Improved C3D Network Model Construction and it's Posture Recognition Study in Swimming Sports

Xiaozhi Peng[1], Yang Li[2*]
[1]Department of Public Course Teaching, Ningbo Polytechnic, Ningbo 315800, China
[2]Department of Physical Education, Hanyang University, Seoul 04763, Korea
Email of corresponding author: lee972023@163.com

Keywords: C3D networks, swimming action recognition, global average pooling, residual networks, deep learning

Received: March 21, 2024

To solve the problem of low recognition accuracy caused by the C3D network limited by the large number of parameters, the study proposes an improved C3D network-based pose recognition model. The improvement of the C3D network is realized by using global average pooling instead of fully connected layer, and the attention residual network on the basis of improved C3D is further designed, and the attention staged residual network model is constructed by introducing the spatio-temporal channel attention mechanism. Comparative validation showed that the improved C3D network increased the accuracy by 13.49% over the C3D network on the HMDB51 dataset. When the various models were compared, it was found that the suggested model, which had an area under the receiver operating characteristic curve as high as 0.98, improved the study's accuracy over the two well-known networks by an average of 14.34%. The accuracy of the proposed model increased the accuracy of the study over the popular networks by an average of 14.50% for the recognition of the postures of all the swimming categories in the homemade swimming sports dataset. The findings show that the number of parameters in the enhanced C3D network proposed in the study has been successfully reduced, and that the attention residual network model based on the enhanced C3D network has a superior application value in sports pose recognition. It also offers some advantages in terms of fine-grainedness and recognition accuracy.

Povzetek: Članek predstavlja izboljšan model omrežja C3D za prepoznavanje položajev v plavalnih športih. Izboljšave vključujejo zamenjavo popolnoma povezanih plasti z globalnim povprečnim združevanjem in uporabo omrežja preostale pozornosti, kar zmanjšuje število parametrov in povečuje natančnost modela. Eksperimentalni rezultati kažejo, da izboljšano omrežje C3D in model ASRNM dosegata visoko natančnost in robustnost v primerjavi z obstoječimi metodami.

1 Introduction

The application of Human posture recognition (HPR) technology is expanding across diverse sectors and settings, particularly in sports, owing to the swift advancement of computer vision technology and associated hardware facilities [1]. Therefore, conducting research on HPR based on visual assistive technology is very important and has a lot of practical application value. Current research on gesture recognition mainly utilizes Deep learning (DL) algorithms, including 2-Dimensional (2D) convolution-based dual-stream networks, 3-Dimensional (3D) convolution-based neural networks, and recurrent convolutional networks [2-3]. Among them, 3D convolutional network can directly carry out spatio-temporal (ST) feature extraction without feature fusion, so scholars at home and abroad mostly utilize it for HPR technology design [4]. Convolutional 3-Dimensional (C3D) networks, as the most commonly used method in 3D networks, have an increased time dimension compared to 2D convolution [5-6]. However, the excessive number of parameters and the relatively simple network structure lead to poor performance accuracy, which is difficult to meet the demand for high accuracy in HPR in current sports. Therefore, the study designed a novel network structure based on the C3D network. Attentional staged residual network modeling (ASRNM) for the improved C3D network was constructed on the basis of firstly replacing the fully connected layer (FCL) with global average pooling (GAP) and replacing the improved C3D network by utilizing Gaussian error linear unit (GELU) activation function, and then experimentally verified it in the HPR of the swimming motion.

The study is divided into four main sections. The first section summarizes the findings of domestic and foreign research on HPR based on vision technology, as well as its drawbacks. In the second part, the swimming posture (S-Pos) recognition model based on the improved C3D network is studied and designed. In the third part, the proposed improved C3D network and S-Pos recognition model are experimented and analyzed. In Part IV, the experimental results are summarized and future research directions are indicated.

As an important computer vision technology, HPR technology provides rich information about body

movement by recognizing and analyzing human posture. Researchers domestically and internationally have made notable advancements in HPR technology. This technology is currently extensively used in industries such as human-computer interaction, intelligent robotics, virtual reality, and medical diagnosis. [7] To realize higher precision 3D human posture reconstruction, Verma and Rajeev proposed a deep architecture model by combining traditional 2D network and 3D network. Additionally, they developed a stack-hourglass network for 2D keypoint heat map prediction, and on the MPII and Human 3.6M datasets, it performed similarly to state-of-the-art techniques [8]. To address the complexity of the convolutional neural network structure and the issue of extracting deep features that only provide global information, Sahoo et al suggested a two-stage residual convolutional network design for learning features from color gesture photos. Using a multi-class support vector machine classifier based on a linear kernel for gesture pose detection allowed for the avoidance of the need for a particular preprocessing step [9]. In an effort to lessen worker load and increase motion detection accuracy for construction industry workers, Chen et al suggested an inherited sensor fusion method for danger prevention. A multi-sensor-based construction site motion identification system was further created using a selective depth detection method based on ordinary depth optimization. The accuracy and effectiveness of body motion detection particular to construction sites was enhanced by merging various signal types to rectify and evaluate worker motion [10]. The ability to tele operate robots can be enhanced by recognizing and reproducing human-like behaviors, however current center-of-mass dynamic balancing is difficult to achieve. To address the issue of variable time series length, Balmik et al. developed a robot-oriented adaptive balancing technique that computes the robot joint angles using pitch and roll control algorithms and uses a proposed 7-layer one-dimensional convolutional neural network to recognize human actions [11].

As an optimization of 3D convolutional network, C3D network brings new ideas to HPR research, and experts and scholars apply it widely in HPR, which promotes the value of HPR technology in real life to a certain extent. For 2D skeleton data, Weng et al. suggested a new 3D graph convolutional network model with ST attention mechanism. The C3D network successfully extracted the ST aspects of the skeleton descriptors, which included joint coordinates, frame differences, and angles, enabling the precise identification and categorization of persons crossing the street [12]. Many labeled data sets and labor are needed for the current skeleton-based recognition techniques, which mostly learn the ideal representation based on human-created criteria. In order to achieve this, Yu et al. presented an adaptive skeleton-based neural network that uses a data-driven methodology to automatically learn the best ST representation. This method effectively allowed memory blocks to learn long-term associations and short-term frame dependencies by encasing a C3D network in a unique attention model [13]. A human aberrant behavior recognition system based on dual-channel C3D and DL was developed by L. Jiang et al. to tightly regulate construction order, work efficiency, and quick response to emergencies at the infrastructure site. A better model was used to integrate this system with a convolutional neural network, yielding validation findings of 98.01% identification rate for particular angles, 97.27% for horizontal angles, and 95.68% for vertical angles [14]. For the challenging problem of recognizing complicated student behavior in films, Jisi and Yin suggested a new feature fusion network for student behavior detection in education. The method combined spatial affine transform network and C3D network with weighted sum method for ST feature fusion, which resulted in superior recognition accuracy over other state-of-the-art algorithms in a wide range of datasets [15]. The above related work is summarized in Table 1.

Table 1: Summary of related work

Methodologies	Data sets	Results	Reference
Reconstruction of 3D poses based on early and late fusion strategies with the introduction of an enhanced stack-hourglass network	MPII and Human 3.6M datasets	Performance comparable to state-of-the-art methods	[8]
Reducing the number of CNN layers and fusing global and local information from different layers	Ha-GRID	The method overcomes the need for a specific pre-processing step	[9]
Image optimization using selective depth detection and construction of a construction site motion recognition system based on sensors	Customized dataset (motion data of 5 adult males aged 20-30 years)	Improving the accuracy and efficiency of detecting construction site-specific body movements	[10]
NAO adaptive balancing technique based on 7-layer	NAO behavior recognition dataset	95% recognition accuracy compared to Hidden Markov	[11]

1D-CNN		Models and Neural Networks	
A novel 3D graph convolutional network model with spatio-temporal attention mechanism	Homemade dataset (ZCP's crosswalk pedestrian dataset); NTU RGB+D dataset	This method outperforms 2D-CNN in recognition results	[12]
Neural network based on adaptive skeleton to automatically learn the optimal spatio-temporal representation through a data-driven approach	MSR-Action-3D dataset; SBU Kinect Interaction dataset; NTU RGB-D dataset; NW-UCLA dataset; UWA3D dataset	State-of-the-art performance was achieved in five challenging benchmarks	[13]
DL and dual-channel C3D based human abnormal behavior recognition system	Self-made dataset	Abnormality recognition rate reaches over 95%	[14]
Fusion of spatio-temporal features through a combination of spatial affine transform networks and C3D networks with a weighted sum approach	HMDB51 dataset; UCF101 dataset; Real student behavior data	Student behavior recognition results are effectively improved and superior to other algorithms	[15]

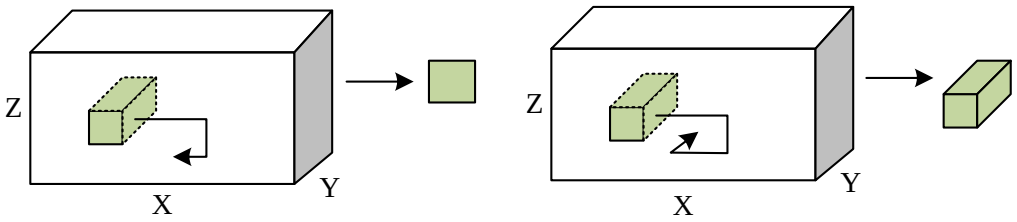
Combined with Table 1, it is evident that scholars both domestically and internationally have conducted extensive research on human gesture recognition technology based on DL. However, as the number of video frames and image pixels continues to increase, current human gesture recognition requires more advanced image features. Meanwhile, the conventional C3D network has a high number of parameters, hindering the effective extraction of deep features from large datasets. Therefore, this study proposes constructing a deep learning recognition model for sports gesture recognition based on an enhanced C3D network. To increase the recognition accuracy of the network model under the enormous number of parameters, the study creatively substitutes the GAP with a FCL and improves the C3D network by replacing the activation function.

