





## Intelligent Recognition of Students' Incorrect Movements in Physical Education using Virtual Reality-based Computer Pattern Recognition

Long Wang<sup>1</sup> and Shuping Xu<sup>2\*</sup>

<sup>1,2</sup>Physical Education Department of North China Electric Power University, Beijing 102206, China  
[1 longwang2222@aliyun.com](mailto:longwang2222@aliyun.com), [2 13601259375@163.com](mailto:13601259375@163.com)

Corresponding author: Shuping Xu, [13601259375@163.com](mailto:13601259375@163.com)

**Abstract:** The use of convolutional neural network methods for computer vision research requires a large amount of labeled data. With the emergence of labeled data sets in different fields and the successive release of deep learning open source frameworks, the development of deep learning has been further promoted. This paper combines computer pattern recognition algorithms to intelligently recognize the wrong actions of students, and improve the standard of students' actions. Moreover, starting from the inherent characteristics of the human body, this paper designs a brand multi-person analysis network based on human body posture region extraction and posture correction. In addition, this paper constructs an intelligent recognition system for students' wrong actions based on computer pattern recognition. Through experimental research, it can be known that the intelligent recognition system of students' wrong actions based on computer pattern recognition proposed in this paper has good results.

**Keywords:** computer pattern recognition; students; wrong actions; intelligent recognition; virtual reality

**DOI:** <https://doi.org/10.14733/cadaps.2023.S14.192-207>

### 1 INTRODUCTION

As the magical product of nature, human beings have a wealth of imagination and innovation, and their exploration of themselves and the outside world has never stopped. Vision plays a very important role as one of the ways for humans to receive external information, and human research on vision has never stopped, and making computers have the ability to resemble human vision is also the focus of research. With the development of science and technology, computer vision technology has appeared in people's field of vision. In recent years, smart phones, smart TVs, smart surveillance and other smart terminals have been rapidly popularized, and the number of images has increased rapidly [5]. If every image has to be processed artificially, the workload will be extremely huge, so it becomes particularly important to use a computer to process the information

in the image or video. At the same time, the emergence of innovative achievements such as drones, autonomous driving, virtual reality, and augmented reality all require the support of computer vision technology. With the development of deep learning, convolutional neural networks are widely used in the field of computer vision. The earliest Lenet-5 used convolutional neural networks to recognize handwritten digits, which can effectively cope with the changes brought about by the rotation and translation of handwritten digital images. Afterwards, the Alexnet network shined in the ImageNet image classification competition. Compared with the five-layer network structure composed of Lenet-5's three-layer convolutional layer and two-layer fully connected layers, the Alexnet network has an eight-layer network structure. It includes five layers of convolutional layers and three layers of fully connected layers, and has more parameters, and it also has better performance [12]. With the vigorous development of smart chips, the gradual increase in computing power has promoted the emergence of more network models. At the same time, the depth of the model is getting deeper and deeper, and the structure of the model is getting more and more complex, but the training time is also shortened accordingly [1].

This paper combines the computer pattern recognition algorithm to intelligently recognize the wrong actions of students, improve the standard of students, and improve the intelligent assistance for subsequent students' action correction.

## 2 RELATED WORK

Literature [7] MaskRCNN is based on the framework. The original detection branch is retained to return to the frame box of each person, and then the mask branch is transformed into the extracted region of interest covering a single person for human body analysis, and the final synthesis example can be seen Analytical results of multi-person human bodies. In [8], in order to enhance the feature semantic information and maintain the feature resolution, the proposed area separation sampling is adopted. Human instances often occupy a relatively large proportion of images. If RoIPool operation is still performed on feature maps with coarser resolution, a lot of human details will be lost. Literature [9] adopts the proposed area separation sampling strategy, which uses pyramid features in the RPN stage, but RoIPool only executes it at the finest level. Secondly, in order to obtain more detailed information to distinguish different human body parts in the example, the RoI resolution of the analytical branch is enlarged. Literature [10] proposed a geometry and context encoding module to expand the receptive field and capture the relationship between different parts of the human body. The first part is used to obtain multi-level receptive fields and context information, and the second part is used to learn geometric correlation. Through this module, better quality case perception analysis results can be produced. Literature [11] uses CE2P and MaskR-CNN to design a framework called M-CE2P. MaskR-CNN is mainly used for instance perception. MaskR-CNN extracts all human body sub-images in the input image and adjusts their size to fit the CE2P input size. Then, all instance-level sub-images are fed into CE2P to train the model for partial instance views. In the testing phase, a sub-image with a single human input image instance is extracted through MaskR-CNN, and further fed to the training model for analytical prediction. The predicted confidence map is adjusted to the original size through bilinear interpolation for subsequent prediction of the entire image. Literature [14] fills the confidence map of each sub-image with zeros to keep the same size as the confidence map from the background, and is further merged together through the summation of the elements on the foreground channel and the minimization on the background channel to output the global Analyze the results. CE2P is a simple and effective single-person analysis framework for context embedding and edge perception. The algorithm includes three key modules, which are analyzed in an end-to-end manner: a high-resolution embedding module, which is used to amplify feature maps to restore details; a global context embedding module, which is used to encode multi-scale context information; and analyze object boundaries. Used to integrate the features of the contour prediction module. In the work of [15], the more advanced single-person

