Journal of Information Hiding and Multimedia Signal Processing ©2025 ISSN 2073-4212
Ubiquitous International Volume **16**, Number **4**, December **2025**

Research on Student Classroom Posture Recognition Based on Deep Learning

Shi Xiangui[1,2,*], Vladimir Y. Mariano[1]

[1]College of Computing and Information Technologies,
National University Manila, Philippines
[2]School of Big Data and Artificial Intelligence,
Anhui Xinhua University, Hefei 230088, China
shixiangui@axhu.edu.cn, vymariano@national-u.edu.ph

*Corresponding author: Shi Xiangui

Received June 9, 2025, revised July 11, 2025, accepted July 21, 2025.

Abstract. *Efficient classroom behavior analysis is the key to optimize teaching strategies and improve teaching quality. Traditional manual observation methods have low efficiency and strong subjectivity. This paper proposes a fusion YOLOv5 target detection with OpenPose attitude estimation of double model framework, realize the automatic recognition and classification of students classroom attitude. The system first uses OpenPose to extract key points of human bones from classroom images, and then YOLOv5 accurately locates students in the classroom in real time. Finally, feature vectors are constructed based on the Angle and distance of key points, and a classifier is trained to recognize four typical classroom postures: "gaze ahead", "bow", "lie down", and "turn around". In the self-built real classroom scene data sets of experiments show that the system average recognition accuracy reaches 92.7%. Focus evaluation, the method of classroom teaching interaction analysis provides the objective and efficient technical support, has significant education application value.*
Keywords: Student posture recognition; Classroom behavior analysis; YOLOv5; OpenPose; Computer vision.

1. **Introduction.** Under the background of digital transformation of education, using intelligent means to understand classroom dynamics and evaluate teaching effects has become a research hotspot. The attitude of students in the classroom is an important explicit indicator of their participation, concentration and learning status. The traditional method relying on teachers' observation records has narrow coverage, strong subjectivity and low efficiency. The development of computer vision technology, especially the breakthrough in deep learning, has made it possible to automate and non-intrusive classroom behavior analysis [1, 2].

As a leading human pose estimation algorithm, OpenPose can efficiently locate key points on multiple parts of the human body. However, it faces challenges such as multi-person occlusion, complex background, and high computational overhead in panoramic classroom scenes. YOLO series object detection algorithms are famous for their excellent real-time performance and accuracy. In this paper, YOLOv5 is combined with OpenPose innovatively to build a collaborative recognition framework: OpenPose high-precision pose analysis, and YOLOv5 quickly lock the student target area. The division of labor system in complex significantly increased robustness and processing efficiency of the classroom environment. This study aims to develop an accurate, real-time, and practical student classroom posture recognition system, which can provide data support for fine teaching management and personalized learning intervention [3, 4].

Early research on classroom behavior recognition was mostly based on wearable sensors, but it can easily interfere with normal teaching activities. Computer vision based methods have gradually become mainstream, evolving from simple motion analysis and background subtraction to deep learning based

behavior recognition. The YOLO series of object detection is widely used due to its "single-stage" detection and extremely fast speed. YOLOv5 achieves a good balance between accuracy and speed, making it suitable for real-time classroom monitoring scenarios. OpenPose adopts Part Affinity Fields (PAFs) technology for pose estimation, which can effectively handle multi person pose estimation and output the coordinates of 18 key points of the human body, providing a foundation for high-level pose understanding. The existing limitations are that single OpenPose processing of panoramic classroom images requires a large amount of computation and is sensitive to occlusion and small targets, while relying solely on object detection lacks fine-grained pose information. Further exploration is needed to integrate the advantages of both in the application of classroom scenes [5, 6].

2. Key Technologies.

2.1. OpenPose algorithm.

OpenPose is an open-source human pose estimation system based on deep learning developed by Carnegie Mellon University. It can detect and estimate the key point positions of the human body, face, and hands in real-time from images or videos, and construct skeleton models. Its core technical principle is to use dual stream convolutional neural networks combined with Part Affinity Fields (PAFs) technology to achieve multi person pose estimation. The workflow is divided into three stages:

(1) **Feature extraction:** using pre trained convolutional neural networks (such as VGG19) to extract multi-level features of the image;
(2) **Key Point and Association Field Prediction:** Generate Part Confidence Maps (PCM) to locate key points in body parts, and describe the connection relationships between key points through Part Association Fields (PAFs);
(3) **Multi stage optimization:** Improve the accuracy of keypoint detection and connection through iterative optimization [7].

