Journal of Computer Science and Electrical Engineering
ISSN: 2663-1946
DOI: https://doi.org/10.61784/jcsee3070

ENHANCING THE YOLOv11 MODEL FOR TEACHING BEHAVIOR RECOGNITION

Yao Tian, Cheng Peng*
School of Computer Science and Technology, Xinjiang Normal University, Urumqi 830054, Xinjiang, China.
Corresponding Author: Cheng Peng, Email: pcxjnu@163.com

Abstract: Traditional methods of teaching behavior recording suffer from inefficiency, long data mining times, and large computational workloads for statistical analysis. Large models and artificial intelligence offer new technical solutions that can significantly improve teaching quality and optimize the teaching process. This study, based on the improved YOLOv11 model, presents a fine-grained teaching behavior recognition technology aimed at addressing the challenges in smart classroom environments. In response to the complexity of classroom environments and the high similarity of teaching behaviors, an improved YOLOv11 algorithm is proposed. The algorithm introduces the MSCB (Multi-Scale Context Block) and SCSA (Spatial-Channel Self-Attention) modules to enhance the robustness of the model's recognition capabilities. Experimental results show that the improved model performs better in teacher behavior detection, with higher accuracy and efficiency, offering a new approach to teaching behavior recognition.
Keywords: Teaching behaviors; Object detection; Fine-grained recognition; YOLOv11; Intelligent teaching

1 INTRODUCTION

Teaching behavior is a crucial means by which teachers convey instructional information and organize classroom activities. By demonstrating various teaching behaviors, teachers can capture students' attention, enhance the effectiveness of their verbal communication, stimulate students' motivation to learn, support their understanding of the content, and improve overall classroom teaching effectiveness[1]. According to constructivist learning theory[2], knowledge is constructed through the interaction between individuals and their environment. Identifying teaching behaviors allows researchers to explore how teachers promote active learning and knowledge construction through their actions. However, both pre-service and in-service teachers often exhibit problematic teaching behaviors, whether in teaching competitions or regular classroom settings[3]. To address this issue, scholars have employed various methods to collect and analyze different types of teaching behaviors.

In recent years, research on teaching behavior recognition has generally fallen into two categories. The first category involves behavior recognition methods based on video image data. Zhao Gang[4] and others proposed a teacher set recognition and extraction algorithm, introducing a behavior recognition network based on three-dimensional bilinear pooling capable of identifying eight types of teaching behaviors. Guo Junqi[5] designed a 3D convolutional neural network tailored to classroom scenarios for recognizing teaching behaviors, achieving high recognition efficiency on a self-constructed dataset. Ding Ning[6], building upon existing coding systems, constructed a high-quality image dataset of teacher body movements and used the VGG16 network model for image recognition.

The second category comprises methods based on teacher skeletal information. Wang Tao[7], grounded in the cognition of body movement characteristics, proposed a teaching activity analysis model based on these features. Zheng Yuhuang[8] introduced a teaching behavior evaluation method based on posture recognition, obtaining teacher pose information through the HRNet deep learning network. Pang Shiyan[9] and others used the human pose estimation algorithm OpenPose to extract coordinate information. However, methods based on skeletal data are easily affected by the way key points are extracted, and the quality of the data obtained can significantly impact the accuracy of final behavior classification.

Although there have been technological breakthroughs in automated classroom teaching evaluation within smart classroom environments, research on automated recognition of teaching behaviors still requires continuous improvement and optimization[10]. This line of research not only advances theoretical development but also promotes teachers' professional growth[11]. Based on this, this study adopts a video image data-based approach and proposes an efficient and accurate method for recognizing teaching behaviors using an improved YOLOv11 model. The aim is to provide a new technical tool for analyzing teaching behaviors, thereby supporting teachers' professional development and enhancing teaching quality.

2 CONSTRUCTION OF TEACHING BEHAVIOR DATASET

2.1 Necessity of Constructing a Teaching Behavior Dataset

Although deep learning technology has demonstrated capabilities surpassing traditional algorithms in most fields, its success still fundamentally relies on the availability of sufficient data. In the field of teaching behavior recognition, current research mainly focuses on static single-frame images, making the acquisition of a large volume of high-quality image data a critical prerequisite. While there are some publicly available action recognition datasets, such as

Kinetics[12], HMDB51[13], and UCF101[14], there remains a lack of large-scale public datasets specifically targeted at teaching behavior recognition. In response to this issue, this study constructs a teaching behavior dataset based on two scenarios: teaching competitions and real classroom environments.

