. A feature pyramid structure is used to further fuse features at different scales. The feature maps of different resolutions are fused with each other through up-sampling and down-sampling operations to enhance the model's ability to perceive features at different scales, as shown in Figure 5.

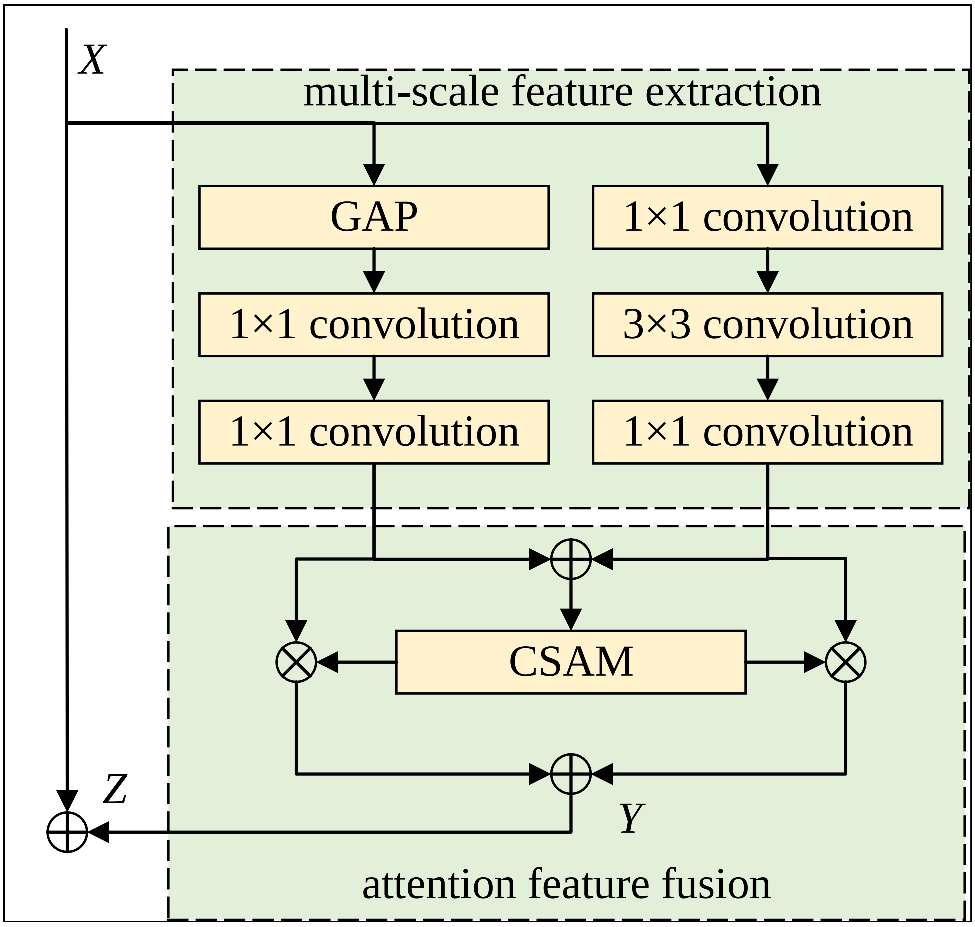

Figure 5. Background information feature extraction module

For the global information, the spatial information of the input feature $X \in \mathbb{R}^{H\times W\times C}$ is firstly compressed using GAP (Global Average Pooling). Then two 1×1 convolutional layers are used to learn the global information, and each convolution is followed by batch normalisation and ReLU nonlinear activation to obtain the global feature $F_g \in \mathbb{R}^{l\times l\times C}$ which contains the global information, and this process can be expressed by Equation 5.

$$F_g = f_g\left(GAP(X); W_g\right) \tag{5}$$

where $f_g(\cdot)$ denotes the global convolution, and W_g is the set of parameters of the global convolution. After ignoring the batch normalisation and ReLU activation function, $GAP(\cdot)$ is calculated as follows.

$$u_c = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} X_c(i, j) \tag{6}$$

where $X_c(i, j)$ denotes the information about the coordinate position (i, j) of the c channel in the input feature X, and u_c is the output corresponding to the c channel.

To improve the extraction of local features that include comprehensive information $F_l \in \mathbb{R}^{H\times W\times C}$, local information extraction makes use of a method that is comparable to the residual.

$$F_l = f_l(X; W_l) \tag{7}$$

where $f_l(\cdot)$ denotes the local convolution, and W_l is the set of parameters for the local convolution.