2 Swimming posture recognition model construction based on improved C3D network

In order to improve the accuracy of HPR in swimming movement, the study proposes an improved C3D network and further designs an S-Pos recognition model based on the improved C3D network. Firstly, the FCL replacement as well as the activation function replacement are performed on the basis of the C3D network. Secondly, the improved convolutional network is further extended into a fully pre-activated residual structure network, and the ST channel attention focusing mechanism is introduced to construct the S-Pos recognition model.

2.1 Improvement of 3D Convolution-Based C3D network

With the rapid development of the world's swimming sports, swimming is loved and welcomed by more and more people. How to accurately identify and evaluate swimming movements has emerged as a research hotspot in the field of sports monitoring in relation to the instruction and training of swimming sports. Among them, the DL algorithm has become a commonly used method in the research of S-Pos recognition. The C3D network in DL can realize the direct extraction of ST features, which effectively circumvents the defects of the dual-stream network that consumes a large number of resources in order to realize the extraction of the temporal features individually [16-17]. However, the huge number of parameters can cause the convolutional network to be difficult to extract ST features completely during the extraction process, and the effectiveness of feature extraction is limited by the narrow number of convolutional network layers. Therefore, to address the problem of low accuracy of C3D network for HPR, a novel C3D network is proposed in the study. C3D network extracts ST features more efficiently than 2D convolution. Traditional 2D convolution processes video frames by ignoring the relationship between video frame sequences, whereas 3D convolutional feature map (FM) contains not only the information between pixels within a single video frame, but also the correlation between the video frame motion data [18-20]. A comparison of the operational maps of the two convolutions is shown in Figure 1.



(a) 2D convolution operation (b) 3D convolution operation

Figure 1: Comparison of two convolutional network operation graphs

C3D network, as a classical network for 3D convolution, is able to synchronize the preservation of temporal and spatial information of the video action when it performs the convolution operation. Its main convolution formula is shown in equation (1).

$$\aleph_{ij}^{abc} = f\left(\sum_{n}\sum_{x=0}^{X_{i-1}}\sum_{y=0}^{Y_{i-1}}\sum_{z=0}^{Z_{i-1}}\omega_{ijn}^{xyz}\chi_{(i-1)n}^{(a+x)(b+y)(c+z)} + \alpha_{ij}\right) \quad (1)$$

In equation (1), $\aleph_{ij}^{abc}$ denotes the convolution result of the j th convolution kernel (CK) of the i layer in position (a,b,c). a, b and c denote the spatial 3D coordinates, and $f(\bullet)$ denotes the convolution function. X_i denotes the width of the CK in layer i, and Y_i denotes the height. Z_i denotes the depth, and ω_{ijn}^{xyz} denotes the weight of the convolution operation of this layer with the n th FM of the previous layer at position (x,y,z). χ denotes the input value of the previous layer at the same position, and α_{ij} denotes the amount of bias. The structure of the C3D network is relatively simple, consisting mainly of a FCL, 3D convolution, and maximum pooling. Since a time dimension is added to the 3D convolution, it requires a larger number of parameters than the 2D convolutional layers (CLs). This allows multiple 3D CLs to be stacked, driving the full number of parameters of the network to be correspondingly large. At the same time, the network training speed depends on the distribution of transmitted data in the CLs, but in C3D networks the CLs do not have data normalization processing, so the traditional C3D networks are not as effective for recognition in HPR [21-22]. Figure 2 depicts the precise structure of the C3D network.

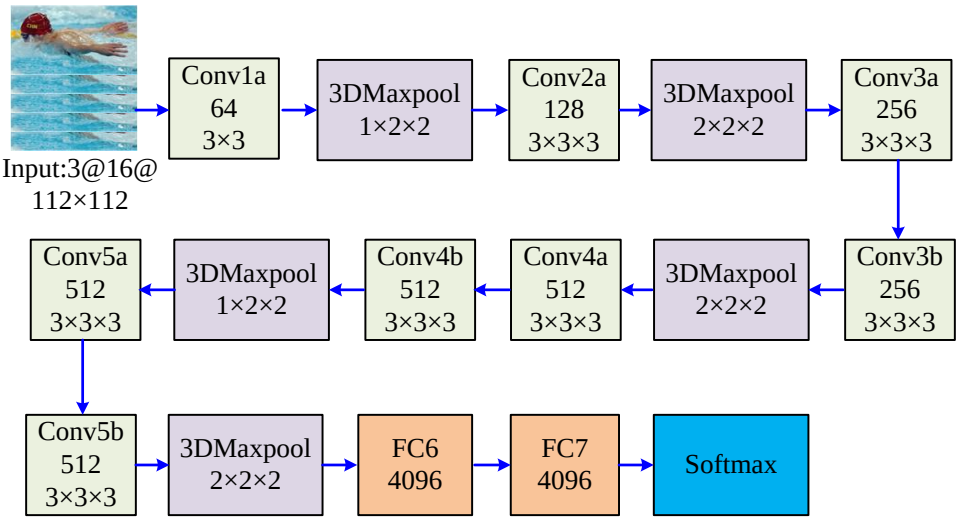


Figure 2: C3D network structure diagram

In Figure 2, the stochastic gradient descent technique optimizes the training of the entire network, and an FCL is used at the network's conclusion. In order for the classifier to classify the data, the extracted features must be mapped to the label space by the FCL. The FCL carries out the feature purification process, meaning that the number of one-dimensional feature vector inputs in the FCL represents a multiple of the number of neurons. The C3D network has an excessive number of parameters and is not suitable for network porting in embedded devices due to every node in the FCL being connected to every other node in the layer before it. Therefore, the study proposes to replace the FCL by utilizing GAP, which reduces the parameter computation by synthesizing the feature information of the weighted average of the FM. In this instance, Figure 3 displays the FCL schematic diagram.

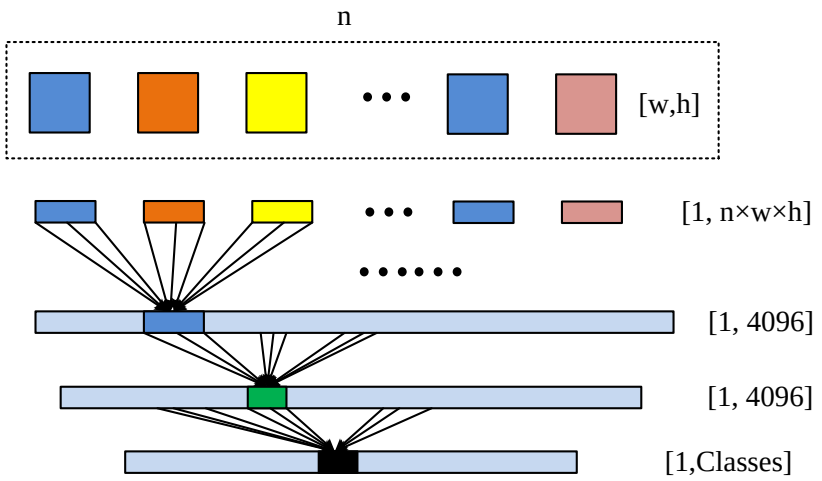


Figure 3: Fully connected layer diagram

GAP itself does not require training parameter computation, by purifying the output features extracted from the CL [23]. Firstly, by sampling the images within each feature channel equally and ensuring that each channel has an output feature image of size 1×1×1. Secondly, the output feature images are transferred to the classification connectivity layer according to the corresponding feature channel. By collecting the spatial information of the feature image using average sampling, GAP's proposed FM may significantly retain the spatial information of the feature image. In addition, the process of transmitting the feature image to the classification layer according to the corresponding channel after GAP processing can effectively increase the mapping connection between the feature image and the classification and weaken the complexity of the FM being interpreted as a category confidence map. The GAP schematic is shown in Figure 4.

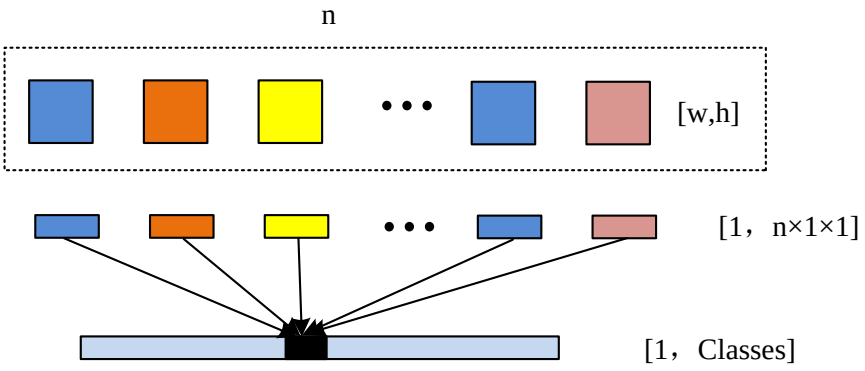


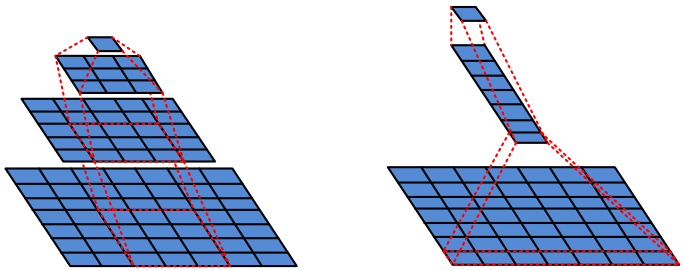
Figure 4: Schematic diagram of GAP

Obviously, only the replacement of FCL is not effective to achieve the improvement of C3D network, the study further introduces 3D dot convolution layer and batch normalization (BN) for enhancing the network's ability to combine features. On this basis, all the activation functions in the network are replaced with GELU functions. The 3D CL is responsible for ST feature extraction in the C3D network, which has an additional temporal dimension than the 2D convolution, and different features are extracted depending on the parameters of the convolutional kernel. Therefore, each CK corresponds to a FM affected by the input FM and the CK, which is calculated as shown in equation (2).

$$\begin{cases} X_{out} = (X_{in} + 2P - F)/S + 1 \\ Y_{out} = (Y_{in} + 2P - F)/S + 1 \\ Z_{out} = J \end{cases} \tag{2}$$

The width, height, and depth of the output FM are indicated by the letters X_{out}, Y_{out}, and Z_{out} in equation (2), respectively. The width and height of the input FM are indicated by X_{in} and Y_{in}, while the pixel padding value of the FM edge is indicated by P. The symbols F, S, and J represent the CK size, step size, and number of CKs, respectively. During the convolution operation on an image, the FMs are locally linked in spatial dimension, all linked in depth, and the weights of neurons at the same depth are shared [24-25]. Therefore,

in order to be able to make the C3D network with a lighter degree of network structure, it is investigated to construct an asymmetric 3D CL by merging and splitting convolutional kernels. The schematic diagram of merging and splitting of CKs is shown in Figure 5.