The network structure is shown in Figure 1.

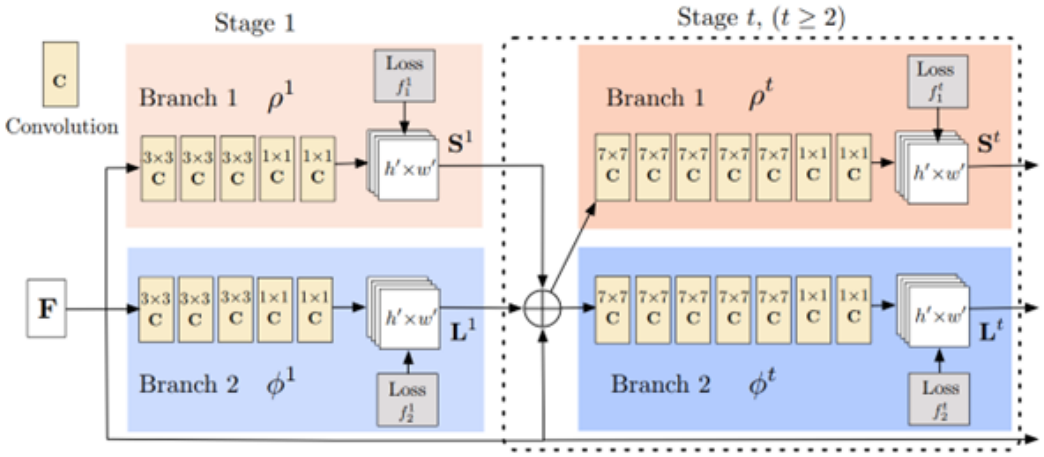

FIGURE 1. OpenPose network structure diagram

OpenPose support is unique in many scenes gesture detection, without prior segmentation individuals, able to identify each person's bone structure in the image, suitable for complex scenarios in the human body posture estimation tasks. The OpenPose system can detect 18 key points of the human body, including the position of the head, shoulder, elbow, knee, and foot. In addition, it can recognize 21 key points of the hand and 70 key points of the face [8].

OpenPose advantage is that it can achieve real-time processing in complex attitude estimation task, the wide range of applications, including motion capture, human behavior analysis, gesture recognition, health monitoring, virtual reality (VR) and augmented reality (AR) and human-computer interaction, etc.

Classroom scenarios and overlapping phenomenon more serious, there are contain Numbers for Openpose algorithm for estimating people pose effect is better, so in this article, the method is chosen as the classroom students attitude skeleton extraction of data. The coordinates of human body are obtained by this method the key information are shown in Table 1 below, a total of 18 key point, use a scale of 0-17 number respectively [9]. The positions of key points in the human skeleton are shown in Figure 2.

TABLE 1. OpenPose human skeleton keypoint list

0: Nose	6: Left elbow	12: Left knee
1: Neck	7: Left wrist	13: Left ankle
2: Right shoulder	8: Right crotch	14: Right eye
3: Right elbow	9: Right knee	15: Left eye
4: Right wrist	10: Right ankle	16: Right ear
5: Left shoulder	11: Left crotch	17: Left ear