In order to efficiently emphasize information-rich aspects, feature fusion that makes use of the attention mechanism may be used. Because Attentional Feature Fusion (AFF) only makes use of channel attention, the purpose of this study is to develop an attentional feature fusion module for the purpose of fusing multi-scale features. This is accomplished by adding a spatial attention mechanism to AFF. The attentional feature fusion module that has been presented may be seen in detail in portion (b) of Figure **??**. After receiving the global feature F_g and the local feature F_l produced by the multiscale feature extraction module as inputs, the attention feature fusion module fuses the two features and then transmits them to the CSAM in order to get the attention weights [**?**, **?**]. After that, the attention weights are obtained by multiplying them with the input characteristics in order to acquire the attention features. Finally, the attention features are fused to obtain the background information feature Y. This feature contains both global and local background information and can enhance the important information to strengthen the expressive ability of the feature. Equation 8 can be used to represent this process.

$$Y = M\left(F_g \oplus F_l\right) \otimes F_g + \left(1 - M\left(F_g \oplus F_l\right)\right) \times F_l \tag{8}$$

where $\oplus$ denotes the addition of the broadcasting mechanism, $\otimes$ denotes the multiplication of the broadcasting mechanism, $Y \in \mathbb{R}^{H\times W\times C}$ is the fused background information feature, and $M(\cdot)$ denotes the CSAM. $1 - M(F_g \oplus F_l)$ denotes the dotted line in part (b) of Figure **??**, because the value of the attentional weight $M(F_g \oplus F_l)$ is a real number between 0 and 1, so $1 - M(F_g \oplus F_l)$ also belongs to the real number between 0 and 1, and therefore the F_g and F_l can be Attentional weighting.

CSMA uses both channel attention and spatial attention. Given an intermediate feature X' as input, the attention computation process can be shown as follows.

$$X'' = s\left(M_c\left(X'\right) \oplus M_s\left(X'\right)\right) \times X' \tag{9}$$

where X' is the input feature, X'' is the output feature, $M_c(\cdot)$ is the channel attention, $M_s(\cdot)$ is the spatial attention, and $s(x)$ is the sigmoid activation function.

(1) Using relationships between feature channels to generate channel attention.

The input features are first spatially compressed using global average pooling, and then the dependencies between the channels are learnt through two 1 × 1 convolutional layers as follows.

$$M_c\left(X'\right) = f_c\left(GAP\left(X'\right); W_c\right) \tag{10}$$

where $f_c(\cdot)$ is the channel attention convolution, and W_c is the corresponding set of weight parameters.

(2) Using spatial dependencies of features to generate spatial attention. First, average and maximum pooling is performed along the channel direction, and the average and maximum pooled features are stitched together to form a new feature. In order to acquire spatial attention $M_s\left(X'\right)$, a convolutional layer with dimensions of 3×3 is used to acquire knowledge about the spatial connection between the features, as follows.

$$M_s\left(X'\right) = f_s\left(\left[AP_c\left(X'\right); MP_c\left(X'\right)\right]; W_s\right) \tag{11}$$

where $AP_c(\cdot)$ is the average pooling along the channel direction, $MP_c(\cdot)$ is the maximum pooling along the channel direction, $f_s(\cdot)$ is the spatial attention convolution, and W_s is the corresponding weight parameter.

4. **Experimental validation and analysis.**

4.1. **Experimental development environment.**

The required environment for the experiments in this paper is shown in Table 1. The experiments are mainly based on Pytorch 1.11.0 deep learning network framework to build, train and test the network. The experiments are conducted on a server with RTX 3090 graphics card and Python version 3.8.

The suggested method's efficiency in boosting network scale invariance and correcting for the loss of feature information is verified by experiments done on both the mainstream multi-person pose estimation dataset and the self-constructed dance dataset. The MPII dataset [**?**] is a public dataset consisting of 5602 photographs depicting multi-person postures. Out of these, 3844 images are used for training the network model, while the remaining 1758 images are reserved for model testing. The collection contains almost 28,000 examples of single human postures, each tagged with 15 keypoints. The MPII dataset evaluates the model accuracy using the mAP (Mean Average Precision) of the keypoints. The resolution of all the images in the dance dataset is repaired to 640 × 480. Manual labelling of the human keypoints of the dataset images is carried out with the help of Labelme annotation software, and 15 keypoints are labelled for each person's image, as shown in Figure **??**.