2.2 Classification of Teaching Behaviors

To conduct in-depth research on teaching behaviors, many scholars have started building their own datasets. Pang Shiyan[15] and others, from the perspective of teachers' non-verbal behaviors, constructed a dataset that includes adaptive behaviors, directive behaviors, intentional behaviors, instrumental behaviors, explanatory behaviors, and evaluative behaviors. Liu Qingtang[16] and colleagues built a dataset using the S-T analysis method, but only categorized behaviors into the two coarse-grained categories of teacher behaviors and student behaviors. These classifications lack sufficient granularity and provide limited guidance for evaluation. Therefore, drawing on the S-T[17] and TBAS[18] teaching behavior analysis methods, as well as the aforementioned studies, this paper selects a set of teaching behaviors for study that are visually distinguishable, pedagogically meaningful, and frequently observed in classroom instruction. The selected categories, as listed in Table 1, include "writing," "teaching," "point the board," "show things," "gesture," and "guide students," comprising a total of 1,871 images.

Table 1 Self-Constructed Teaching Behavior Dataset

Behavior Category	Action Classification	Action Description	Quantity
Writing	Writing on the blackboard	The teacher emphasizes key teaching points by writing on the blackboard	331
Teaching	Lecturing while facing students	The teacher explains content verbally without using any teaching aids	303
Point the Board.	Pointing to the board while explaining	The teacher explains key content by pointing to written material on the blackboard	291
Show Things	Holding and explaining with props	The teacher uses physical objects to visually aid understanding of the content	345
Gesture	Gesturing with both hands while explaining	The teacher uses body language and speech to convey instructional content	287
Guide Students	Interacting with students	The teacher asks students questions or comments on their responses	314

2.3 Dataset Construction Based on Teaching Competitions and Real Classrooms

The construction of the dataset mainly involves four steps: data collection, data filtering and cleaning, data preprocessing, and manual annotation. To enhance the diversity of the dataset, classroom videos were sourced from two main categories: one part consists of real classroom recordings taken in physical classrooms, and the other includes publicly available teaching competition videos found online. In the real classroom recordings, cameras were positioned either at the back of the classroom or mounted on the ceiling in the middle of the room, facing the front of the classroom to ensure the teacher was captured in the frame. In contrast, the teaching competition videos are entirely teacher-centered, with only the teacher appearing in the frame, along with elements such as the podium, teacher's desk, blackboard, and multimedia screen.

During the data preprocessing stage, video frames were extracted at regular intervals—one frame every 30 seconds—resulting in a total of 13,000 images. These images were then filtered to remove blurry, duplicate, or otherwise unusable ones. To ensure a balanced number of images for each category of teaching behavior, additional images that met the criteria were selected from the publicly available SCB-Dataset3[19] to supplement the dataset. Ultimately, 1,871 high-quality images were selected. Sample images after filtering are shown in Figure 1. Finally, these images were manually annotated using the LabelImg software.

For the experiment, the dataset was randomly divided into a training set and a test set at a ratio of 8:2. To ensure stability during the training process, data augmentation techniques were applied to the training set. These techniques included mixed enhancement methods such as random adjustments of color (brightness, contrast, saturation), the addition of Gaussian noise, and the application of Gaussian blur. The augmented images are also shown in Figure 1.

Figure 1 Dataset Samples and Augmented Images

3 IMPROVED YOLOv11 ALGORITHM IMPLEMENTATION

3.1 Overview of the YOLOv11 Model

YOLOv11 is the latest object detection model released by the Ultralytics team in 2024[20]. The architecture of YOLOv11 has undergone multiple optimizations aimed at enhancing feature extraction performance and overall efficiency. First, the model replaces the original C2f module with the C3k2 module, significantly improving module adaptability. Then, a new C2PSA module is added after the SPPF module. This module integrates an extended C2f structure with the PSA attention mechanism, effectively enhancing the extraction of key features. In the Neck part, the Concat module fuses multi-level feature maps along the channel dimension, enabling multi-scale feature fusion, which enriches feature representation and improves detection accuracy. These enhancements allow YOLOv11 to maintain high detection efficiency while significantly reducing the number of model parameters and computational load.