(a) Merging of convolution kernels

(b) Asymmetric splitting of convolutional kernels

Figure 5: Merging and splitting of convolutional kernels

The study increases the convolutional kernels of all CLs other than the first in the original C3D network to 3:3:3 3D CLs, and combines the three CLs with increased convolutional kernels into a CL with 3:7:7 convolutional kernels. The feature extraction capability of the CL is improved by increasing the region of the CL that affects a specific unit of the network in the input controls. Finally, the CL with 3:7:7 convolutional kernel is asymmetrically disassembled into two asymmetric 3D CLs with convolutional kernels of 3:1:7 and 3:7:1. The CL's weight parameters can be decreased to enhance the image's spatial information and lessen overfitting during network training. Based on the obtained non-stacked 3D CLs, the study further introduces 3D point CLs for cross-channel information fusion and transfers the fused feature information to the next set of asymmetric 3D CLs. Considering that the parameters change continuously during the network training process and the change of data distribution in the previous layer affects the subsequent data distribution, the study utilizes BN to process the input data in the 3D CL. Considering the BN processing as a network layer processing and with trainable parameters distributed between CLs, when the network learns and trains the data in small batches of data, the BN performs normalization with variance of 1 and mean of 0 based on the small batches of data. In this case, the expression formula for the input data is shown in equation (3).

$$\chi_i \leftarrow \frac{\chi_i - u_B}{\sqrt{\theta_B^2 + \varepsilon}} \tag{3}$$

In equation (3), denotes the input data, and χ_i B denotes the set of m data entered in small batches. θ_B^2 denotes the variance, ε denotes the tiny constant value that avoids the denominator equal to 0, and u_B denotes the mean value. Equation (4) displays the formula for figuring out the mean value of the picture feature data.

$$u_B \leftarrow \frac{1}{m}\sum_{i=1}^{m}\chi_i \tag{4}$$

The formula for calculating the variance of the image feature data is shown in equation (5).

$$\theta_B^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(\chi_i - u_B)^2 \tag{5}$$

The expression function of the output normalized result according to GN processing is shown in equation (6).

$$\gamma_i \leftarrow \lambda\chi_i + \beta \tag{6}$$

In equation (6), γ_i denotes the output result of χ_i after BN processing. λ denotes the learnable parameter of scaling and β denotes the learnable parameter of translation. The most widely used activation function in neural networks is the rectified linear unit (ReLU), which can handle difficult nonlinear issues and enhance the low-expression effect of linear functions for challenging issues [26]. Equation (7) displays the particular formula for the function expression.

$$\gamma = \max(0, \chi) = \begin{cases} 0 & (\chi \le 0) \\ \chi & (\chi > 0) \end{cases} \tag{7}$$

In equation (7), γ denotes the output. However, the ReLU function ignores the link between activation and regularization of the data. Again, the study utilizes the GELU function with regularization as the activation function of the improved C3D network [27]. Its specific expression formula is shown in Equation (8).

$$\gamma = \chi \cdot \frac{1}{2}[1 + erf(\chi / \sqrt{2})] \tag{8}$$

In equation (8), $erf(\bullet)$ denotes the Gaussian error expression function. The GELU function is an activation function that compresses the stochastic process, combining the activation ability of nonlinearity with data regularization to achieve a stochastic regularization effect. Combining the above, the overall network architecture of the improved C3D network proposed in the study is shown in Figure 6.

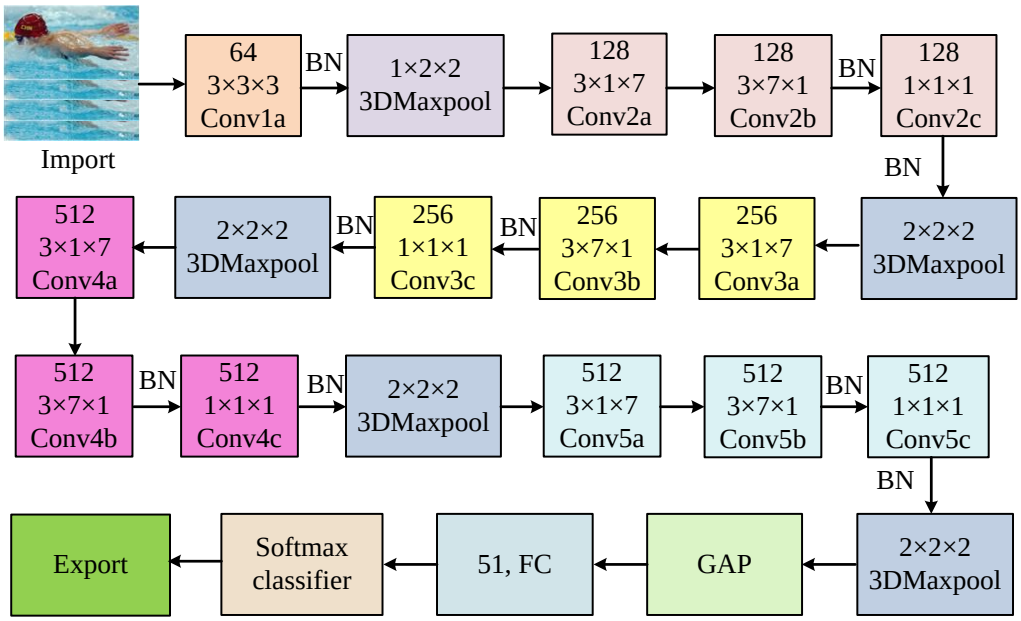


Figure 6: Overall structure of the improved C3D network

Firstly, the input video is segmented into corresponding video frame images after data preprocessing and fed into a 3D CL consisting of 3:3:3 convolutional kernels for comprehensive feature extraction. Next, the extracted data is normalized in small batches using BN, after which the redundant information is removed using the 3D deflation layer. Then, the ST feature information extraction is performed by the asymmetric 3D CL, and then input into the 3D point CL. Finally, all the feature data are passed through the GAP and classification connection layer for discriminative output value calculation, and the final classification result is output in the form of probability through the Softmax classifier.

2.2 ASRNM Based on Improved C3D network

In order to extract deeper ST features of feature images, the study further designs the ASRNM for S-Pos recognition based on the proposed improved C3D network. Firstly, the improved C3D network with fully pre-activated residual's structure is further extended into a FPR network based on the C3D attention, and the Staged Residuals (SR) structure for network optimization. The FPR structure, unlike the original residual structure which can only achieve constant mapping connections on residual blocks, can combine regularization with an activation function as a pre-activation before the information enters the convolutional weights. Based on fully pre-activated residuals-C3D (FPR-C3D), which extends FPR to form a C3D basis in the network, the study replaces the maximum pooling of the network with soft pool (SP). SP obtains the weights of each FM activation value by Softmax exponential normalization, and the final SP output is achieved by weighted summation of the weights for each activation value within the pooling kernel [28]. The weight expression function of the activation values is shown in equation (9).

$$W_g = \frac{\exp(\partial_g)}{\sum_{h \in R} \exp(\partial_h)} \tag{9}$$

In equation (9), W_g denotes the weight assigned to each activation value within the pooling kernel. ∂_g and ∂_h denote the activation values within the pooling kernel of the activation FM, g and h denote the index numbers within the pooling kernel range, and R denotes the pooling kernel range. Equation (10) displays the final formula for the output of SP.

$$\partial = \sum_{g \in R} W_g * \partial_g \tag{10}$$

In equation (10), ∂ denotes the output result after the final SP. Meanwhile, considering that the increase of network depth will negatively affect the BN small BN effect, the study utilizes group normalization (GN) to perform regularization operations on individual 3D CLs. The data normalization process is achieved by calculating the variance and mean used to normalize the features within the grouped channels. The normalization formula is shown in equation (11).

$$\begin{cases} u_g = \dfrac{1}{m}\sum\limits_{k=Q_g}^{m} \chi_k \\ \theta_g = \sqrt{\dfrac{1}{m}\sum\limits_{k=Q_g}^{m} (\chi_k - u_g)^2 + \varepsilon} \\ \chi_g = \dfrac{\chi_g - u_g}{\theta_g} \\ \gamma_g = \lambda\chi_g + \beta \end{cases} \tag{11}$$

In equation (11), Q_g denotes the set of pixels with the mean and variance of the data. The Q_g expression function is shown in equation (12).

$$Q_g = \left\{ k \left| k_D = g_D, \left\lfloor \frac{k_C}{C/T} \right\rfloor = \left\lfloor \frac{g_C}{C/T} \right\rfloor \right. \right\} \tag{12}$$

C and k respectively stand for the channel dimension and the number of input data in equation (12). The batch size is shown by D, and the ability to compute each input data set's mean and variance using the (C, X, Y) axis is indicated by $k_D = g_D$. T denotes the number of groupings, $\lfloor \bullet \rfloor$ denotes rounding down the data, and $\left\lfloor \frac{k_C}{C/T} \right\rfloor = \left\lfloor \frac{g_C}{C/T} \right\rfloor$ denotes that both indexed data are in the same channel grouping. GN regularization computes the input data to circumvent the BN's dependence on memory consumption, which is conducive to improving the accuracy of the network model for HPR. However, before video clips can be classed and identified in the FPR-C3D network for the purpose of recognizing human posture, they must be processed into time-series video frames. Furthermore, the effectiveness of the attention module influences the network's recognition effect [29]. Therefore, the study proposes an improving convolutional block attention model (ICBAM) based on the convolutional block attention model (CBAM), which is extended to the ST domain by adding the temporal dimension, as shown in Figure 7.