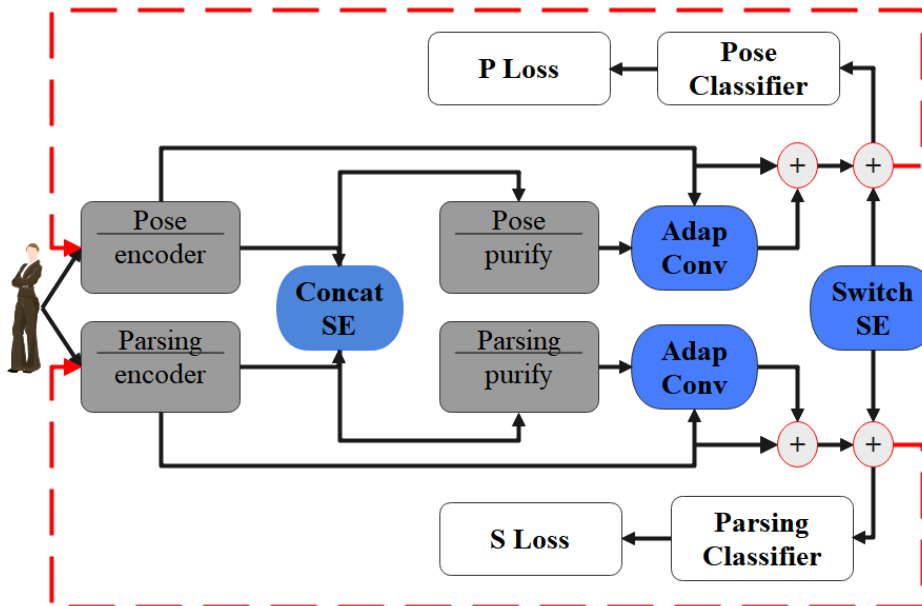
parsing algorithm BraidNet and MaskR-CNN are used for multi-person parsing, which brings a significant performance improvement and exceeds the previous state-of-the-art methods. Literature [16] conducts fine-grained human analysis by learning complementary semantics and details. BraidNet contains two braided structures. The first is a semantic abstract sub-model with a deep and narrow structure, which can learn semantic knowledge through a fully convolutional hierarchical structure to overcome the challenges of diverse human perspectives. In order to capture the details of small objects, a shallow and wide detail-preserving sub-model without down-sampling is designed. This model can reserve enough local structure for small objects. Finally, a set of braid modules is designed between the two sub-models, through which feature information can be exchanged during end-to-end training [17].

Literature [18] reconstructed the instance-level human body analysis virtual reality (VR) environments. from a new perspective, unifying the global body part segmentation and the detection-free detection of instance-aware edge detection into a framework. This method is called partial grouping network (PGN). The region-level pixel grouping of human body parts can be solved by semantic partial segmentation tasks, each pixel is assigned a partial label, and classification attributes are learned. Secondly, given a set of independent human body semantic parts, the instance-level part grouping can determine the instance attribution of all parts according to the predicted instance perception edge. Among them, the parts separated by the instance edges will be classified as different human body instances [20]. PGN seamlessly integrates partial segmentation and edge detection under a unified network. First, it learns the shared feature representation, and then attaches two parallel branches to perform partial segmentation of human body part semantics and instance-aware edge detection. Finally, given human body part segmentation and instance edges, efficient cutting inference can be used to perform a breadth-first search on the line segments obtained by scanning the segmentation map and the edge map together to generate instance-level human body analysis results [21]. Multi-person parsing machine system (MHPM) [22], in response to data challenges and model challenges, proposed a unified multi-person parsing system. In response to the challenge of data scarcity, an MHPMontage model was designed as a data generator in MHPM. The model can intelligently synthesize images from existing data sets to generate real images containing multiple people and their annotations. It is a new type of image synthesis network that uses generative confrontation network (GAN) and spatial transformation network (STN) to learn to automatically synthesize images. Literature [23] designed an MHP solver. The MHP solver converts the input image in the original pixel space into an embedding space, in which the embedding positions of the same person are close to each other, while the embedding distances of different people are far away. At the same time, the MHP solver uses a new global individual push-pull loss (GIPP) method, which operates on the embedding space. Literature [13] can improve the accuracy of multi-person analysis in virtual reality (VR) environments, leading to more seamless and realistic experiences.

### 3 COMPUTER PATTERN RECOGNITION

The network structure is shown in Figure 1. For the input image  $I \in \mathbb{R}^{H \times W \times C}$ , H and W are the height and width of the image, respectively, and  $S = \{s_i\}_{i=1}^{H \times W}$  represents the human body analysis result of the input image I. Among them,  $s_i \in \{0, 1, 2 \dots p\}$  represents the i-th type of semantic pixel annotation, p is the total semantic annotation type, and 0 here represents the background category.  $P = \{(x_i, y_i)\}_{i=1}^N$  represents the human body joint point positioning of the input image I, where  $(x_i, y_i)$  represents the space coordinate of the i-th joint point, and N represents the total number

of joint points. We use the analytical information of the human body and the information of the posture joint points to promote each other to improve accuracy.



**Figure 1:** Structure diagram of the algorithm in this paper.

A network that can optimize the human body analysis process and pose estimation process at the same time is proposed, and  $f_{[\theta, \theta]}(\cdot)$  and  $g_{[\sigma, \sigma]}(\cdot)$  represent analysis and pose models. Then, we get[6]:

$$S^{(t)} = f_{[\theta^{(t)}, \theta^{(t)}]}(F_S^{(t)}), \text{ among } \theta'^{(t)} = g'(F_P^{(t)}, \bar{P}) \quad (1)$$

$$P^{(t)} = g_{[\sigma^{(t)}, \sigma^{(t)}]}(F_P^{(t)}), \text{ among } \sigma'^{(t)} = f'(F_S^{(t)}, \bar{S}) \quad (2)$$