3.2 Algorithm Improvement

In the study of teaching behavior, issues such as large variations in object scale and strong background interference can lead to unstable detection performance when using YOLOv11 directly. At the same time, adopting a larger model would significantly increase computational load and parameter cost. Therefore, to improve both the accuracy and efficiency of teaching behavior detection while maintaining a lightweight structure, this study proposes an improved model architecture based on YOLOv11. The structure of the improved model is shown in Figure 2. Each component of the enhanced network is explained in detail in the following two subsections.

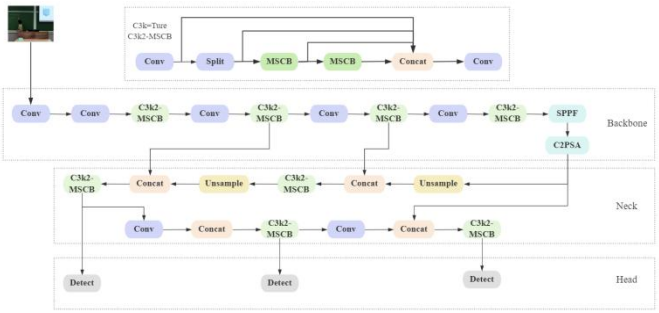

Figure 2 Improved YOLOv11 Model Diagram

3.2.1 MSCB module

In teaching behavior detection, it is necessary to process complex, dynamic, and multimodal behavioral data. To enhance the model's capability in recognizing these diverse behaviors and better adapt to the variability of teaching environments as well as the dynamic nature of behaviors, we introduce the Multi-Scale Convolution Block (MSCB), whose structure is shown in Figure 3.

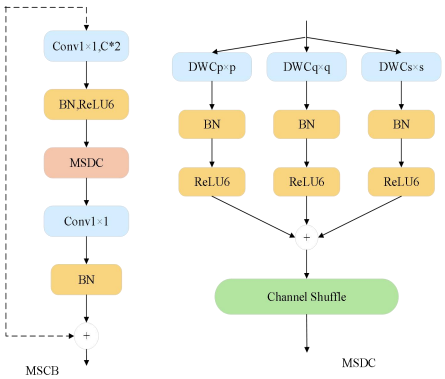

Figure 3 MSCB and MSDC Model Structures

The MSCB module integrates depthwise separable convolution with multi-scale convolution. Its core mechanism, the Multi-scale Depth-wise Convolution (MSDC), is an improved convolutional neural network architecture designed to enhance multi-scale feature extraction capability in convolution operations[21].

The core idea is to perform convolutional operations at multiple scales to capture image features at different hierarchical levels while maintaining the computational efficiency of Depth-wise Convolution. In MSCB, we follow the Inverted Residual Block (IRB) design of MobileNetV2 [22], performing depth-wise convolution at multiple scales and using channel shuffle [23] to shuffle channels between groups.Specifically:First expand the number of channels using a pointwise (1×1) convolutional layer PWC1(·).Followed by batch normalization layer BN(·).Then ReLU6 activation layer R6(·) [24]Next apply multi-scale depth-wise convolution MSDC(·) to capture multi-scale and multi-resolution context.Since depth-wise convolution ignores relationships between channels, channel shuffle operation is used to

integrate inter-channel relationships.Then:Apply another pointwise convolution PWC2(·).Followed by BN(·) to transform back to the original channels.This also encodes correlations between channels.

3.2.2 SCSA module

The SCSA module is a novel co-attention mechanism proposed by combining spatial attention and channel attention [25]. The design of SCSA consists of two main components: Shared Multi-semantic Spatial Attention (SMSA) and Progressive Channel-wise Self-Attention (PCSA). The SCSA mechanism aims to effectively integrate the advantages of both channel and spatial attention while fully utilizing multi-semantic information to enhance performance in visual tasks [26].

As shown in Figure 4, SMSA and PCSA are used in series to achieve spatial-channel co-attention based on dimension decoupling, lightweight multi-semantic guidance, and semantic discrepancy mitigation[27]. The SMSA extracts multi-level spatial information through multi-scale depth-shared 1D convolutions, providing multi-semantic spatial priors for channel attention, which helps enhance the representation of different semantic information[28].