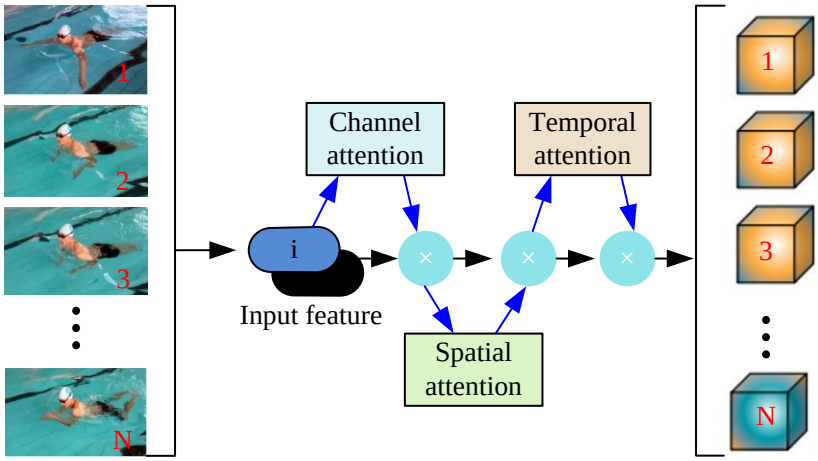


Figure 7: ICBAM principle

In Figure 7, ICBAM first inputs the FMs extracted by 3D convolution and obtains the Identity (ID) channel attention FMs through the attention module of the channels. After adaptive feature refinement, the channel attention FM is obtained by multiplying the ID channel attention FM by the original FM element by element. Equation (13), in particular, displays the calculation formula.

$$P' = V_L(P) \otimes P \tag{13}$$

In equation (13), P denotes the input FM and $V_L(P)$ denotes the ID channel attention FM. V_L denotes the channel attention module and $\otimes$ denotes the element-by-element multiplication. P' denotes the FM obtained by multiplying $V_L(P)$ and P element by element. Pass P' through the spatial attention module to obtain the 2D spatial attention FM $V_o(P')$, multiply $V_o(P')$ with P' element by element to obtain the new adaptive feature refined channel attention FM P''. Equation (14) displays the particular calculating formula.

$$P'' = V_o(P') \otimes P' \tag{14}$$

The P'' is then passed through the temporal attention module $V_\tau(P'')$ in order to distinguish the key video frames. Thus the final obtained FM P''' on the basis of temporal channel attention is shown in equation (15).

$$P''' = V_\tau(P'') \otimes P'' \tag{15}$$

Combined with ICBAM, the proposed FPR-C3D network is further optimized as a full-domain activated residual (FPR-ICBAM-C3D) network based on the C3D attention network. However, FPR is not fully effective in solving the network degradation problem in a huge

number of network layers, and only normalizes the residual branches, which cannot normalize the data in the convolutional weight layer. Therefore, the study introduces SR without a point CL for network optimization. SR enables faster and more efficient transfer of information through the network and enables the synchronization of driving the network to parameter learning and training to optimize the deep network [30-31]. Therefore, the study finally proposes the ASRNM model as shown in Figure 8.

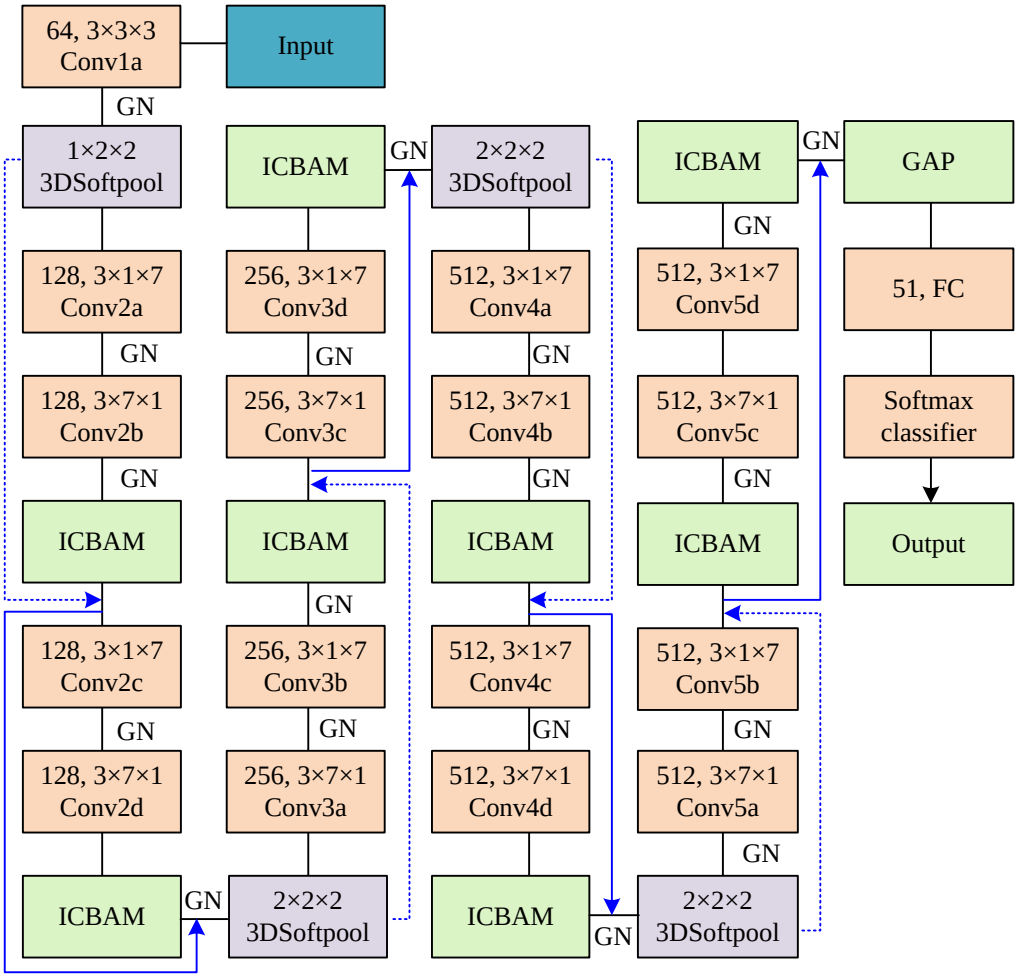


Figure 8: ASRNM network model

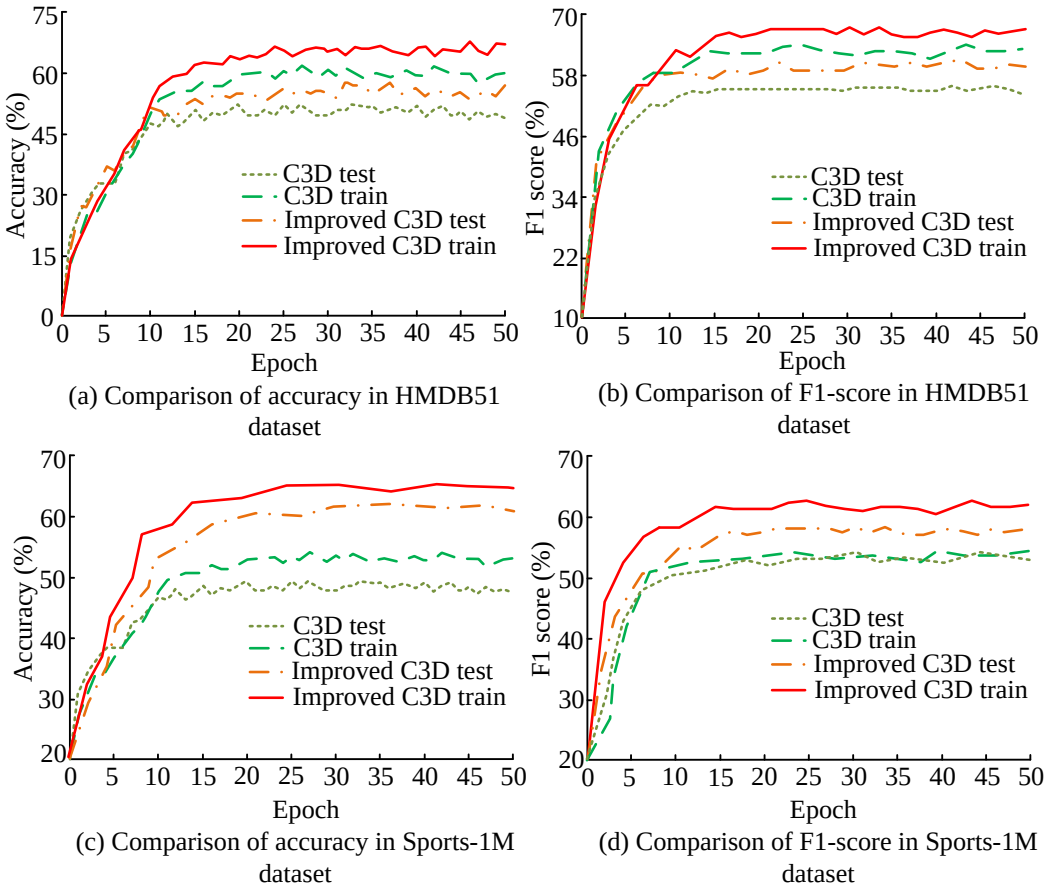
The ASRNM model starts with feature extraction and normalization of the input data by the first 3×3 convolutional and GN layers, and the first step of SR processing. Within the residual block of the initial stage, SP downscaling is performed, followed by feature extraction using asymmetric 3D CLs, GN regularization of the data, and then information extraction of key frames using ICBAM. The initial stage residual block processes the data and then enters the end residual block of SR. Based on the whole SR processed data obtained, the above processing operations are repeated in the next part of SR until all SRs are passed. Finally, the extracted feature information is passed through GAP, Classification Connection Layer and Softmax for the final recognition result output.