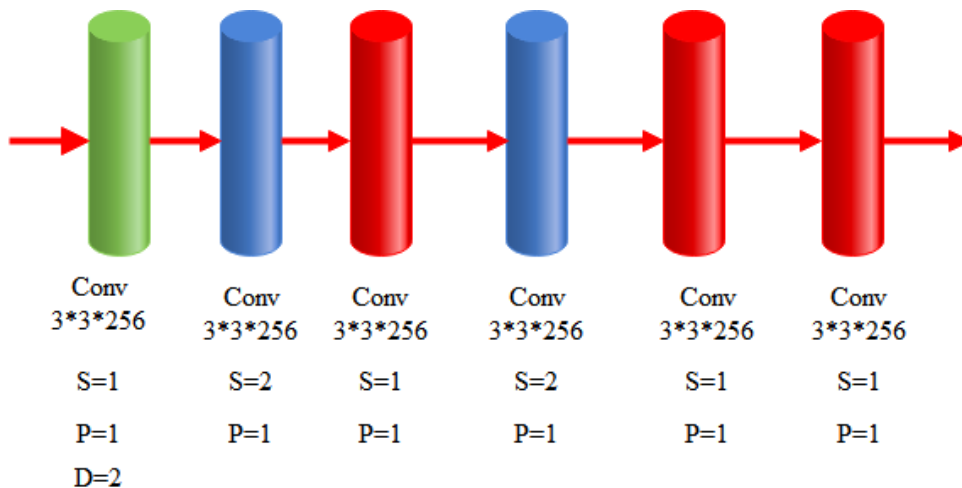
Among them,  $t$  represents the iteration index,  $\bar{P}$  and  $\bar{S}$  respectively represent the true value annotations of the joint points and the human body analysis with respect to the input image  $I$ , and  $F_P^{(t)}$  and  $F_S^{(t)}$  respectively represent the characteristics learned from the joint points and the analytical estimates at the  $t$ -th iteration. The formula emphasizes learning parameters from one task to guide another task. Through feature purification functions  $g'(\cdot)$  and  $f'(\cdot)$ , the parameters  $\theta^{(t)}$  and  $\sigma^{(t)}$  that are suitable for the task of the other party are refined. In addition, the parameters  $\theta^{(t)}$  and  $\sigma^{(t)}$  are closely related to the input image. Different input images will produce different parameters to dynamically adjust the network to modify the model parameters, so that the network fitting has better robustness. Moreover, it iteratively uses the information between the two tasks to guide each other, thereby continuously optimizing the two models[3].

The input image  $I$  is initially sent to the attitude encoder  $E_{\sigma_e^p}^p(\cdot)$  and the analytical encoder  $E_{\theta_e^s}^s(\cdot)$  respectively, which can encode the input  $F_P^{(t)}$  and  $F_S^{(t)}$  into high-level feature representations. Here, the encoder we choose is the hourglass network. After the encoder, we obtain the human bodyanalytic feature  $E_{\theta_e^s}^s(F_S^{(t)})$  and the pose estimation feature  $E_{\sigma_e^p}^p(F_P^{(t)})$ . Then, we send these two features into the stitching SE module that we introduced earlier. The follow-up operation is the same as the classic SE module. After the sigmoid activation function, the original features are multiplied separately. Finally, it is added to the residual module, namely[4]:

$$F_S^{(t)} = SC\left(E_{\theta_e^s}^s(F_S^{(t)}), E_{\sigma_e^p}^p(F_P^{(t)})\right) \tag{3}$$

$$F_P^{(t)} = SC\left(E_{\theta_e^s}^s(F_S^{(t)}), E_{\sigma_e^p}^p(F_P^{(t)})\right) \tag{4}$$

Among them, SC represents the stitched SE module, and  $F_S^{(t)}$  and  $F_P^{(t)}$  are the analytic features and pose estimation features of the human body obtained after stitching the SE modules, respectively. After stitching the SE modules, the pose features and analytical features have the first mutual cooperation to improve their respective accuracy. If we want to further enable the characteristics of the two to help and promote each other, then we need to think about the network structure. When we use the characteristics of one task to completely induce another task, although the two tasks are highly related, we need to think about whether the characteristics of one side completely promote the other side. It is possible that the characteristics of one party will inhibit the expression of the other party, so we need to ensure that the work is completely positive. Then it is necessary to design a new purification module. After the purification module, the characteristics of one party can help the other party's work well. The structure of the purification module is shown in Figure 2.



**Figure 2:** The structure diagram of the purification module.

According to the previously defined purification functions  $g'(\cdot)$  and  $f'(\cdot)$ , we get:

$$\theta^{(t)} = g' \left( F_P^{(t)}, \bar{P} \right) = O_{\theta^{(t)}} \left( F_P^{(t)} \right) \quad (5)$$

$$\sigma^{(t)} = f' \left( F_S^{(t)}, \bar{S} \right) = O_{\sigma^{(t)}} \left( F_S^{(t)} \right) \quad (6)$$

After having the mutual guidance parameters, we refer to the adaptive convolution in the work of MULA, and perform convolution operations on the mutual guidance parameters and the features obtained after passing through the encoder through adaptive convolution. Among them,  $\otimes$  represents the adaptive convolution operation, then[19]:

$$R_P^{(t)} = \theta^{(t)} \otimes F_P^{(t)} \quad (7)$$

$$R_S^{(t)} = \sigma^{(t)} \otimes F_S^{(t)} \quad (8)$$

Among them,  $R_P^{(t)}$  and  $R_S^{(t)}$  respectively represent the posture feature and analytical feature obtained after adaptive convolution operation. Then, the previous feature and the feature after adaptive convolution are added to obtain the respective overall features:

$$\bar{R}_P^{(t)} = R_P^{(t)} + F_P^{(t)} \quad (9)$$

$$\bar{R}_S^{(t)} = R_S^{(t)} + F_S^{(t)} \quad (10)$$

Among them,  $\bar{R}_P^{(t)}$  and  $\bar{R}_S^{(t)}$  respectively represent the posture feature and analytical feature after the addition operation.