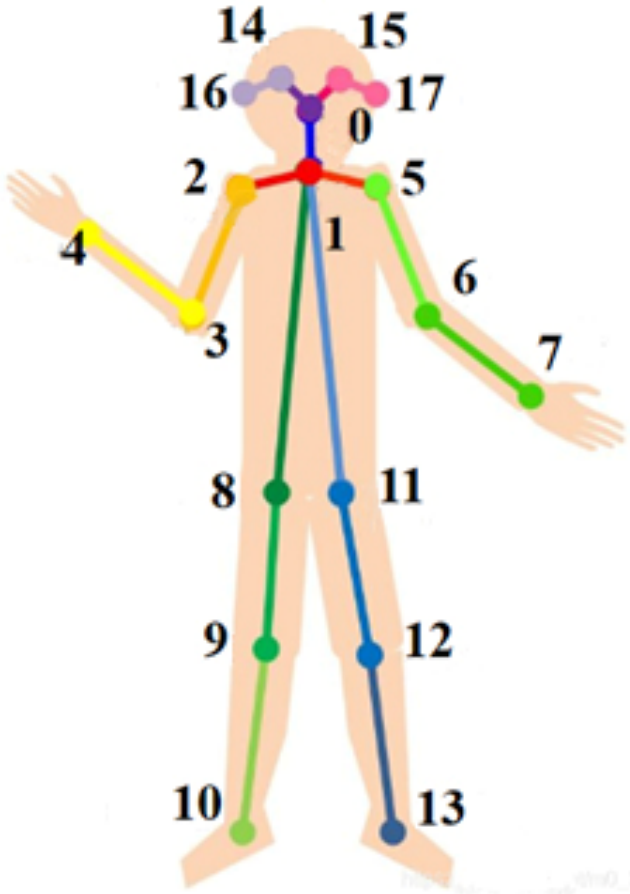

FIGURE 2. OpenPose human keypoint location map

2.2. **YOLO algorithm.** YOLO (You Only Look Once) is an advanced deep learning-based real-time object detection algorithm first proposed by Joseph Redmon et al. in 2015. YOLO is mainly used in the field of object detection and the subfield of machine vision. It can complete the recognition and location of objects in the image by a single view. It has the advantages of fast speed, high accuracy, strong interpretability and wide applicability, and is one of the most important representatives in the field of object detection [10]. Its core feature is to predict the location and class of multiple objects in the image simultaneously through a single forward propagation, regard object detection as a single regression task, and directly output the bounding box coordinates (x, y, w, h) and class probabilities in the convolutional neural network [11].

2.2.1. *The core design idea of YOLO.* The core design ideas of YOLO are mainly reflected in the following aspects, This has led to breakthroughs in real-time performance and accuracy in the field of object detection.

(1) One-Stage detection:: YOLO abandons the traditional multi-stage detection process, such as the R-CNN series, and reconstructs object detection as an end-to-end regression problem. It directly predicts the bounding box coordinates (position and size) and their category probabilities of objects in a single neural network, and directly outputs the bounding box coordinates (x, y, w, h) and category probabilities of the target through a single forward propagation, achieving end-to-end object detection and significantly improving detection speed.

(2) Meshing and multi-scale prediction:: Grid prediction refers to dividing the input image into $S \times S$ grids (such as 7×7), where each grid is responsible for predicting the target whose center point falls within its region, and directly regressing the bounding box coordinates. Multi scale fusion refers to the introduction of feature pyramids (FPN/BiFPN) for prediction on feature maps of different scales, balancing the detection accuracy of small and large targets.

(3) Global context modeling:: YOLO difference from the traditional local area analysis method (such as Faster R-CNN), it will be the whole image input network, the reasoning with full information, each prediction unit (grids), boundary box when it is output will be integrated whole figure rather than the local characteristics, such treatment can effectively reduce the background, and improve the understanding of context [12].

2.2.2. *YOLOv5.* YOLOv5 consists of three parts: Backbone, Neck, and Head, which support multi-scale object detection [13]. The overall network architecture of YOLOv5 algorithm is shown in Figure 3. The core architecture of YOLOv5 includes:

- **(1) Backbone:** A backbone network improved based on CSPNet, using efficient convolutional neural networks as feature extractors, typically based on Cross Stage Partial Network (CSP), i.e. CSP structure design, such as CSPDarknet53, CSPDarknet53x or CSPDarknet53s, reduces computation and parameter count while maintaining strong feature extraction capabilities.
- **(2) Neck:** Usually, Spatial Pyramid Pooling (SPP) module or Path Aggregation Network (PANet) structure is used to further enhance the spatial information of feature maps.
- **(3) Head:** Use prediction heads to output bounding boxes, object confidence, and category probabilities.