3 Experimental analysis of swimming posture recognition model based on improved C3D Network

To validate the effectiveness of the proposed S-Pos recognition technique in the study, the proposed improved C3D network is firstly tested for comparison in the dataset. Based on this, additional performance validation of the ASRNM model is carried out in order to assess its efficacy in comparison with the currently in use network models in the sports dataset. Finally, swimming action pose recognition experiments are conducted in the dataset of swimming sports.

3.1 Improved C3D network validation

The study used the adaptive moment estimation algorithm for the network model training optimization algorithm and set the network iteration period to 50 times, the initial learning rate to 0.00001, the number of groups for the normalized grouping to 32, the weight decay parameter to 5×10-4, and the batch size during the training process to 8 in order to experimentally validate the performance of the improved C3D network. The HMDB51 dataset and Sports-1M dataset are pre-processed based on the above parameters, and then the improved C3D network model is used to compare the validity with the traditional C3D model. The HMDB51 dataset is mainly derived from movie clips and short videos uploaded online by netizens, with a total of 6,766 video data, most of which suffer from camera jitter, poor shooting angles, and low-quality video frame defects, and the use of which better demonstrates the reliability of the network model proposed in the study. The Sports-1M dataset is a collection of sports video clips classified into 487 action categories, totaling 1,100,000 clips. It is a useful tool for validating the effectiveness of the network model proposed in the study. Figure 9 displays the accuracy and F1 value change curves for both datasets.



(a) Comparison of accuracy in HMDB51 dataset

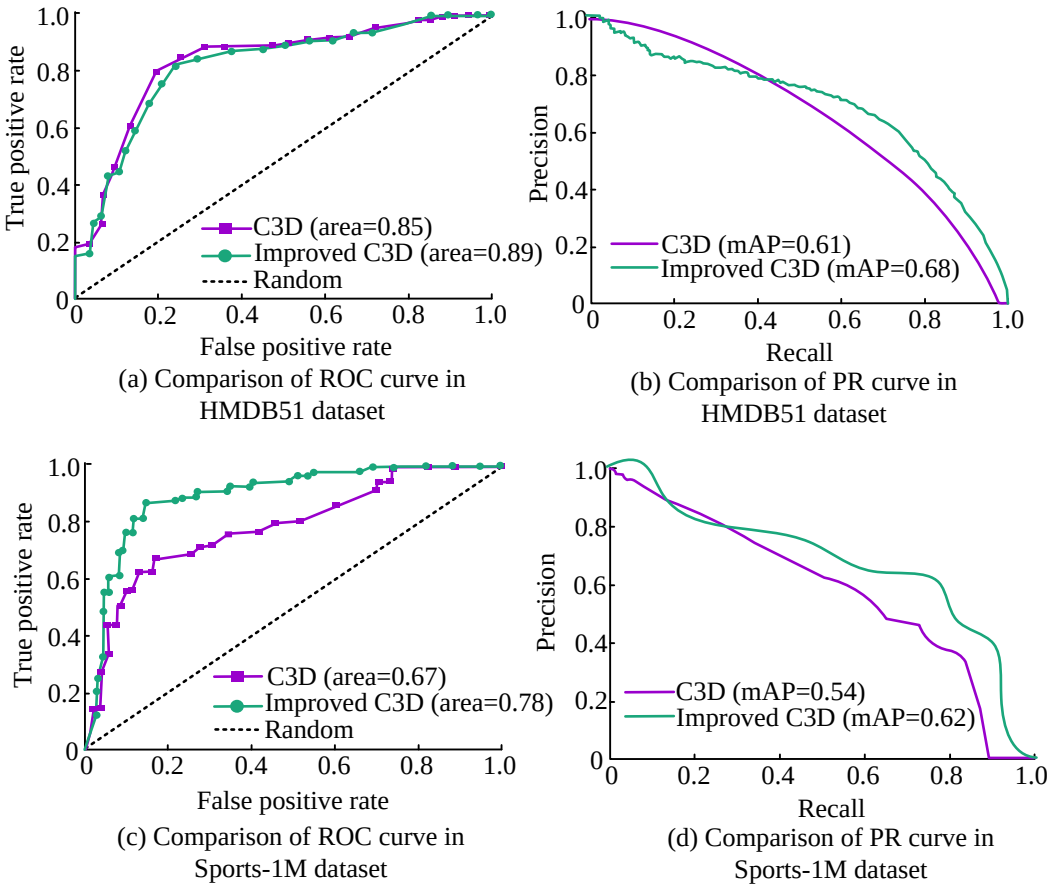
(b) Comparison of F1-score in HMDB51 dataset

(c) Comparison of accuracy in Sports-1M dataset

(d) Comparison of F1-score in Sports-1M dataset

Figure 9: Comparison of accuracy and F1 score in HMDB51 dataset and Sports-1M dataset

In Figure 9(a), the accuracy of both testing and training of the improved C3D network is improved in different ways compared to the traditional C3D network. Compared to the unimproved C3D training, the improved C3D training improved the accuracy by 15.15% after 50 iterations, while the test accuracy improved by 13.49%. Figure 9(b) presents a comparison of the F1 values of the two convolutional networks. Testing and training results indicate that the upgraded C3D network's F1 values outperform the C3D network. This suggests that the model's performance can be enhanced by the enhanced C3D network that the study suggests. After 10 iterations, the improved C3D network model shows a leveling off trend earlier than the C3D network, which indicates that the improved C3D network model can find the optimal solution quickly. The improved C3D network model's superiority is evident when comparing the accuracy and F1 values of both networks in the Sports-1M dataset, as shown in Figures 9(c) and 9(d). The study also compares the C3D network with the enhanced C3D network's receiver operating characteristic (ROC) and precision recall (PR) curves; the comparison's findings are displayed in Figure 10.



(a) Comparison of ROC curve in HMDB51 dataset

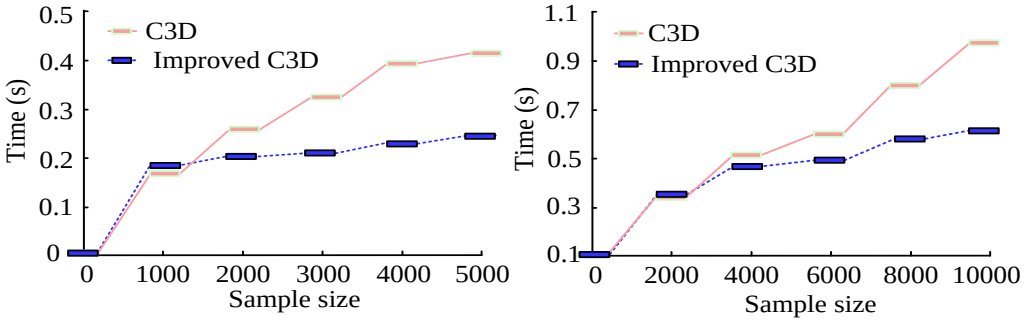
(b) Comparison of PR curve in HMDB51 dataset

(c) Comparison of ROC curve in Sports-1M dataset

(d) Comparison of PR curve in Sports-1M dataset

Figure 10: Comparison of ROC and PR curve

As compared to C3D, the revised C3D network model's area under the ROC curve in Figure 10(a) rises by 4.71%, indicating an improvement in the model's accuracy. Better model accuracy is shown by a greater mean average precision (mAP) of the area under the PR curve. The mAP value of the enhanced C3D network model in Figure 10(b) is 0.68, a 11.48% increase over C3D. Figure 10(c) shows the ROC curves of the two network models in Sports-1M. The improved C3D network model has a higher curve area than C3D. Figure 10(d) illustrates that the improved C3D network model has a 14.81% increase in mAP value over C3D. The sample data distribution has less of an impact on the ROC curve, and the PR curve more accurately represents the performance of the model with a broader sample data distribution. This suggests that the study's enhanced C3D network can successfully address the shortcomings of the conventional C3D network, which has an inadequate classification impact because of an excessive number of characteristics. Finally, the study further compares the training time overhead of the two models, the improved C3D network and the traditional C3D network, in the two datasets, as shown in Figure 11.