After the purification module and adaptive convolution, the mutual assistance of the two features has been very effective, and their respective features have been fully expressed. Next, we use the mutual guiding relationship between the two features to make the final step of fine-tuning the analytical features and posture features. We use the previously designed SE module that exchanges weights to guide the mutual learning between features. The posture feature is multiplied by the analytical feature after passing the SE module of the weight exchange, and the formula is expressed as follows:

$$F_S'^{(t)} = SS \left( \bar{R}_P^{(t)}, \bar{R}_S^{(t)} \right) \quad (11)$$

$$F_P''^{(t)} = SS \left( \bar{R}_P^{(t)}, \bar{R}_S^{(t)} \right) \quad (12)$$

Among them, SS represents the SE module of the exchange weight,  $F_S''^{(t)}$  and  $F_P''^{(t)}$  respectively represent the analytical feature and posture feature after the SE module of the exchange weight. Finally, the obtained features are sent to the classification and refinement module for the calculation of the loss function. There are:

$$L = \sum_{t=1}^T \left( L^S \left( C_{\theta^{(t)}}^S \left( F_S'^{(t)}, \bar{S} \right) \right) + \beta L^P \left( C_{\sigma^{(t)}}^P \left( F_P''^{(t)}, \bar{P} \right) \right) \right) \quad (13)$$

Among them,  $L$  represents the overall loss function,  $L^S$  and  $L^P$  represent the loss functions of the analysis task and the attitude task, respectively.  $C_{\theta^{(i)}}^S$  and  $C_{\sigma^{(i)}}^P$  respectively represent the analytical classification refiner and the pose feature refiner, which are converted from the feature through  $1 \times 1$  convolution and facilitate the calculation of the loss function, and  $\bar{S}$  and  $\bar{P}$  represent the true value of the analytic marker and the pose marker, respectively. Among them,  $\beta$  represents the hyperparameter between the adjustment loss functions.

At this point, the main structure of our network has been introduced. If the entire process of one stage is integrated into one step and sent to the next stage, then there are[2]:

$$F_S^{(t+1)} = M(F_S^{(t)}) \quad (14)$$

$$F_P^{(t+1)} = M(F_P^{(t)}) \quad (15)$$

Among them,  $F_S^{(t+1)}$  and  $F_P^{(t+1)}$  represent the features sent to the next stage after the network processing of the previous stage, and  $M$  represents the various feature processing operations of the network.

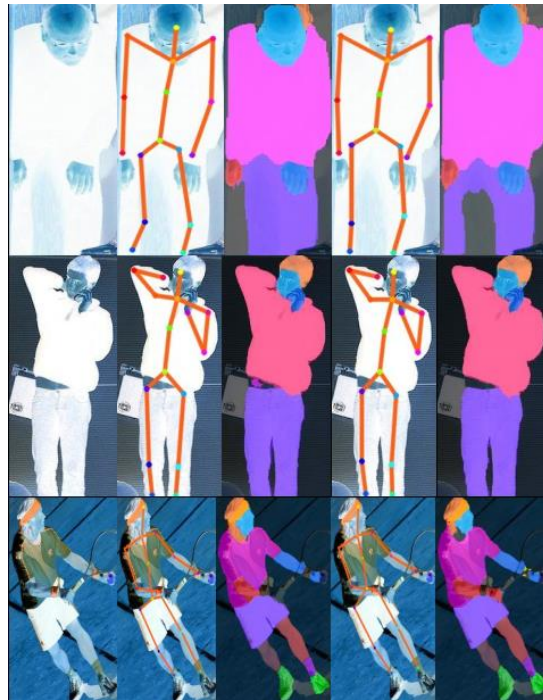
We use the LIP dataset as the training set. The LIP data set is currently the largest data set covering human body analysis and pose estimation. It contains a variety of complex scenes and poses, so it is more challenging. The data set includes 50,462 single-person images of real scenes with 19 types of pixel-level semantic annotations and 16 types of body joint position annotations. Among them, 30462 sheets are used for training, 10000 sheets are used for verification, and the remaining 10000 sheets are used for testing. This article uses random rotation, random zoom, random translation and random folding from the center of the picture, and finally the picture size reaches  $256 \times 256$  after a series of changes and is sent to the above network.

Our task uses Pytorch training on a server with 4 GTX1080TIs, the optimizer uses Adam, the initial learning rate is 0.0025, a total of 250 rounds of training, and at the 150th, 200th, and 230th rounds, the learning rate decays to half of the original. We use mIOU index to evaluate the effect of human body analysis, and PCKh to measure the performance of human body posture estimation. We use cross-entropy loss for analysis tasks, and mean square error loss for attitude tasks. In particular, we choose a value of 0.01 for adjusting the hyperparameter. Figure 3 shows some experimental results. It can be seen that the algorithm in this paper works better in the first row of leg segmentation, the second row of non-clothing misclassification, and the third row of right-hand classification.

We compared the performance of mainstream human body analysis and pose estimation on the LIP dataset. It can be seen that the results of our method are very good (Table 1 and Table 2). Figure 4 is the experimental result of MULA and the method in this paper on the pedestrian re-identification data set Market-1501. It can be seen that the effect of this paper is better.

First of all, we compare the following three methods: the method in this paper, the method without the splicing SE, and the method without the weighted SE module. It can be seen that the splicing SE module has a positive effect on the accuracy of attitude estimation, and the weighted SE module has a positive effect on the accuracy of human body analysis. The reason is that in the initial stage of the network, the segmentation feature is obvious for the improvement of the posture feature, and all the splicing is very effective. At the back of the purification module, the pose feature has been well expressed, and the segmentation feature at this time may inhibit the pose feature.