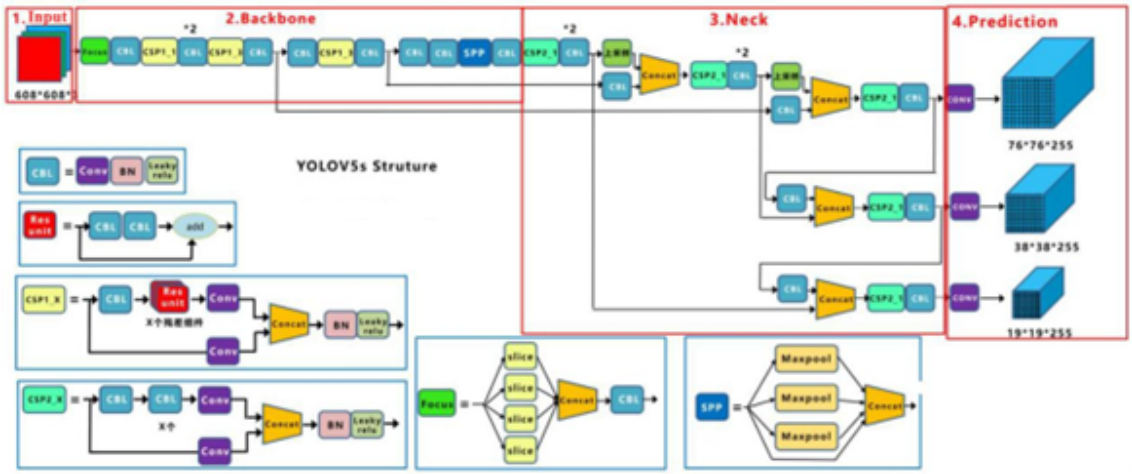

FIGURE 3. Overall structure diagram of YOLOv5 network model

3. Classroom posture recognition method based on OpenPose-YOLO.

3.1. System process design.

(1) **Input:** install high-definition cameras in the classroom, collect teaching video data during class, and extract and process these video frames to generate classroom teaching images.

(2) **Student skeleton data generation:** the cropped classroom teaching images are input into the OpenPose model, and the pixel coordinates of the key points of the student body in the classroom environment are extracted.

(3) **Feature extraction:** calculate the pose features based on the coordinates of key points.

(4) **Student pose classification detection:** The pre-trained YOLOv5s model is used to detect all students in the image and output the Bounding Box. The extracted feature vectors are input into the pre-trained classifier to output the pose category labels (look, down, lie, turn) [14, 15].

The processing flow of this system is shown in Figure 4 below.

FIGURE 4. Flow chart of the system experiment

3.2. **Model selection and implementation.**

(1) **OpenPose:** uses the OpenPose model based on COCO dataset for training weights, and applies it to the classroom teaching images to extract the key point data of students' human bones in the classroom environment.

(2) **YOLOv5:** The lightweight YOLOv5s version was selected to achieve a balance between accuracy and speed. The weights pre-trained on the COCO dataset were used, and the classroom scene data were used for fine-tuning to optimize the detection performance of student targets.

3.3. **Dataset construction.**

(1) **Data source:** Axis series high-definition cameras were used to collect classroom teaching videos of different periods in real classrooms. The video frames are extracted at a certain frame rate, and the sorted high-resolution pictures are used as the basic image data set of this paper. The filtered data set has the advantages of rich data categories, clear data, and obvious student postures.

(2) **Data labeling:** LabelImg tool was used to manually label the pose categories of students, and the key point coordinates output by OpenPose were recorded for feature calculation.

(3) **Scale and the division:** the video according to the frame rate and in time to catch the segmentation, a total of 6300 classroom teaching image segmented, manual pictures after screening, finally determined the 2880 images containing a higher proportion of students attitude information as the research of the original data set.

4. Experimental results and analysis.

4.1. **Training and testing data.** Since there is not a public dataset of student classroom posture, this study uses a high-definition camera installed in the classroom to capture classroom teaching videos and extract video frames from them. The dataset includes classroom teaching images in different classes and under different lighting conditions, which is beneficial to enhance the robustness of the network.

The preparation work of the dataset mainly includes three parts: data collection and screening, image preprocessing, and annotation. Data collection mainly involves collecting classroom teaching image data, using cameras installed in the classroom to capture and obtain video data, and then extracting video frames at a certain frame rate to obtain a large number of classroom teaching images. Afterwards, the selected images are filtered to remove unqualified images such as excessive blur and occlusion. Image preprocessing mainly involves filtering, adjusting, enhancing, and annotating raw images to meet the input requirements of the network. In order to better extract image features and prevent overfitting of the model, data augmentation was performed on the collected image samples to expand the scope of the dataset and enhance the model's generalization ability. The main methods used for data augmentation include horizontal mirroring, vertical flipping, vertical mirroring, and other transformation methods. In addition, the division of the training set, testing set, and validation set should strive to maintain the consistency of the data distribution as much as possible, in order to avoid introducing additional biases during the data division process that may affect the final results [16, 17].