(a) Training time overhead in HMDB51 dataset

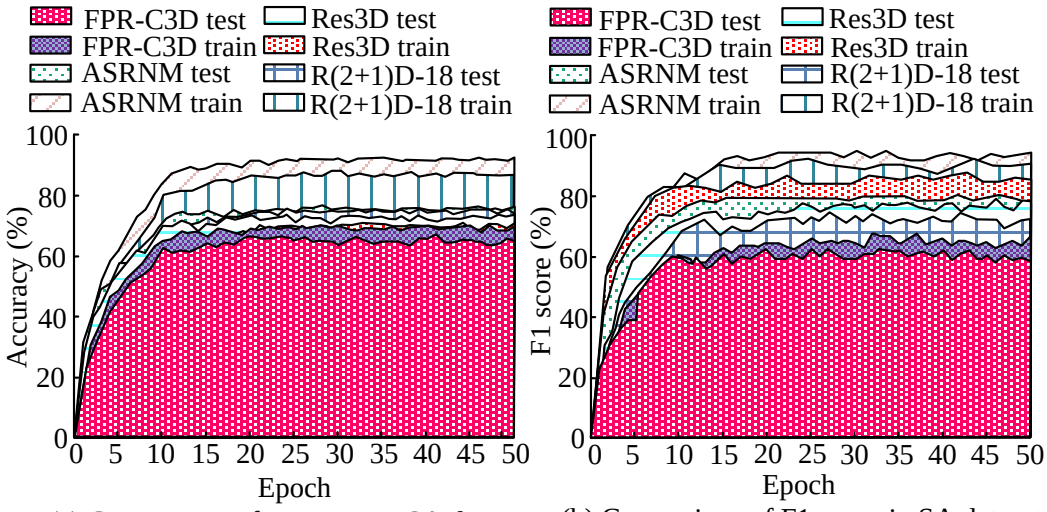
(b) Training time overhead in Sports-1M dataset

Figure 11: Comparison of the training time overhead of the two network models in the dataset

Figure 11(a) shows that the training time overhead of the improved C3D network model is significantly lower than that of the C3D network model in the HMDB51 dataset. As the number of samples increases, the time overhead of both network models also increases, but the increase is smaller for the improved C3D network model. Figure 11(b) shows that the time overhead of the improved C3D network model is slightly higher than that of C3D when the sample data is at 2000. This may be due to the fact that the improved C3D network takes some time to adapt to the computation of the samples after the reduction of the number of references. However, as the data samples increase, the increase in the time of the improved C3D network model decreases. By improving the number of parameters and the activation function of the C3D network, the use of GELU as the activation function facilitates the generalization of the model. This results in a reduction in the model time overhead, leading to faster and more accurate recognition of the swimming action.

3.2 Verification of ASRNM based on improved C3D network

The improved C3D network demonstrated its good classification performance in the HMDB51 dataset, but considering the one-sidedness of a single dataset and the effectiveness of S-Pos identification, the study utilized the kinetic400 dataset and the UCF101 dataset to build a sports action (SA) dataset for the performance validation of the FPR-C3D network, ASRNM model. performance validation. 5302 video clips in all, broken down into 43 categories with 108 clips in each, make up the SA dataset. The suggested network model is evaluated in terms of accuracy and F1 value performance against the currently in use networks using the SA dataset; the comparison results are displayed in Figure 12.



(a) Comparison of accuracy in SA dataset (b) Comparison of F1-score in SA dataset

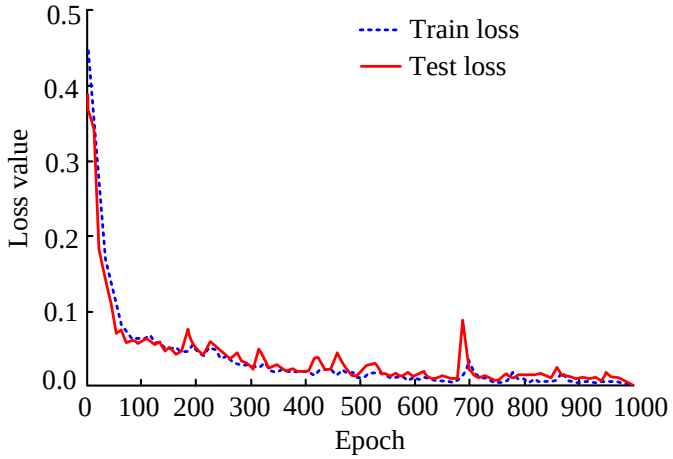
Figure 12: Comparison of accuracy and F1 score in SA dataset

In Figure 12(a), the research-proposed ASRNM model has the fastest and largest increase in accuracy during training, and the accuracy after 50 iterations is the highest among all compared models. The FPR-C3D network model has the lowest accuracy among the four methods, which confirms the need for the study to propose optimized convolutional networks using the SR residual structure. Res3D and R(2+1)D-18, the more popular networks, had higher accuracy than the FPR-C3D network, but the accuracy of the ASRNM model training increased by 24.37% and 4.31% over the two popular networks, respectively. The F1 values of the four models are compared in Figure 12(b), which demonstrates that the ASRNM model continues to be the most superior in terms of both speed and magnitude of improvement. The F1 curves of the four models in the SA dataset improve more quickly, and the change in their F1 values tends to stabilize when the iteration is 10 times. When the training went through 50 iterations, the ASRNM model increased the F1 value by 38.67% over the FPR-C3D network. The change in accuracy and F1 value curves also shows that the ASRNM model has less fluctuation, which indicates its superior generalization. The four models' AUC, mAP, number of model parameters (Params), and floating-point perations per second (Flops) are compared in order to further demonstrate the superiority of the model suggested in the study. The precise findings are displayed in Table 2.

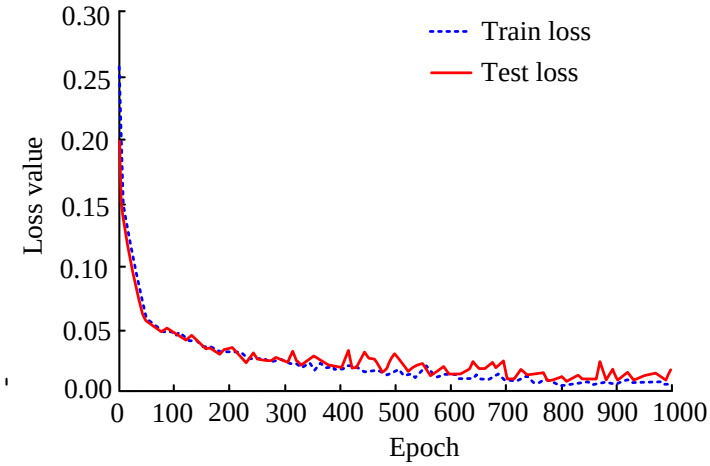
Table 2: Comparison of experimental results of different models on the SA dataset

Model	AUC	mAP	Params (×106)	Flops (×109)
C3D	0.92	0.64	78.21	38.66
Improved C3D	0.94	0.69	26.98	40.85
FPR-C3D	0.95	0.72	47.95	45.34
ASRNM	0.98	0.88	47.95	45.37
Res3D	0.97	0.79	33.20	37.54
R(2+1)D-18	0.96	0.86	33.31	38.75

It is evident from comparing the models' AUC and mAP values that the ASRNM model performs the best on the SA dataset. Its mAP value rises by 2.33%-37.50% across multiple approaches, and its AUC value reaches as high as 0.98. This suggests that in the SA dataset, the ASRNM model performs better overall than the Res3D and R(2+1)D-18 networks. The ASRNM model outperforms the other two models in terms of computational complexity and parameter count due to the inclusion of a ST channel attention mechanism in the network. This mechanism increases the number of parameters, which in turn increases the computing complexity and Flops value. However, comparing with the traditional C3D network, the ASRNM model parameter computation is reduced by 38.69% and the computational complexity is only increased by 17.36%. Furthermore, the study examines how the ASRNM model's loss value changes after 1000 iterations in the HMDB51 and SA datasets. The precise findings are displayed in Figure 13.



(a) Loss curve of ASRNM model on HMDB51 dataset

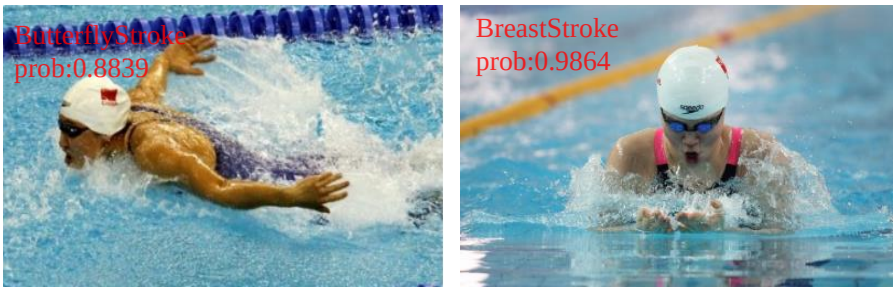


(b) Loss curve of ASRNM model on AS dataset

Figure 13: Loss curves of ASRNM model in HMDB51 dataset and SA dataset

In the HMDB51 dataset, the ASRNM model converges after 900 iterations, as shown in Figure 13(a), and the loss value tends to be near 0. The loss value of the SA model is almost equal to 0.01 after 800 iterations, but as the number of iterations rises, it gets closer and closer to 0. On the whole, the initial loss value of the ASRNM model is relatively low, which indicates that its model classification performance is better and has

superior classification effect in sports pose recognition. When the aforementioned information is combined, it becomes clear that the study's ASRNM model has a strong classification capability for sports gesture identification. As a result, the research produced a visual representation of the prediction outcomes for the SA dataset based on the confusion matrix of the ASRNM model, as seen in Figure 14.