**Figure 3:** A sample of MULA and the method in this paper on the LIP data set.



**Figure 4:** A sample of MULA and the method in this paper on the LIP data set.

<i>Attitude estimation method</i>	<i>PCKh</i>
<i>BUPTMM-POSE</i>	<i>81.002</i>
<i>PSN</i>	<i>82.921</i>
<i>JPPNet</i>	<i>84.941</i>



<i>MULA</i>	<i>86.355</i>
<i>The method in this paper</i>	<i>87.567</i>

**Table 1:** Comparison of mainstream attitude estimation methods on LIP dataset.

<i>Human body analysis method</i>	<i>mIOU</i>
<i>DeepLabV2</i>	<i>42.016</i>
<i>Attention</i>	<i>43.329</i>
<i>Attention+SSL</i>	<i>45.147</i>
<i>SS-NAN</i>	<i>47.571</i>
<i>MULA</i>	<i>49.49</i>
<i>The method in this paper</i>	<i>50.197</i>

**Table 2:** Comparison of mainstream human body analysis methods on LIP data set.

<i>Experimental method</i>	<i>PCKh</i>	<i>mIOU</i>
<i>The method without the weighted SE module</i>	<i>87.567</i>	<i>49.793</i>
<i>The method without the splicing SE</i>	<i>86.759</i>	<i>49.894</i>
<i>The method in this paper</i>	<i>87.567</i>	<i>50.197</i>

**Table 3:** The first set of ablation experiments.

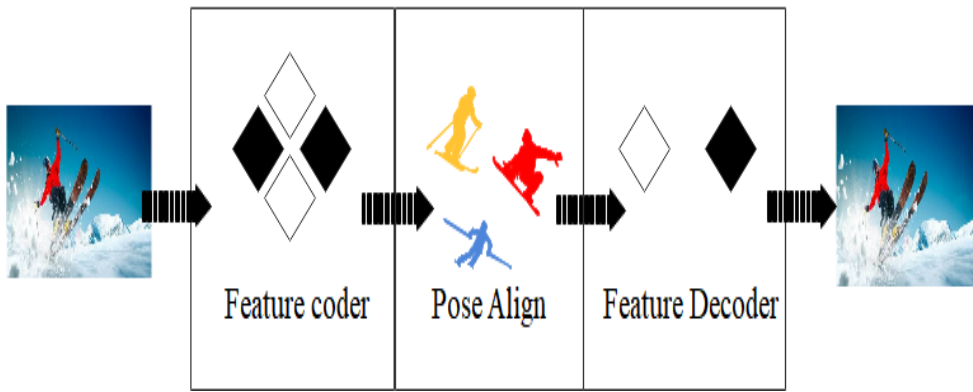
Next, we compare the second set of ablation experiments, that is, the comparison between the new purification module and the method in the original text. It can be seen from the table that the introduction of the hole convolution expands the receptive field and improves the analysis effect, and the use of convolutional downsampling with a step size of 2 is better than the maximum pooling in our task.

<i>Experimental method</i>	<i>PCKh</i>	<i>mIOU</i>
<i>MULA</i>	<i>86.355</i>	<i>49.49</i>
<i>This article only adds purification</i>	<i>86.456</i>	<i>49.591</i>

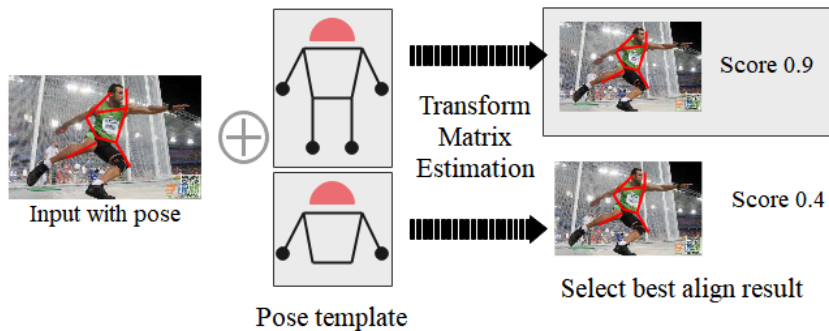
**Table 4:** The second group of ablation experiments.

#### 4 INTELLIGENT RECOGNITION OF SPORTS PLAYERS' WRONG ACTIONS BASED ON COMPUTER PATTERN RECOGNITION

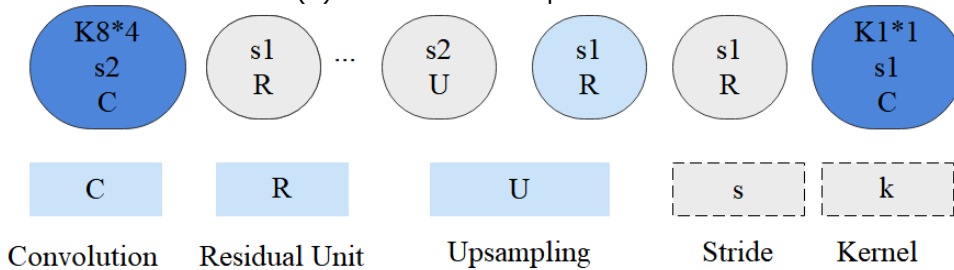
On the one hand, according to the posture of the human body, affine transformation is used to correct the irregularities of the human body. On the other hand, the key points of the human body are used to achieve better segmentation and analysis of the cross-occluded human body, as shown in Figure 5.