OpenPose is used to extract the features of students' human bones from classroom teaching images, and the postures of students' bones are mainly divided into four categories: look, down, lie, and turn. The classification table of students' classroom postures is shown in Table 2 below. After the above data processing, 2880 data pictures were collected and organized, and LabelImg open source data labeling tool

was used to label the student's human posture. The four types of posture images are shown in Figure 5 below.

TABLE 2. Categories of student classroom postures

Posture	**State name**	**Criteria for determination**
look	Look ahead	The key points of the head should be basically aligned with the central axis of the torso, facing forward
down	Head down	The head key point is significantly lower than the central axis of the torso, and the head is oriented toward the table top
lie	Lie on your stomach	Chest, shoulders, and head keypoints are almost parallel to the desktop
turn	Turn around	The key point of the head or the central axis of the torso is significantly offset from the direction of the blackboard

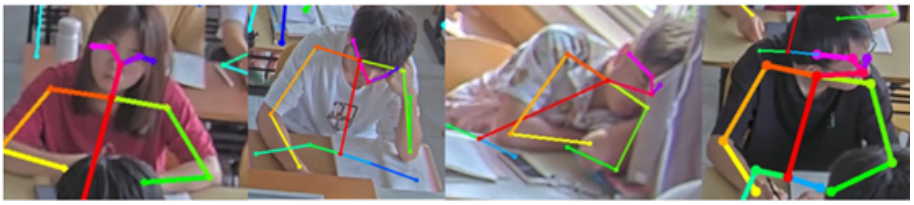

FIGURE 5. Images of four classes of classroom postures

Divide the dataset into three parts: training set, testing set, and validation set, and divide them into training set, validation set, and testing set in a 6:2:2 ratio. Ensure that the distribution of samples in different posture categories is relatively balanced. Using the 'leave out method', the dataset D is directly divided into three mutually exclusive sets, one of which serves as the training set S, one as the testing set T, and the other as the validation set V, i.e. $D = S \cup T \cup V$, $S \cap T \cap V = \emptyset$. Train the network parameters through the training set, and then pass the model error detection to the validation set to further determine the parameters. Repeat the training until the error on the validation set reaches the minimum [18].

4.2. **Evaluation index. Recall:** The recall rate refers to the proportion of positive samples correctly predicted by the classifier to all positive samples, and its calculation formula is as follows:

$$Recall = \frac{TP}{TP + FN} \tag{1}$$

Among them, TP is the number of positive samples correctly identified, and FN is the number of positive samples incorrectly identified as negative samples.

Precision: the ratio of TP in the recognized picture is calculated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Where FP (false positives) is the number of negative samples that are incorrectly identified as positive samples.

The relationship between precision and recall in this experiment is shown in Table 3 below.

The Loss functions used by YOLOv5s mainly include Bounding Box Regression Loss, Objectness Loss, Classification Loss, and so on. These loss functions work together to optimize the performance of the

TABLE 3. Evaluation metrics for this experiment

Attitude category	Precision	Recall
Sit and listen	0.94	0.96
Bow your head	0.91	0.88
Lie on your stomach	0.89	0.92
Turn around	0.92	0.91

model in detection tasks. Bounding box regression loss further considers the distance and shape difference between bounding boxes based on CIoU [19]. It is calculated as follows:

$$CIoU = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha \cdot v \tag{3}$$

Among them, IoU is the intersection to union ratio, b and b^{gt} represent the central points of the predicted box and the ground truth box respectively, ρ is the Euclidean distance, c is the diagonal length of the smallest enclosing box, and $\alpha \cdot v$ is the aspect ratio penalty term. The binary cross entropy loss is used to calculate the target confidence loss, and its calculation formula is as follows:

$$BCE(p, p^{(gt)}) = -[p^{gt} \cdot \log(p) + (1 - p^{gt}) \cdot \log(1 - p)] \tag{4}$$

Among them, p is the predicted probability (confidence level) and p^{gt} is the ground truth label (1 means there is a target, 0 means there is no target). The calculation formula for binary cross entropy loss is the same as the formula for target confidence loss mentioned above.

4.3. **Model training and detection results.** The model training hardware environment was NVIDIA Geforce RTX 4090 GPU, Intel(R) Xeon(R) Gold 6132 CPU, 128GB RAM. The development language is Python, and PyTorch is used as the deep learning framework. After about 400 rounds of training, the loss function of YOLOv5s network model tends to be stable [20].