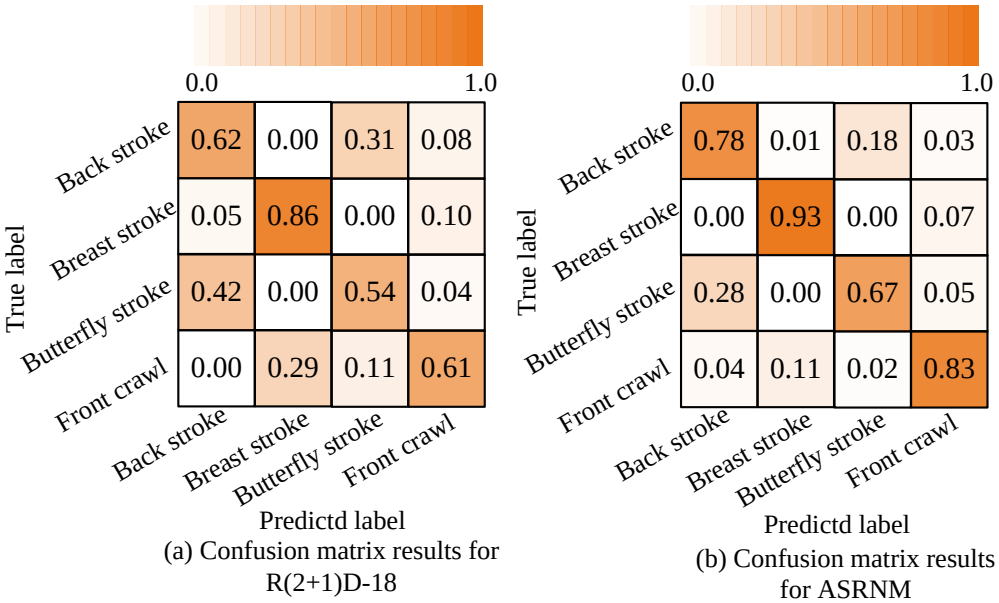
(a) Butterfly stroke recognition results (b) Breast stroke recognition results

Figure 14: Visualization of prediction results

Figures 14(a) and (b) demonstrate the recognition of the butterfly and breaststroke, respectively. When there is intra-image reference blurring or video blurring, the ASRNM model detects it correctly.

3.3 Experimental validation of gesture recognition in swimming

Based on the above validation results, the study further compares the fine-grained recognition effects of the ASRNM model and the R(2+1)D-18 network in the swimming motion (Swim) dataset, which is a dataset constructed from four swimming motions, namely, breaststroke, butterfly, backstroke, and freestyle in the SA dataset, with 494 video clips, and utilizes this dataset to perform the fine-grained recognition of swimming motions. The reliability and accuracy of the model's classification in environments such as light and water refraction can be verified, as well as the recognition accuracy of gestures with similar movements. The recognition results of the two algorithms on the Swim dataset are shown in Figure 15.



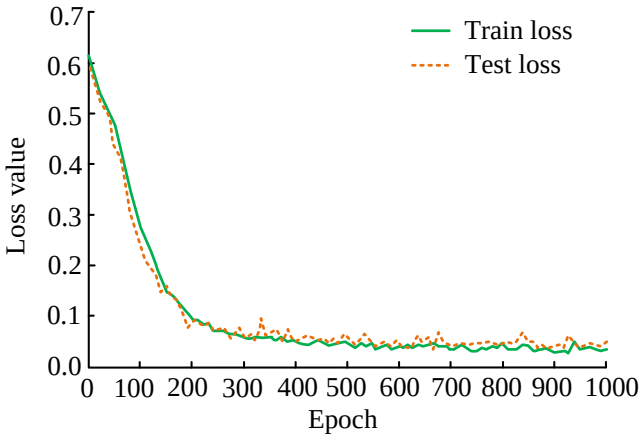
(a) Confusion matrix results for R(2+1)D-18

(b) Confusion matrix results for ASRNM

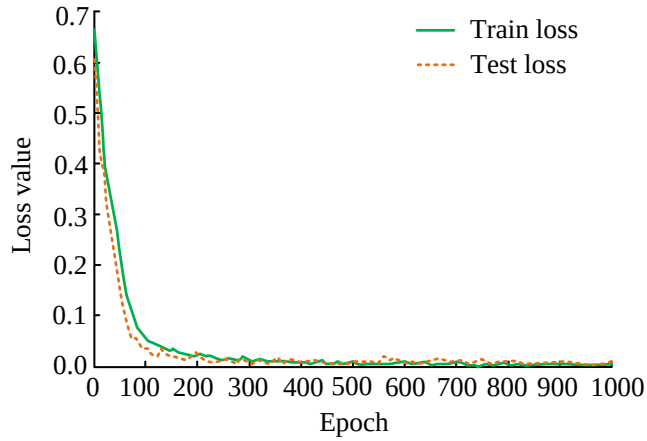
Figure 15: Matrix plot for Swim datasets

Figure 15(a) displays the confusion matrix results produced by the R(2+1)D-18 network on the Swim dataset, whereas Figure 15(b) displays the confusion matrix results produced by the ASRNM model on the same dataset. The comparison shows that the ASRNM model increases the accuracy of pose recognition for the four swimming categories by an average of 14.50% over the R(2+1)D-18 network. This suggests that the study's ASRNM model can recognize swimming poses with greater fine-grainedness. In the meantime, the study

examines the two approaches' variations in loss values on the Swim dataset in more detail. The comparative results are displayed in Figure 16.



(a) Loss curve of R(2+1)D-18 model on Swim dataset



(b) Loss curve of ASRNM model on Swim dataset

Figure 16: Comparison of loss values between the two methods in the Swim dataset

The loss curve of the R(2+1)D-18 network for 1000 iterations in the Swim dataset is shown in Figure 16(a). A comparison of the ASRNM model's loss curves in Figure 16(a), (b) reveals that, at 1000 iterations, the R(2+1)D-18 network's loss value is near to 0.03, whereas the ASRNM model's loss value is close to 0 at roughly 300 iterations. This suggests that the study's proposed ASRNM model performs better and needs fewer iterations.

4 Discussion

The development and application of artificial intelligence have led to the emergence of HPR, which involves interdisciplinary disciplines for automatic extraction of human feature poses in video images via DL technology. However, the traditional C3D network model requires a large number of parameters in the process of feature extraction, which can lead to a decline in model recognition accuracy. Therefore, this study proposes an improved C3D network model. The validation of the model's performance indicates that the accuracy and F1 value of the enhanced C3D network proposed in the study are significantly higher than those of the original C3D network. This finding was consistent with the results reported in literature [8] and [15]. Then, the accuracy and F1 value of the model varied depending on the validation dataset used. When comparing the validation results of literature [15] in the HMDB51 dataset, the accuracy of the proposed improved C3D network model was lower. This may be due to the fact that the study's recognition performance was affected to some extent by reducing the number of parameters in the network model. However, the validation resulted from the Sports-1M dataset further confirm the feasibility of the improved C3D network model for recognizing swimming sport poses.

The practical value of gesture recognition technology in sports applications is significant. In this study, an S-Pos recognition model was constructed based on the improved C3D network, combined with the SR structure of the ST channel attention mechanism. The recognition model achieved a high mAP value of 0.88 and an AUC of 0.98 in the homemade SA dataset. While previous HPR studies achieved up to 95% accuracy in some benchmark datasets, the validation of the mAP value was not analyzed further. However, when

comparing AUC, the study showed that the ASRNM model is superior for recognizing sports. Additionally, the confusion matrix validation results for the R(2+1)D-18 network on gesture recognition of swimming actions further affirmed the value of the ASRNM model in sports applications such as swimming.

However, recognizing SA poses can be limited by factors such as video quality, illumination, and the distance between the athlete and the camera. The ASRNM model construction is performed based on the improved C3D network. The model is then validated from different perspectives using the HMDB51 dataset, Sports-1M dataset, and SA dataset. The HMDB51 dataset comprises 6,849 videos with a video resolution in the range of 320*240 and includes 51 types of actions, such as general facial actions, human body actions, and general body actions. The dataset's performance validation confirms the effectiveness of the proposed model for recognizing human gestures in low-quality videos. The Sports-1M dataset consists mainly of videos from YouTube, which vary in quality, shooting background, lighting, and camera distance. The proposed improved C3D network shows an average improvement of 25% compared to the original C3D network. This suggests that the improved C3D network has potential applications in sports. The ASRNM model proposed still demonstrates superior performance in studying the homemade SA sports dataset and recognizing swimming actions.

5 Conclusion

The paper suggests a pose recognition model based on the enhanced C3D network in an effort to address the issue of the unsatisfactory recognition impact of the network under the large number of parameters. Firstly, the FCL as well as the activation function are replaced to improve the C3D network, and the improved C3D network is further extended into the FPR-C3D network, and the S-Pos recognition model is constructed by utilizing the ST channel attention focusing mechanism and the SR structure. The validation of the improved C3D network revealed that its training and testing accuracies in the HMDB51 dataset increased by 15.15% and 13.49%, respectively, compared to the traditional C3D network. The accuracy of the ASRNM model in the SA dataset increased by 24.37% and 4.31% over the two popular networks, respectively, according to a comparison of the performance of the various models. Its AUC value was as high as 0.98 and its mAP value increased by 2.33%-37.50% over several methods. The confusion matrix results for the Swim dataset revealed that the ASRNM model increased the accuracy of pose recognition for the four swimming categories by an average of 14.50% over the R(2+1)D-18 network. The aforementioned findings demonstrate that the study's improved C3D network has successfully had its parameter count reduced. Additionally, the ASRNM model, which is based on the improved C3D network, is much lighter than the traditional C3D model and performs better in terms of accuracy and fine-grainedness when it comes to sports pose recognition.

6 Limitations and future work

Experimental validation on various datasets confirms the effectiveness of the improved C3D network model and the ASRNM model for SA recognition, such as swimming. However, the study's shortcoming is the low validation accuracy and F1 value scores for large datasets, although it is still superior to the C3D network. Additionally, the ASRNM model proposed in the study could not be pre-trained on large datasets due to hardware limitations. As a result, the number of networks capable of effectively recognizing contrasts is limited. This is a crucial aspect to improve and optimize in the next step of the study. To improve the accuracy of period human pose recognition, further reducing the number of network participants and using larger computer equipment for pre-training the recognition model will be considered.

The next step of the research will be to design a visualization system for swimming sport recognition based on the ASNRM model. Conducting research on the recognition of various sports movements is not only conducive to the development of sports, but also helps to promote the intelligence, science, and rationality of physical exercise. The utilization of DL and other technologies for researching movement recognition is significant for sports. This data can be used to plan athlete training and recuperation. Additionally, implementing intelligent technology to recognize human movement postures can improve incorrect national movement postures, promoting the healthy development of national sports and achieving the ambitious goal of strengthening the national body.