Table 1. Environment required for the experiment

Experimental environment	**Version number and name**
Operating system	Ubuntu 18.04 64-bit
CPU	Intel(R) Xeon(R) Platinum 8157 CPU @2.30GHz
CPU model	NVIDIA GeForce RTX 3090
Memory	24 GB
Python version	3.8
Deep learning acceleration tools	CUDA 11.3
Deep learning frameworks	Pytorch 1.11.0

During the training phase of the MPII dataset, the validation set consisted of 350 randomly chosen samples from the multi-person pose training set. The remaining samples,

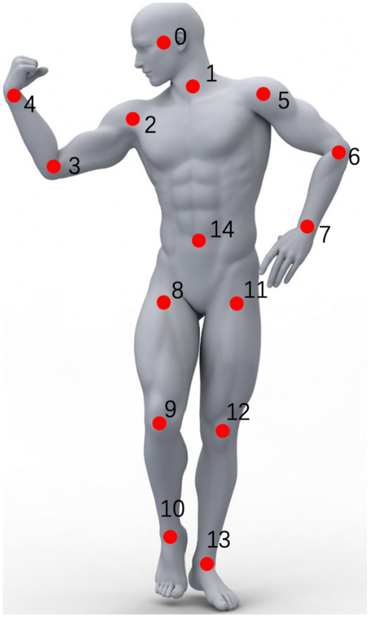

Figure 6. 15 Key Points for the Human Body

which included both multi-person pose and single-person pose samples, were utilized for the training of the model. The RMSprop optimiser was used to update the parameters. The initial learning rate was set to 0.002 and a total of 200 rounds of training were performed. The learning rate was doubled at rounds 100, 160, 220, and 280. The parameter settings were the same on the dance dataset.

In the process of key point extraction for the dance students, the pose estimation network is used to obtain the coordinate information of the joint points of the smart classroom dance students and their heat maps after feature extraction. For the purpose of obtaining the human skeleton, the key point coordinates of the pictures that were collected are joined in a certain sequence in order to generate the synthetic image that is going to be used for posture estimation. Finally, the obtained heat maps are inputted into the built improved HRNet for training and testing.

4.2. **Visualisation results.** The visualisation of the estimation results of human posture on the MPII dataset and dance dataset is shown in Figure 7. It can be seen that the proposed method performs excellently for conventional human posture estimation results. The suggested feature pyramid and backdrop information features have the potential to successfully tackle the scale sensitivity issue that HRNet is experiencing. Additionally, the CBAM attention mechanism is able to further enhance the precision of pose estimation. For the MPII dataset, the improved HRNe method achieves a mAP of 73.6%.

4.3. **Comparison of results with mainstream algorithms.** The improved HRNet is tested on a self-built dance test set and compared with the current mainstream bottom-up approach, and the experimental results are shown in Table 2.

While earlier approaches have shown promising results when working with large-scale, relatively immobile keypoints like the head and shoulder, detection accuracy for smaller, more mobile keypoints like the wrist and ankle remains poor. The rationale for this is because these control points are both spatially malleable and readily obstructed, which requires HRNet to get multi-scale background knowledge and improve the characterisation capability, which is difficult to do simultaneously with previous methods. In this work, HRNet is improved using background information features and CBAM attention

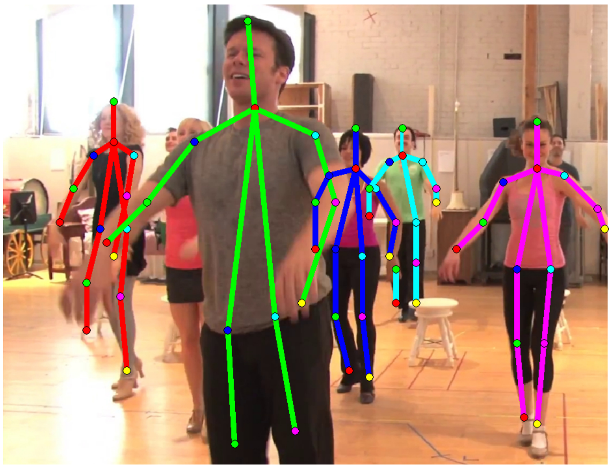

Figure 7. Visualisation of Multi-person pose estimation

Table 2. Results of the bottom-up method on the MPII test set

Method	**Head**	**Shoulders**	**Elbow**	**Wrist**	**Buttocks**	**Lap**	**Ankles**	**mAP**
OpenPose	88.1	84.5	75.2	64.2	73.5	68.1	59.8	73.6
AE	87.7	85.6	69.7	62.7	72.9	71.8	66.7	73.9
PensonLab	90.5	86.9	77.0	66.1	74.7	68.2	61.0	74.9
HRNet	91.4	88.6	78.2	69.1	75.5	70.9	64.0	76.8
SPM	91.1	88.8	79.7	68.9	76.6	71.0	64.8	77.3
HigherHRNet	89.0	86.7	79.7	71.7	76.0	74.2	67.6	77.8
Ours	91.6	88.2	81.4	73.6	76.2	75.7	69.7	79.5