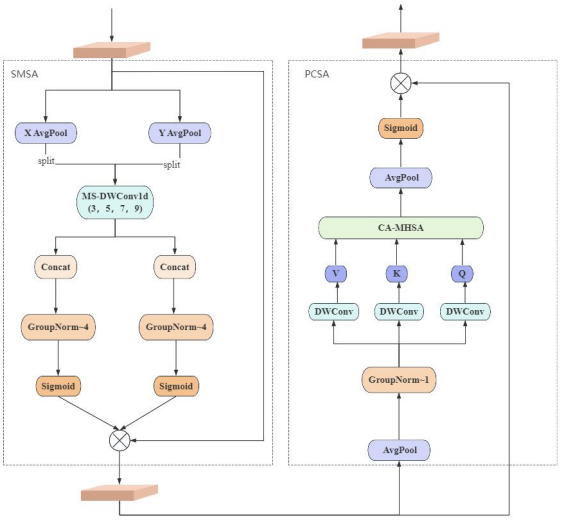

Figure 4 SMSA and PCSA Model Structures

Using a progressive compression strategy, discriminative spatial information is injected into PCSA to effectively guide channel re-calibration. This compression strategy reduces computational complexity while preserving critical spatial structural information, enabling channel attention to leverage more spatial priors during computation[29]. The PCSA module employs an input-aware self-attention mechanism that effectively computes inter-channel similarity, thereby alleviating semantic discrepancies among different sub-features within SMSA[30].

4 EXPERIMENTS

4.1 Experimental Environment and Parameters

Experiments were conducted on an Ubuntu 22.04 system using an NVIDIA RTX 3090 GPU with 24GB of memory. The model training involved 200 epochs with a batch size of 16, and the input image size was set to 640x640.

4.2 Evaluation Metrics

The following metrics are used to evaluate the performance of the model. P represents Precision, and R represents Recall, calculated as follows:

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

mAP50 refers to the mean Average Precision when the IoU threshold is 0.5. The mAP value can be calculated using the following formulas:

$$mAP = \frac{1}{C}\sum_{i=1}^{e} AP_i$$

$$AP = \int_0^1 (Pr\,(Re))d(Re)$$

mAP50-95 indicates the average of mAP values calculated at IoU thresholds ranging from 0.5 to 0.95 in increments of 0.05.

Parameter count measures the scale and complexity of the model and is calculated by summing the number of weight parameters across all layers.
GFLOPs (Giga Floating Point Operations per Second) are used to evaluate the computational complexity and runtime efficiency of the model.

4.3 Ablation Study

To validate the effectiveness of each improvement strategy, we conducted systematic ablation experiments based on YOLOv11. The experimental results are presented in Table 2.

Table 2 Ablation Study

Models	Params/M	GFlOPs	P/%	R/%	mAP50/%	mAP50-95/%
YOLOv11	2.6	6.3	84	82	86.6	50
YOLOv11＋MSCB	2.4	6.2	85	78	86.9	51
YOLOv11＋SCSA	2.5	6.3	83	81	87	50
YOLOv11＋MSCB＋SCSA	2.3	6.2	86	80	88	50

4.4 Comparative Experiments

To verify the effectiveness of our proposed improved network for teaching behavior detection, we performed comparative experiments with other classic deep learning detection networks on our self-built dataset. All comparative experiments employed identical training hyperparameter settings. The results are shown in Table 3.
As demonstrated in Table 3, our proposed improved model outperforms other detection networks in teaching behavior detection performance. Considering both average precision and processing speed for teaching behavior detection, our method exhibits superior overall performance compared to other detection networks, making it more suitable for practical deployment in teaching behavior detection applications.

Table 3 Comparative Experiment of Algorithms

Models	Params/M	GFlOPs	P/%	R/%	mAP50/%	mAP50-95/%
YOLOv8	2.7	6.8	79	80	85	50
YOLOv9	6.2	22.1	81	80	85	49
YOLOv10	2.7	8.2	84	77	83	47
YOLOv11	2.6	6.3	84	82	86.6	50
Ours	2.3	6.2	86	80	88	50