(a) Overview of Pose 2par



(b) Pose correction operation

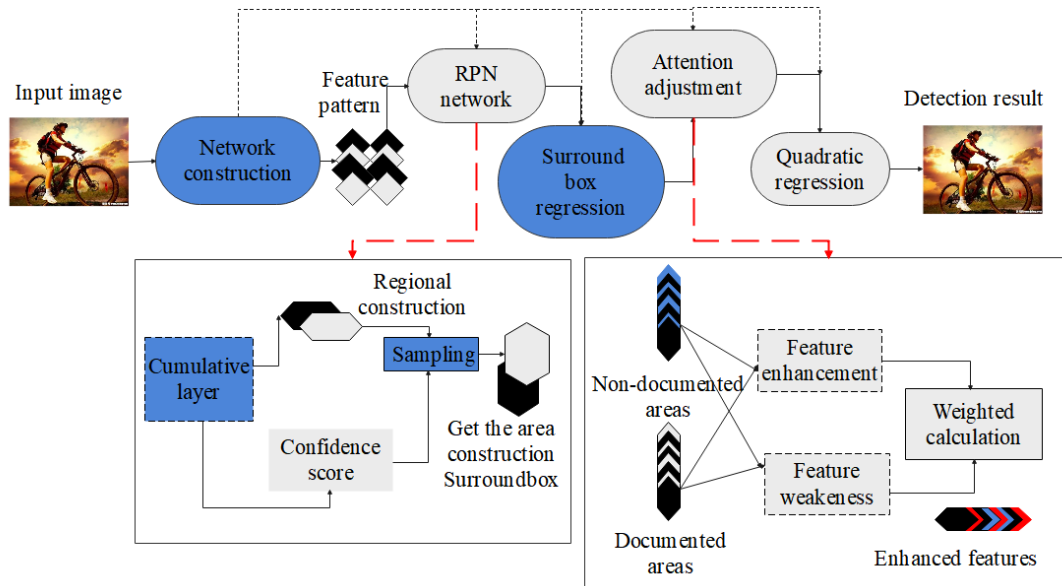


(c) Structure of ParModule

**Figure 5:** Diagram of each structure of the algorithm.

Since the top-down detection method is adopted, it can be observed from the data analysis that the occlusion of the lamp damages the integrity of the human body structure, which causes deviations in the feature extraction during the target detection stage. This paper proposes an attention enhancement mechanism. Based on the previous target detection, the non-occluded area is taken as the key focus area for feature extraction, and feature enhancement operations are performed to reduce the weight value of the feature of the occluded area. In this paper, the processed features are regressed to achieve the effect of feature enhancement. The attention adjustment is after the

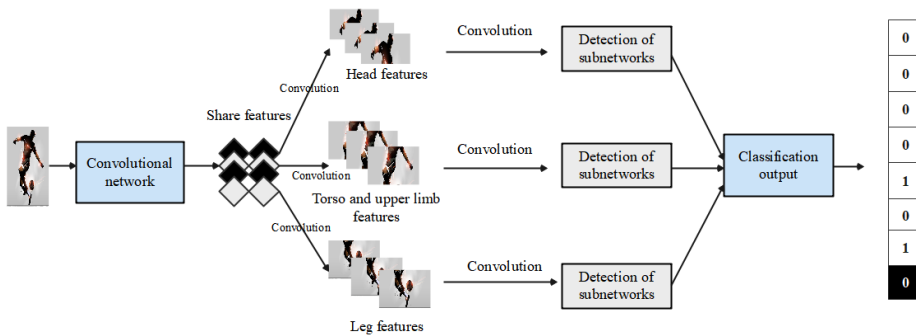
RPN network extracts the bounding box, and then uses the result of the weighted calculation to fine-tune the bounding box. The results with higher confidence are selected, and the detection results are used to obtain attention information, so as to distinguish the occluded area from the non-occluded area. The attention enhancement mechanism is used to perform weighted calculation on the shared feature map obtained in the early stage of the feature extraction network, so as to obtain the enhanced feature map. Using the enhanced feature map, this paper performs a regression operation on the obtained results to further obtain better detection results. The specific structure diagram of attention enhancement is shown in Figure 6.



**Figure 6:** Schematic diagram of attention enhancement structure.

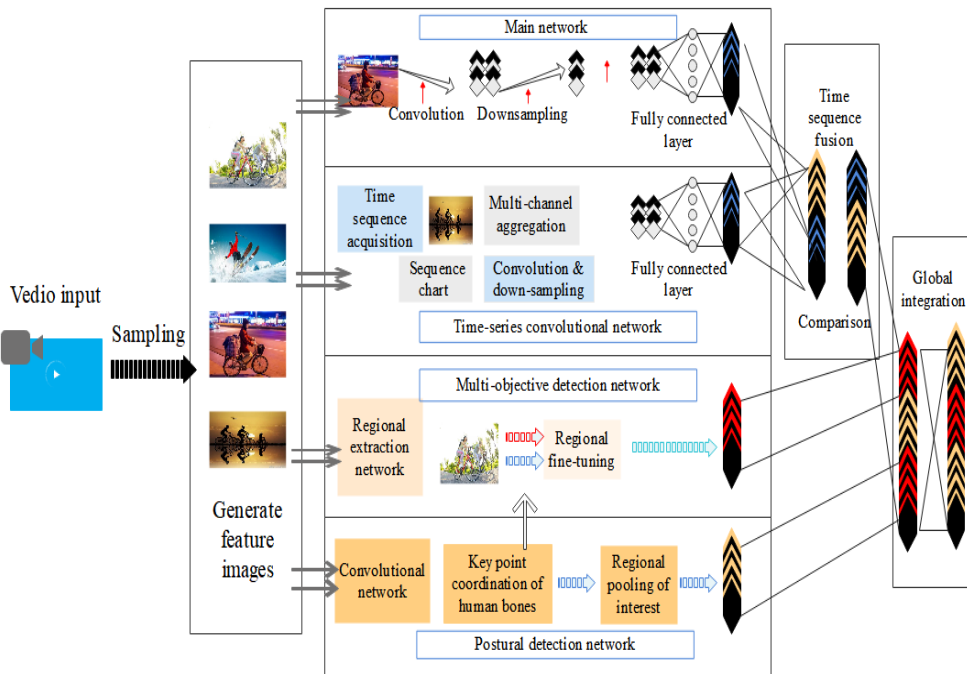
In the training process, different parts can be positioned and classified according to the different label information of the data set, so that the detection sub-network can pay attention to the feature information of different body parts. As shown in Figure 7, the overall network structure is composed of two consecutive phase groups. In the initial data input stage, a general convolutional neural network is used for feature extraction to obtain shared features. After that, the shared features are classified by tagging information, and head features, trunk and upper limb features, and leg features are obtained respectively. After that, the acquired features of each part are respectively transferred to a specific classification sub-network, and after the classification sub-network is calculated, the classification output is finally performed.

This paper distinguishes students from other personnel through target detection, and uses the human body key point detection algorithm to identify the key points of students. Moreover, this paper recognizes the body actions of remote mobilization related to the angle calculation between each key point and the time series. After that, this paper uses the improved target detection algorithm, combined with the relevant skills of small target detection, to further identify other related items. Finally, this paper integrates all the test results to judge the compliance of the athlete's motion detection. The specific detection process is shown in Figure 8.