mechanism, which enables HRNet to handle multi-scale human keypoints and improve the model receptive field, and achieves the best results on the dance test set with a mAP of 79.5%. Compared to the HigherHRNet method, this paper's method achieves greater improvements in detection accuracy for keypoints at the elbow, wrist, knee, and ankle, with improvements of 2.13%, 2.65%, 2.02%, and 3.11%, respectively, suggesting that it is better able to handle keypoints in complex environments.

5. **Conclusion.** In this paper, a pose estimation method based on multi-scale attention feature fusion HRNet is proposed. In order to solve the problem of partial feature information loss, CBAM is introduced. The CBAM attention mechanism can improve the model representation in the channel and emphasise the features of spatial location in the spatial dimension to improve the network representation in two ways. CBAM is introduced into the parallel sub-network of HRNet network with different resolutions so as to improve the extraction rate of channel and spatial feature information in the network. A feature extraction method for background information that can extract both global and local features is proposed. A combination of deep and shallow network features will be used and an attention mechanism will be utilised to enhance the extraction of global and local information. Statistical information from the global feature map will be analysed to determine which background regions are important and the feature map will be weighted accordingly to enhance the feature representation in these regions.

The test results on the dance dataset show that compared with the HigherHRNet method, this paper's method has a greater improvement in the detection accuracy of key points such as elbow, wrist, knee, and ankle, with an improvement of 2.13%, 2.65%, 2.02%, and 3.11%, respectively, which suggests that it is able to better deal with the key points in complex environments.

REFERENCES

[1] A. Camurri, I. Lagerlöf, and G. Volpe, "Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques," *International Journal of Human-Computer Studies*, vol. 59, no. 1-2, pp. 213-225, 2003.

[2] E. Karampoula, and H. Panhofer, "The circle in dance movement therapy: A literature review," *The Arts in Psychotherapy*, vol. 58, pp. 27-32, 2018.

[3] E. Van Dyck, D. Moelants, M. Demey, A. Deweppe, P. Coussement, and M. Leman, "The impact of the bass drum on human dance movement," *Music Perception: An Interdisciplinary Journal*, vol. 30, no. 4, pp. 349-359, 2012.

[4] L. M. McGarry, and F. A. Russo, "Mirroring in dance/movement therapy: Potential mechanisms behind empathy enhancement," *The Arts in Psychotherapy*, vol. 38, no. 3, pp. 178-184, 2011.

[5] R. Dieterich-Hartwell, "Dance/movement therapy in the treatment of post traumatic stress: A reference model," *The Arts in Psychotherapy*, vol. 54, pp. 38-46, 2017.

[6] P. P. Capello, "Dance as our source in dance/movement therapy education and practice," *American Journal of Dance Therapy*, vol. 29, pp. 37-50, 2007.

[7] M. Hu, and J. Wang, "Artificial intelligence in dance education: Dance for students with special educational needs," *Technology in Society*, vol. 67, 101784, 2021.

[8] X. Li, M. Karuppiah, and B. Shanmugam, "Psychological perceptual analysis based on dance therapy using artificial intelligence techniques," *International Journal on Artificial Intelligence Tools*, vol. 30, 2140012, 2021.

[9] I. Rallis, A. Voulodimos, N. Bakalos, E. Protopapadakis, N. Doulamis, and A. Doulamis, "Machine learning for intangible cultural heritage: a review of techniques on dance analysis," *Visual Computing for Cultural Heritage*, pp. 103-119, 2020.

[10] M. Joshi, and S. Chakrabarty, "An extensive review of computational dance automation techniques and applications," *Proceedings of the Royal Society A*, vol. 477, no. 2251, 20210071, 2021.

[11] J. Radianti, T. A. Majchrzak, J. Fromm, and I. Wohlgenannt, "A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda," *Computers & Education*, vol. 147, 103778, 2020.

[12] Q. Song, and Y. S. Wook, "Exploration of the application of virtual reality and internet of things in film and television production mode," *Applied Sciences*, vol. 10, no. 10, 3450, 2020.

[13] W. Liu, Q. Bao, Y. Sun, and T. Mei, "Recent advances of monocular 2d and 3d human pose estimation: a deep learning perspective," *ACM Computing Surveys*, vol. 55, no. 4, pp. 1-41, 2022.