References

[1] M. Estiri, J. H. Dahooie, and E. K. Zavadskas, “Providing a framework for evaluating the quality of health care services using the healthqual model and multi-attribute decision-making under imperfect knowledge of data,” Informatica, vol. 34, no. 1, pp. 85-120, 2023. https://doi.org/10.15388/23-INFOR512.

[2] H. Tang, L. Ding, S. Wu, B. Ren, N. Sebe, and P. Rota, “Deep unsupervised key frame extraction for efficient video classification,” ACM Transactions on Multimedia Computing Communications and Applications, vol. 19, no. 3, pp. 1-17, 2023. https://doi.org/10.1145/3571735.

[3] S. Salimian, S. M. Mousavi, and Z. Turskis, “Transportation mode selection for organ transplant networks by a new multi-criteria group decision model under interval-valued intuitionistic fuzzy uncertainty,” Informatica, vol. 34, no. 2, pp. 337-355, 2023.

https://doi.org/10.15388/23-INFOR513.

[4] C. Pham, L. Nguyen, A. Nguyen, N. Nguyen, and V. T. Nguyen, "Combining skeleton and accelerometer data for human fine-grained activity recognition and abnormal behaviour detection with deep temporal convolutional networks," Multimedia Tools and Applications, vol. 80, no. 19, pp. 28919-28940, 2021. https://doi.org/10.1007/s11042-021-11058-w.

[5] T. Huang, X. Ben, C. Gong, B. Zhang, R. Yan, and Q. Wu, "Enhanced spatial-temporal salience for cross-view gait recognition," T-CSVT, vol. 32, no. 10, pp. 6967-6980, 2022. https://doi.org/10.1109/TCSVT.2022.3175959.

[6] T. C. Koh, C. K. Yeo, X. Jing, and S. Sivadas, "Towards efficient video-based action recognition: context-aware memory attention network," SN Applied Sciences, vol. 5, no. 12, pp. 1-12, 2023. https://doi.org/10.1007/s42452-023-05568-5.

[7] C. Zheng, W. Wu, C. Chen, T. Yang, S. Zhu, J. Shen, N. Kehtarnavaz, and M. Shah, "Deep learning-based human pose estimation: A survey," ACM Computing Surveys, vol. 56, no. 1, pp. 1-37, 2023, https://doi.org/10.1145/3603618.

[8] P. Verma and S. Rajeev, "Two-stage multi-view deep network for 3D human pose reconstruction using images and its 2D joint heatmaps through enhanced stack-hourglass approach," The Visual Computer, vol. 38, no. 7, pp. 2417-2430, 2022. https://doi.org/10.1007/s00371-021-02120-7.

[9] J. P. Sahoo, S. P. Sahoo, S. Ari, and S. K. Patra, "RBI-2RCNN: Residual block intensity feature using a two-stage residual convolutional neural network for static hand gesture recognition," Signal Image and Video Processing, vol. 16, no. 8, pp. 2019-2027, 2022. https://doi.org/10.1007/s11760-022-02163-w.

[10] T. S. Chen, N. Yabuki, and T. Fukuda, "Motion recognition method for construction workers using selective depth inspection and optimal inertial measurement unit sensors," CivilEng, vol. 4, no. 1, pp. 204-223, 2023. https://doi.org/10.3390/civileng4010013.

[11] A. Balmik, A. Paikaray, M. Jha, and A. Nandy, "Motion recognition using deep convolutional neural network for Kinect-based NAO teleoperation," Robotica, vol. 40, no. 9, pp. 3222-3253, 2022. https://doi.org/10.1017/S0263574722000169.

[12] L. Weng, W. Lou, X. Shen, and F. Gao. "A 3D graph convolutional networks model for 2D skeleton - based human action recognition," IET Image Processing, vol. 17, no. 3, pp. 773-783, 2023. https://doi.org/10.1049/ipr2.12671.

[13] J. Yu, H. Gao, Y. Chen, D. Zhou, J. Liu, and Z. Ju, "Adaptive spatiotemporal representation learning for skeleton-based human action recognition," IEEE Transactions on Cognitive and Developmental Systems, vol. 14, no.4, pp. 1654-1665, 2021. https://doi.org/10.1109/TCDS.2021.3131253.

[14] L. Jiang, B. Zou, S. Liu, W. Yang, M. Wang, and E. Huang, "Recognition of abnormal human behavior in dual-channel convolutional 3D construction site based on deep learning," Neural Computing and Applications, vol. 35, no. 12, pp. 8733-8745, 2023. https://doi.org/10.1007/s00521-022-07881-3.

[15] A. Jisi and S. Yin, "A new feature fusion network for student behavior recognition in education," JASE, vol. 24, no. 2, pp. 133-140, 2021. https://doi.org/10.6180/jase.202104_24(2).0002.

[16] A. Jan and G. M. Khan, "Real-world malicious event recognition in CCTV recording using Quasi-3D network," Journal of Ambient Intelligence and Humanized Computing, vol. 14, no. 8, pp. 10457-10472, 2023. https://doi.org/10.1007/s12652-022-03702-6.

[17] Z. Chen, M. Liang, Z. Xue, and W. Yu, "STRAN: Student expression recognition based on spatio-temporal residual attention network in classroom teaching videos," Applied Intelligence, vol. 53, no. 21, pp. 25310-25329, 2023. https://doi.org/10.1007/s10489-023-04858-0.

[18] H. Wu, J. Luo, X. Lu, and Y. Zeng, "3D transfer learning network for classification of Alzheimer's disease with MRI," International Journal of Machine Learning and Cybernetics, vol. 13, no. 7, pp. 1997-2011, 2022. https://doi.org/10.1007/s13042-021-01501-7.

[19] X. Hong, T. Zhang, Z. Cui, and J. Yang, "Variational gridded graph convolution network for node classification," JAS, vol. 8, no. 10, pp. 1697-1708, 2021. https://doi.org/10.1109/JAS.2021.1004201.

[20] S. Kumawat, M. Verma, Y. Nakashima, and S. Raman, "Depthwise spatio-temporal STFT convolutional neural networks for human action recognition," TPAMI, vol. 44, no. 9, pp. 4839-4851, 2021. https://doi.org/10.1109/TPAMI.2021.3076522.

[21] S. Liu, Y. Ren, L. Li, X. Sun, Y. Song, and C. C. Hung, "Micro-expression recognition based on SqueezeNet and C3D," Multimedia Systems, vol. 28, no. 6, pp. 2227-2236, 2022. https://doi.org/10.1007/s00530-022-00949-z.

[22] J. Guo, Y. Liu, Q. Yang, Y. Wang, and S. Fang, "GPS-based citywide traffic congestion forecasting using CNN-RNN and C3D hybrid model," Transportmetrica A: Transport Science, vol. 17, no. 2, pp. 190-211, 2021. https://doi.org/10.1080/23249935.2020.1745927.

[23] H. Gao, Y. Liu, and S. Ji, "Topology-aware graph pooling networks," TPAMI, vol. 43, no. 12, pp. 4512-4518, 2021. https://doi.org/10.1109/TPAMI.2021.3062794.

[24] S. M. S. Abdullah and A. M. Abdulazeez, "Facial expression recognition based on deep learning convolution neural network: A review," JSCDM,

vol. 2, no. 1, pp. 53-65, 2021. https://doi.org/10.30880/jscdm.2021.02.01.006.

[25] B. Gülmez, "A novel deep neural network model based Xception and genetic algorithm for detection of COVID-19 from X-ray images," Annals of Operations Research, vol. 328, no. 1, pp. 617-641, 2023. https://doi.org/10.1007/s10479-022-05151-y.

[26] I. Jahan, M. F. Ahmed, M. O. Ali, and Y. M. Jang, "Self-gated rectified linear unit for performance improvement of deep neural networks," ICT Express, vol. 9, no. 3, pp. 320-325, 2023. https://doi.org/10.1016/j.icte.2021.12.012.

[27] Y. Xie, A. N. J. Raj, Z. Hu, S. Huang, Z. Fan, and M. Joler, "A twofold lookup table architecture for efficient approximation of activation functions," IEEE Transactions on Very Large-Scale Integration (VLSI) Systems, vol. 28, no. 12, pp. 2540-2550, 2020. https://doi.org/10.1109/TVLSI.2020.3015391.

[28] Y. Wang, D. J. Tan, N. Navab, and F. Tombari, "Softpool++: An encoder–decoder network for point cloud completion," IJCV, vol. 130, no. 5, pp. 1145-1164, 2022. https://doi.org/10.1007/s11263-022-01588-7.

[29] S. Liu, X. Wang, L. Zhao, B. Li, W. Hu, J. Yu, and Y. D. Zhang, "3DCANN: A spatio-temporal convolution attention neural network for EEG emotion recognition," IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 11, pp. 5321-5331, 2021. https://doi.org/10.1109/JBHI.2021.3083525.

[30] T. Ge and O. Darcy, "Study on the design of interactive distance multimedia teaching system based on VR technology," International Journal of Continuing Engineering Education and Life Long Learning, vol. 32, no. 1, pp. 65-77, 2022. https://doi.org/10.1504/IJCEELL.2022.121221.

[31] T. V. Henriksen, N. Tarazona, A. Frydendahl, T. Reinert, F. Gimeno-Valiente, J. A. Carbonell-Asins, and C. L. Andersen, "Circulating tumor DNA in stage III colorectal cancer, beyond minimal residual disease detection, toward assessment of adjuvant therapy efficacy and clinical behavior of recurrences," Clinical Cancer Research, vol. 28, no. 3, pp. 507-517, 2022. https://doi.org/10.1158/1078-0432.CCR-21-2404.