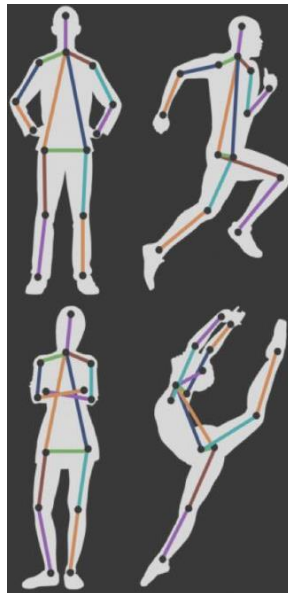
**Figure 7:** Overall structure diagram of the classification process.

### Human action compliance testing



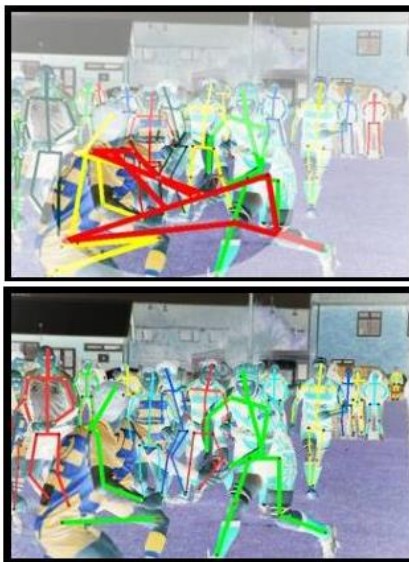
**Figure 8:** The structure diagram of the athlete's action specification detection.

The above constructs the result of the sports athlete's action specification based on computer pattern recognition. Figure 9 below shows the schematic diagram of the key node detection proposed in this paper.



**Figure 9:** Schematic diagram of the key node detection of students.

Figure 10 shows the intelligent recognition results of students' actions obtained by computer pattern recognition in this paper.



**Figure 10:** Schematic diagram of intelligent recognition results of students' actions.

The model proposed in this paper is studied through multiple sets of simulation tests, and the accuracy of the model in this paper on the recognition of students' action and the effect of action correction are calculated, and the results shown in Table 5 below are obtained.

<i>Number</i>	<i>Action recognition</i>	<i>Action correction</i>	<i>Number</i>	<i>Action recognition</i>	<i>Action correction</i>
1	91.00	79.57	31	93.31	84.05
2	91.38	83.87	32	93.51	75.67
3	88.02	79.10	33	88.10	73.44
4	92.50	80.98	34	89.53	80.34
5	88.50	75.14	35	91.67	68.81
6	88.67	69.46	36	94.84	81.52
7	92.71	69.23	37	89.56	77.33
8	88.48	79.11	38	88.49	76.08
9	91.89	75.53	39	93.40	80.59
10	91.82	82.40	40	88.43	72.46
11	90.62	81.93	41	89.43	85.46
12	92.54	78.64	42	90.18	76.37
13	90.86	80.14	43	91.64	80.73
14	94.31	70.17	44	88.01	82.07
15	88.60	83.44	45	90.04	76.82
16	88.28	69.54	46	93.06	79.60
17	94.66	78.97	47	91.09	79.47
18	91.81	74.60	48	93.83	78.45
19	89.07	71.22	49	92.33	75.18
20	91.88	77.88	50	93.34	71.72
21	91.21	74.91	51	94.33	84.88
22	89.62	68.36	52	94.81	69.84
23	89.32	71.81	53	94.23	79.61
24	94.31	81.70	54	92.26	77.50
25	94.45	74.60	55	94.62	82.11
26	91.97	75.35	56	89.16	83.01
27	93.41	76.41	57	94.68	84.12
28	89.11	77.91	58	89.41	73.95
29	92.76	79.58	59	92.59	83.96
30	88.71	75.06	60	91.80	79.94

**Table 5:** The accuracy of the model recognition of students' action and the effect of action correction.

From the above research, it can be seen that the intelligent recognition system of students' wrong actions based on computer pattern recognition proposed in this paper has good results.

## 5 CONCLUSION

As a relatively basic task in computer vision tasks, human body analysis has a wide range of application value. It has been successfully applied in the fields of human body weight recognition, human body posture estimation, video surveillance, automatic recommendation and so on. Considering practical applications, human body analysis in multi-person scenarios has also received more and more attention. Multi-person analysis refers to segmenting a crowd scene image into semantically consistent areas belonging to body parts or clothes, and distinguishing different instances at the same time, so as to assign a semantic label and the instance to which each pixel in the image belongs. In this task, it can be divided into two subtasks: human body analysis and instance perception. This paper combines the computer pattern recognition algorithm to intelligently recognize the wrong actions of students, improve the standard of students, and improve the intelligent assistance for subsequent correction of students' actions. Through experimental research, it can be known that the intelligent recognition system of students' wrong actions based on computer pattern recognition proposed in this paper has good results.