[14] N. Sarafianos, B. Boteanu, B. Ionescu, and I. A. Kakadiaris, "3d human pose estimation: A review of the literature and analysis of covariates," *Computer Vision and Image Understanding*, vol. 152, pp. 1-20, 2016.

[15] S. Liu, Y. Li, and G. Hua, "Human pose estimation in video via structured space learning and halfway temporal evaluation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 7, pp. 2029-2038, 2018.

[16] D. C. Luvizon, H. Tabia, and D. Picard, "Human pose regression by combining indirect part detection and contextual information," *Computers & Graphics*, vol. 85, pp. 15-22, 2019.

[17] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive neural networks for high performance skeleton-based human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1963-1978, 2019.

[18] I. Petrov, V. Shakhuro, and A. Konushin, "Deep probabilistic human pose estimation," *IET Computer Vision*, vol. 12, no. 5, pp. 578-585, 2018.

[19] C. Guo, J. Zhou, W. Du, and X. Zhang, "Multi-scale collaborative network for human pose estimation," *International Journal of Humanoid Robotics*, vol. 16, no. 4, 1941003, 2019.

[20] L. Yang, Y. Qin, and X. Zhang, "Lightweight densely connected residual network for human pose estimation," *Journal of Real-Time Image Processing*, vol. 18, pp. 825-837, 2021.

[21] R. Li, A. Yan, S. Yang, D. He, X. Zeng, and H. Liu, "Human Pose Estimation Based on Efficient and Lightweight High-Resolution Network (EL-HRNet)," *Sensors*, vol. 24, no. 2, 396, 2024.

[22] T.-Y. Wu, H. Li, S. Kumari, and C.-M. Chen, "A Spectral Convolutional Neural Network Model Based on Adaptive Fick's Law for Hyperspectral Image Classification," *Computers, Materials & Continua*, vol. 79, no. 1, pp. 19-46, 2024.

[23] F. Zhang, T.-Y. Wu, and G. Zheng, "Video salient region detection model based on wavelet transform and feature comparison," *EURASIP Journal on Image and Video Processing*, vol. 2019, 58, 2019.

[24] F. Zhang, T.-Y. Wu, J.-S. Pan, G. Ding, and Z. Li, "Human motion recognition based on SVM in VR art media interaction environment," *Human-centric Computing and Information Sciences*, vol. 9, 40, 2019.

[25] L. Zhu, H. Zhu, S. Yang, P. Wang, and H. Huang, "Pulmonary nodule detection based on hierarchical-split HRNet and feature pyramid network with atrous convolution," *Biomedical Signal Processing and Control*, vol. 85, 105024, 2023.

[26] H. Xia, L. Wu, Y. Lan, H. Li, and S. Song, "HRNet: A hierarchical recurrent convolution neural network for retinal vessel segmentation," *Multimedia Tools and Applications*, vol. 81, no. 28, 39829-39851, 2022.

[27] Z. Zheng, S. Yu, and S. Jiang, "A domain adaptation method for land use classification based on improved HR-Net," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1-11, 2023.

[28] B. Lu, Y. Sun, Z. Yang, R. Song, H. Jiang, and Y. Liu, "HRNet: 3D object detection network for point cloud with hierarchical refinement," *Pattern Recognition*, vol. 149, 110254, 2024.

[29] M. Li, X. Li, J. Sun, and Y. Dong, "HRNet encoder and dual-branch decoder framework-based scene text recognition model," *International Journal of Antennas and Propagation*, vol. 2022, 2996862, 2022.

[30] Y. Sun, and W. Zheng, "HRNet-and PSPNet-based multiband semantic segmentation of remote sensing images," *Neural Computing and Applications*, vol. 35, no. 12, pp. 8667-8675, 2023.

[31] X. Zhao, C. Song, H. Zhang, X. Sun, and J. Zhao, "HRNet-based automatic identification of photovoltaic module defects using electroluminescence images," *Energy*, vol. 267, 126605, 2023.

[32] Y. Luo, Z. Ou, T. Wan, and J.-M. Guo, "FastNet: Fast high-resolution network for human pose estimation," *Image and Vision Computing*, vol. 119, 104390, 2022.

[33] S. Wei, H. Su, J. Ming, C. Wang, M. Yan, D. Kumar, J. Shi, and X. Zhang, "Precise and robust ship detection for high-resolution SAR imagery based on HR-SDNet," *Remote Sensing*, vol. 12, no. 1, 167, 2020.