5 CONCLUSION AND OUTLOOK

This study successfully improved the YOLOv11 model for efficient and accurate teaching behavior recognition. The introduction of the MSCB and SCSA modules enhanced the model's ability to handle complex, dynamic, and multi-modal behavior data. The experimental results validate the effectiveness of the proposed model in teaching behavior detection, showing improved precision while maintaining model efficiency, making it suitable for practical classroom deployment. Future developments in this field can expect more accurate and intelligent behavior recognition systems, leveraging advancements in deep learning and computational power.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Xu Lan, Deng Yingfeng. Research on the Path of Empowering High-Quality Development of Vocational Education Through the “Three Education” Reform—Based on the Background of Industrial Digital Transformation. Vocational Education Forum, 2022(7).
[2] Ren Jiemin. Positioning of Teachers' Roles in Multimedia Teaching Environment Under Constructivist Learning Theory. Curriculum Education Research, 2018(33): 193–194.
[3] Zhang Zhe, Chen Xiaohui, Qin Pengxi, et al. Meta-analysis of Factors Influencing Teachers' Use of Intelligent Technology in Teaching. Modern Distance Education, 2019(2).
[4] Zhao Gang, Zhu Wenjuan, Hu Biling, et al. A simple teacher behavior recognition method for massive teaching videos based on teacher set. Applied Intelligence, 2021, 51(12).
[5] Guo J, Lü J, Wang R, et al. Deep learning model-driven teacher-student classroom behavior recognition. Journal of Beijing Normal University (Natural Science Edition), 2021, 57(06): 905–912.
[6] Ding N. Intelligent analysis and recognition of teacher body movements in secondary school classroom videos. Master's Thesis, Central China Normal University, 2020.

[7] Wang T. Research on classroom teaching behavior analysis methods based on human motion detection. Master's Thesis, Chang'an University, 2020.
[8] Zheng Y. A posture recognition-based method for teacher teaching behavior evaluation. Software Engineering, 2021, 24(04): 6–9.
[9] P Shiyan, Z Anran, L Shuhui, Z Zhiqi. Automatic recognition of teachers' nonverbal behavior based on dilated convolution. 2022 IEEE 5th International Conference on Information Systems and Computer Aided Education (ICISCAE), 2022: 162–167.
[10] Ma X. Research and application of teacher behavior recognition for smart classrooms. Master's Thesis, Yunnan Normal University, 2023.
[11] Liu Y. Research on the evaluation of teaching effectiveness for primary and secondary school teachers in digital education environments. Master's Thesis, Northwest Normal University, 2023.
[12] Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the Kinetics dataset. Computer Vision and Pattern Recognition (CVPR), IEEE Conference on, 2017: 4724–4733.
[13] Kuehne H, Jhuang H, Garrote E, et al. HMDB: A large video database for human motion recognition. In Proc. ICCV, 2011: 2556–2563.
[14] Soomro K, Zamir A, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild. Computer Science, 2012.
[15] Pang S, Hao J, Hu H, et al. Teacher behavior recognition method based on spatiotemporal graph convolutional neural networks. Journal of Central China Normal University (Natural Science Edition), 2023, 57(05): 715–723.
[16] Liu Q, He H, Wu L, et al. Classroom teaching behavior analysis method based on artificial intelligence and its application. China Educational Technology, 2019(09): 13–21.
[17] Fu D, Zhang H, et al. Educational information processing. Beijing: Beijing Normal University Press, 2011.
[18] Mu S, Zuo P. Research on classroom teaching behavior analysis methods in an information-based teaching environment. Educational Technology Research, 2015, 36(09): 62–69.
[19] Yang F, Wang T. SCB-dataset: A dataset for detecting student classroom behavior, 2023.
[20] Jegham N, Chan Y, Marwan A, et al. Evaluating the evolution of YOLO (You Only Look Once) models: A comprehensive benchmark study of YOLO11 and its predecessors. arXiv preprint arXiv:2411.00201, 2024.
[21] Rahman M, Marculescu R. Medical image segmentation via cascaded attention decoding. IEEE/CVF Winter Conference on Applications of Computer Vision, 2023: 6222–6231.
[22] Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L. Mobilenetv2: Inverted residuals and linear bottlenecks. IEEE Conference on Computer Vision and Pattern Recognition, 2018: 4510–4520.
[23] Zhang X, Zhou X, Lin M, Sun J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. IEEE Conference on Computer Vision and Pattern Recognition, 2018: 6848–6856.
[24] Krizhevsky A, Hinton G. Convolutional deep belief networks on CIFAR-10. Unpublished manuscript, 2010, 40(7): 1–9.
[25] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in Neural Information Processing Systems, 2017, 30.
[26] Ma N, Zhang X, Zheng H, Sun J. Shufflenet v2: Practical guidelines for efficient CNN architecture design. Proceedings of the European Conference on Computer Vision (ECCV), 2018: 116–131.
[27] Han K, Wang Y, Tian Q, et al. Ghostnet: More features from cheap operations. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 1580–1589.
[28] Chen J, Kao S, He H, et al. Run, don’t walk: Chasing higher FLOPs for faster neural networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 12021–12031.
[29] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7132–7141.
[30] Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 13713–13722.