On the test set, the average accuracy of YOLOv5s model after fine-tuning reaches 94.8%, which is significantly better than that of directly using the pre-trained model (89.5%), proving the importance of scene adaptation. On the independent validation set, the average accuracy of pose classification is 92.7%.

The bounding box regression Loss, Objectness Loss and Classification Loss used in the training process converge gradually with the progress of training, which indicates that the model is gradually optimized and learns better feature representations. The decreasing trend of these losses also reflects the continuous improvement of the prediction accuracy of the model, and there is no obvious overfitting phenomenon, indicating that the setting of hyperparameters and the early stopping mechanism have played a good effect. The detection effect is shown in Figure 6 below.

5. **Conclusion.** This paper successfully constructed and verified a student classroom posture recognition system based on YOLOv5 and OpenPose collaborative framework. The system accurately extracts the key points of the human body through OpenPose, and then efficiently locates the student target area by YOLOv5. Finally, the posture features are classified. On the real classroom dataset, the system achieves an average posture recognition accuracy of 92.7%, which proves its effectiveness, real-time performance and practicability in complex classroom environments.

This study provides a reliable technical solution for objective and automatic analysis of classroom student behavior, which has important educational application value. Future work will focus on improving the robustness of the model under extreme occlusion and perspective, exploring more refined posture definition and high-level behavior understanding, and deepening its practical application in scenarios such as smart classroom and personalized teaching.

Acknowledgment. This study was partially funded by the First Class Professional Construction Project of Anhui Xinhua University (No. 2020ylzyx03), the Quality Engineering Project of Anhui Xinhua University (No. 2024zygzj04), the National Innovation Training Project (No. 20241221632), and the Anhui Provincial Innovation Training Project (No. 202312216021/202312216022).

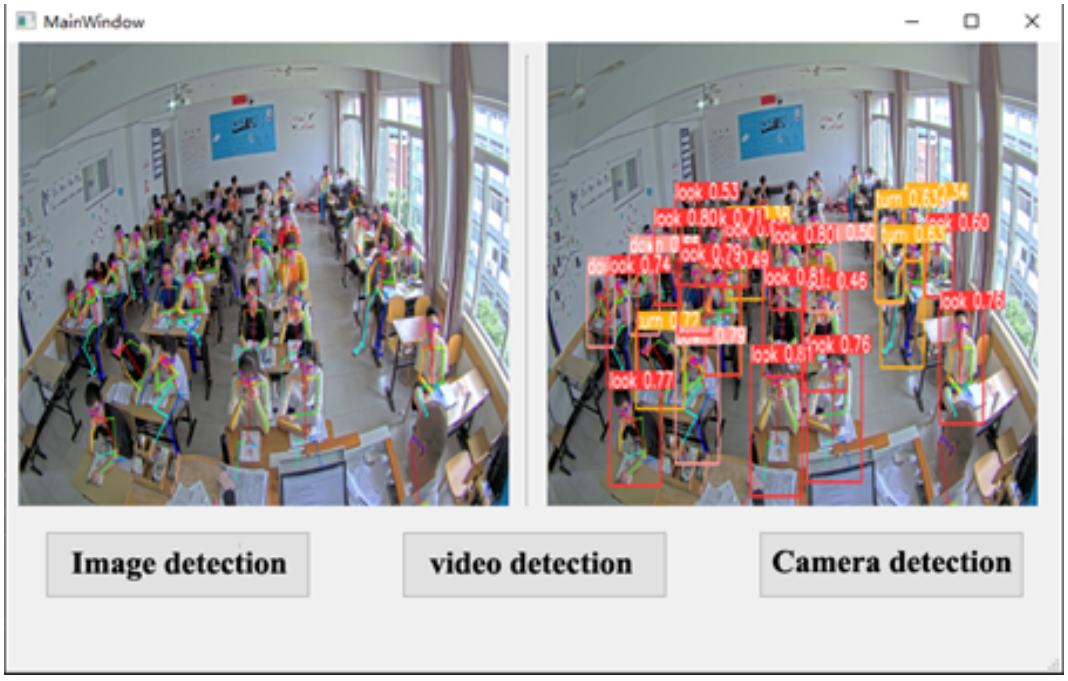

FIGURE 6. Test effect diagram

REFERENCES

[1] Liu Hongda, Sun Xuhui, and Li Yibin, "A Review of Deep Learning Models for Image Classification Based on Convolutional Neural Networks," *Computer Engineering and Applications*, vol. 61, no. 11, pp. 1-21, 2020.