Long Wang, <https://orcid.org/0009-0007-7093-0923>

Shuping Xu, <https://orcid.org/0009-0000-0252-5204>

## REFERENCES

- [1] Aso, K.; Hwang, D. H.; Koike, H.: Portable 3D Human Pose Estimation for Human-Human Interaction using a Chest-Mounted Fisheye Camera, In Augmented Humans Conference 2021, 2021, 116-120. <https://doi.org/10.1145/3458709.3458986>
- [2] Azhand, A.; Rabe, S.; Müller, S.; Sattler, I.; Heimann-Steinert, A.: Algorithm based on one monocular video delivers highly valid and reliable gait parameters, Scientific Reports, 11(1), 2021, 1-10. <https://doi.org/10.1038/s41598-021-93530-z>
- [3] Bakshi, A.; Sheikh, D.; Ansari, Y.; Sharma, C.; Naik, H.: Pose Estimate Based Yoga Instructor, International Journal of Recent Advances in Multidisciplinary Topics, 2(2), 2021, 70-73.
- [4] Colyer, S. L.; Evans, M.; Cosker, D. P.; Salo, A. I.: A review of the evolution of vision-based motion analysis and the integration of advanced computer vision methods towards developing a markerless system, Sports Medicine-open, 4(1), 2018, 1-15. <https://doi.org/10.1186/s40798-018-0139-y>
- [5] Dang, Q.; Yin, J.; Wang, B.; Zheng, W.: Deep learning based 2d human pose estimation: A survey, Tsinghua Science and Technology, 24(6), 2019, 663-676. <https://doi.org/10.26599/TST.2018.9010100>
- [6] Díaz, R. G.; Laamarti, F.; El Saddik, A.: DTCoach: Your Digital Twin Coach on the Edge During COVID-19 and Beyond, IEEE Instrumentation & Measurement Magazine, 24(6), 2021, 22-28. <https://doi.org/10.1109/MIM.2021.9513635>
- [7] Ershadi-Nasab, S.; Noury, E.; Kasaei, S.; Sanaei, E.: Multiple human 3d pose estimation from multiview images, Multimedia Tools and Applications, 77(12), 2018, 15573-15601. <https://doi.org/10.1007/s11042-017-5133-8>
- [8] Gu, R.; Wang, G.; Jiang, Z.; Hwang, J. N.: Multi-person hierarchical 3d pose estimation in natural videos, IEEE Transactions on Circuits and Systems for Video Technology, 30(11), 2019, 4245-4257. <https://doi.org/10.1109/TCSVT.2019.2953678>
- [9] Hua, G.; Li, L.; Liu, S.: Multipath affinity stacked—hourglass networks for human pose estimation, Frontiers of Computer Science, 14(4), 2020, 1-12. <https://doi.org/10.1007/s11704-019-8266-2>



- [10] Li, M.; Zhou, Z.; Liu, X.: Multi-person pose estimation using bounding box constraint and LSTM, *IEEE Transactions on Multimedia*, 21(10), 2019, 2653-2663. <https://doi.org/10.1109/TMM.2019.2903455>
- [11] Liu, S.; Li, Y.; Hua, G.: Human pose estimation in video via structured space learning and halfway temporal evaluation, *IEEE Transactions on Circuits and Systems for Video Technology*, 29(7), 2018, 2029-2038. <https://doi.org/10.1109/TCSVT.2018.2858828>
- [12] Martínez-González, A.; Villamizar, M.; Canévet, O.; Odobez, J. M.: 2 Efficient convolutional neural networks for depth-based multi-person pose estimation, *IEEE Transactions on Circuits and Systems for Video Technology*, 30(11), 2019, 4207-4221. <https://doi.org/10.1109/TCSVT.2019.2952779>
- [13] McNally, W.; Wong, A.; McPhee, J.: Action recognition using deep convolutional neural networks and compressed spatio-temporal pose encodings, *Journal of Computational Vision and Imaging Systems*, 4(1), 2018, 3-3.
- [14] Mehta, D.; Sridhar, S.; Sotnychenko, O.; Rhodin, H.; Shafiei, M.; Seidel, H. P.; Theobalt, C.: Vnect: Real-time 3d human pose estimation with a single rgb camera, *ACM Transactions on Graphics (TOG)*, 36(4), 2017, 1-14. <https://doi.org/10.1145/3072959.3073596>
- [15] Nasr, M.; Ayman, H.; Ebrahim, N.; Osama, R.; Mosaad, N.; Mounir, A.: Realtime Multi-Person 2D Pose Estimation, *International Journal of Advanced Networking and Applications*, 11(6), 2020 4501-4508. <https://doi.org/10.35444/IJANA.2020.11069>
- [16] Nie, X.; Feng, J.; Xing, J.; Xiao, S.; Yan, S.: Hierarchical contextual refinement networks for human pose estimation, *IEEE Transactions on Image Processing*, 28(2), 2018, 924-936. <https://doi.org/10.1109/TIP.2018.2872628>
- [17] Nie, Y.; Lee, J.; Yoon, S.; Park, D. S.: A Multi-Stage Convolution Machine with Scaling and Dilation for Human Pose Estimation, *KSII Transactions on Internet and Information Systems (TIIS)*, 13(6), 2019, 3182-3198. <https://doi.org/10.3837/tiis.2019.06.023>
- [18] Petrov, I.; Shakhuro, V.; Konushin, A.: Deep probabilistic human pose estimation, *IET Computer Vision*, 12(5), 2018, 578-585. <https://doi.org/10.1049/iet-cvi.2017.0382>
- [19] Sárándi, I.; Linder, T.; Arras, K. O.; Leibe, B.: Metrabs: Metric-scale truncation-robust heatmaps for absolute 3d human pose estimation, *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(1), 2020, 16-30. <https://doi.org/10.1109/TBIOM.2020.3037257>
- [20] Szűcs, G.; Tamás, B.: Body part extraction and pose estimation method in rowing videos, *Journal of Computing and Information Technology*, 26(1), 2018, 29-43. <https://doi.org/10.20532/cit.2018.1003802>
- [21] Thành, N. T.; Công, P. T.: An Evaluation of Pose Estimation in Video of Traditional Martial Arts Presentation, *Journal of Research and Development on Information and Communication Technology*, 2019(2), 2019, 114-126. <https://doi.org/10.32913/mic-ict-research.v2019.n2.864>
- [22] Xu, J.; Tasaka, K.; Yamaguchi, M.: Fast and Accurate Whole-Body Pose Estimation in the Wild and Its Applications, *ITE Transactions on Media Technology and Applications*, 9(1), 2021, 63-70. <https://doi.org/10.3169/mta.9.63>
- [23] Zarkeshev, A.; Csiszár, C.: Rescue Method Based on V2X Communication and Human Pose Estimation, *Periodica Polytechnica Civil Engineering*, 63(4), 2019, 1139-1146. <https://doi.org/10.3311/PPci.13861>