[2] Ji Changqing, Gao Zhiyong, and Qin Jing, "A Review of Image Classification Algorithms Based on Convolutional Neural Network," *Journal of Computer Applications*, vol. 42, no. 04, pp. 1044-1049, 2022.

[3] Tian Yu and Liu Hong, "Pedestrian joint Detection Algorithm Based on Improved OpenPose," *Sensors and Microsystems*, vol. 43, no. 09, pp. 144-148, 2024.

[4] Zhang Haoyue, Li Ziyun, and Xu Taoming, "Research on the Design of an Intelligent Desk for Sitting Posture Reminder Based on OpenPose," *Science & Technology Innovation*, no. 13, pp. 59-61+64, 2024.

[5] Song Wei, Wang Heying, Guo Zhongfeng, et al., "Object Detection Algorithm Based on Improved YOLOv5s," *Mechanical Engineer*, no. 09, pp. 64-67, 2024.

[6] Liu Cuiwei, "Application Research on Using OpenPose Bone Tracking Technique to Assist Aerobics Teaching," Guangzhou University, 2023.

[7] N. Tasnim, M. K. Islam, and J.-H. Baek, "Deep learning based human activity recognition using spatio-temporal image formation of skeleton joints," *Applied Sciences*, vol. 11, no. 6, p. 2675, 2021.

[8] Li Boyu, "Human Key Point Extraction and Parameter Measurement Based on OpenPose," *Digital Technology and Application*, vol. 41, no. 06, pp. 13-16+22, 2023.

[9] Guo Yuan, Guo Chenxu, and Shi Xin, "Research on Learning Sitting Posture Monitoring System Based on OpenPose," *Electronic Design Engineering*, vol. 32, no. 18, pp. 37-41, 2024.

[10] Song Lihang, Zhang Yi, and Shi Yinhao, "Detection Method of Nectarine maturity in Natural Environment Based on Improved YOLOv5s," *Chinese Journal of Agricultural Mechanization*, vol. 45, no. 09, pp. 250-257+2, 2024.

[11] Tan Suyan, Wang Zuxuan, and He Gao Da, "Real-time Panoramic Multi-scale Classroom Behavior Recognition Based on CA-YOLOv9 Network," *Modern Educational Technology*, vol. 34, no. 07, pp. 123-130, 2024.

[12] Hu Haitao, "Research on Student Behavior Intelligence Recognition and Analysis Based on Classroom Video," Qilu University of Technology, 2024.

[13] P. Yu, X. Wang, J. Liu, et al., "Bridge target detection in remote sensing image based on improved YOLOv4 algorithm," in *2020 4th International Conference on Computer Science and Artificial Intelligence*, pp. 139-145, 2020.

[14] Yang Nan, "Research on Distracted Driving Behavior Recognition Based on Improved YoloV5 and Openpose," University of Science and Technology Liaoning, 2023.
[15] Li Zhangwei, Hu Anshun, and Wang Xiaofei, "A Review of Vision-based Object Detection Methods," *Computer Engineering and Applications*, vol. 56, no. 08, pp. 1-9, 2020.
[16] Y. Chen, L. Wang, C. Li, Y. Hou, and W. Li, "Convnets-based action recognition from skeleton motion maps," *Multimedia Tools and Applications*, vol. 79, pp. 1707-1725, 2020.
[17] H. V. Koay, J. H. Chuah, C.-O. Chow, Y.-L. Chang, and B. Rudrusamy, "Optimally-weighted image pose approach (owipa) for distracted driver detection and classification," *Sensors*, vol. 21, no. 14, p. 4837, 2021.
[18] B. Sheng, F. Xiao, L. Sha, et al., "Deep spatial–temporal model based cross-scene action recognition using commodity WiFi," *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 3592-3601, 2020.
[19] Xie Weining, "Steel Surface Defect Detection Based on Improved YOLOv8 Model," Zhejiang University of Science and Technology, 2024.
[20] W. Peng, X. Hong, H. Chen, et al., "Learning graph convolutional network for skeleton-based human action recognition by neural searching," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 03, pp. 2669-2676, 2020.