A Hybrid YOLO-Based Approach for Fine-Grained Detection of Classroom Student Behaviors

Conference Paper · December 2024

DOI: 10.1109/ICCIT64611.2024.11022537

5 authors, including:

Mahtab Uddin

60 PUBLICATIONS **257** CITATIONS

SEE PROFILE

Hridoy Das

United International University

5 PUBLICATIONS **0** CITATIONS

SEE PROFILE

Hazera Khtaun

United International University

4 PUBLICATIONS **0** CITATIONS

SEE PROFILE

Apu Kumar Roy

BRAC University

1 PUBLICATION **0** CITATIONS

SEE PROFILE

2024 27th International Conference on Computer and Information Technology (ICCIT)
20-22 December 2024, Cox's Bazar, Bangladesh

A Hybrid YOLO-Based Approach for Fine-Grained Detection of Classroom Student Behaviors

Hridoy Das*, Hazera Khatun Hira*, Mahtab Uddin†, Apu Kumar Roy‡, Asif Mahmud*
*Dept. of Computer Science and Engineering, United International University, Dhaka, Bangladesh
Emails: {hdas201278, hhira201276, amahmud201178}@bscse.uiu.ac.bd

†Institute of Natural Sciences, United International University, Dhaka, Bangladesh
Email: mahtab@ins.uiu.ac.bd (Corresponding Author)

‡Dept. of Computer Science and Engineering, BRAC University, Dhaka, Bangladesh
Email: apu.kumar.roy@g.bracu.ac.bd

***Abstract*—This paper introduces a system for real-time identification of student behavior in classrooms, employing YOLOv8 for precise object detection and Convolutional Neural Networks (CNNs) for behavior analysis, this technology enables the identification of body language to categorize behaviors such as attentiveness and participation. The scalable solution is guaranteed to work effectively for various classroom sizes due to the implemented approach. The system possesses the potential to revolutionize classroom management, as evidenced by its initial findings demonstrating its accuracy. Furthermore, this technology facilitates the provision of immediate feedback to teachers, while simultaneously enabling them to identify long-term behavioral patterns. It serves as a valuable tool that assists teachers in delivering tailored instruction and fostering enhanced student engagement. The source code and models are publicly released: GitHub Repository.**

***Index Terms*—YOLOv8, YOLOv8+CNN(HEAD), Deep Learning, Computer Vision, Object Detection, Student Classroom Behavior Detection**

I. Introduction

Monitoring student behavior in classrooms is crucial for improving educational engagement, but traditional methods such as manual observation are time-consuming and inconsistent. In the realm of object detection, particularly the YOLO algorithms, stands out as a highly competitive group of models. Notably, YOLOv8 is renowned for its exceptional real-time object detection capabilities, characterized by its remarkable speed and accuracy. However, these models face limitations in recognizing unconscious gestures, such as hand raising or reading, in situations characterized by high levels of crowding. To address these issues, the hybrid YOLOv8+CNN(head) paradigm is developed, resulting in the inclusion of a YOLOv8-fast detection to a CNN-specific layer to extract high-level features that achieve detailed behavior recognition. This technique outperformed the previous one (precision 0.4, recall 0.5) and other models such as YOLOv7 and ResNet50, thus, precision was increased to 0.7 and recall to 0.7. The glory of the YOLO models lies in their speed; however, to the CNN level's benefit is the enhancement of the model's ability to detect subtle actions which consequently endears the model to classroom settings for immediate use. As for the hybrid method, is a practical and real-time monitoring system of student behavior in the classroom, which targets the issue of student misconduct with better accuracy and scalability than the previously practiced methods that were usually found inadequate. The diversity of models and strategies discovered related to the classroom by previous research. Su et al. [1] have delved into this area before;

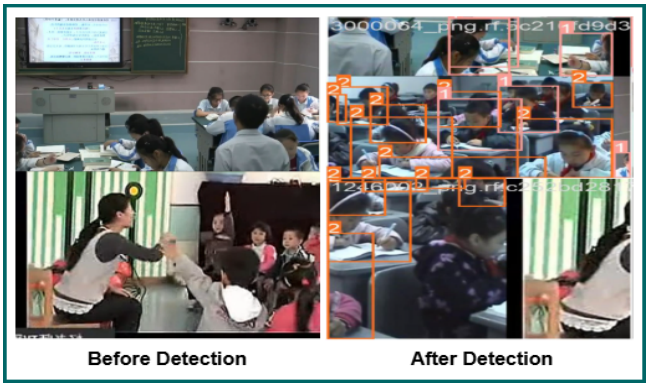

Fig. 1. Comparison Before Vs After Detection.

it was using the pre-trained ResNet50 model that enabled them to adjust and obtain parameters on a certain dataset, classroom behavior phenomenon. Later, the outcomes of the study revealed that ResNet50 achieved parameter adjustments not only at a greater speed than other pre-trained models but also in the identification of different student behaviors. Wei et al. [2] presented an innovative method of using VGG16 and Transfer Learning to classify the seven common classroom behaviors: listening and reading. This technique aimed at improving teaching feedback by correctly recognizing student activities. Likewise, Ji et al. [3] made a further modification to the already existing traditional (CNN) structure and developed three additional network structures and then named them Deep Convolutional Neural Networks(DCNN). Their focus was on

979-8-3315-1909-4/24/$31.00 ©2024 IEEE

investigating specific area effects on a target over time and in space, using image feature maps trained on several datasets during the learning stage to enhance behavior recognition. This study sets out to include a unique contribution to how object detection networks can be utilized to automate the identification of student behaviors in a classroom setting. To facilitate this research, a common dataset was constructed consisting of activities such as hand-raising, writing, and reading, which are characteristic of the classroom. This dataset lays the groundwork for future research on student behavior recognition during instructional contexts that use advanced automated systems improving the efficiency and dynamism of teaching practices. To assure dataset quality, we analyzed all data and performed a final benchmark testing using the hybrid model that we proposed consisting of CNN and YOLOv8. This process has produced consistent training data. Since YOLOv8 is a relatively newer model, we trained it on our dataset as well to improve its detection capability. With the application of multiple layers, CNN is very important in the process of receiving input data as it receives more of the elementary features making identification of other complex objects possible. Such challenges are particularly prevalent in dense classroom or student-to-student occlusions situations. Such issues occasionally result in inconsistencies in the bounding boxes that the algorithm generates. shown in Figure 1. The remainder of the paper is organized as follows: section II provides a comprehensive Literature Review. Section III presents the datasets and methodology. Section IV presents the Result Analysis. Section V presents the discussion, limitations & future work. Finally, the conclusion is presented in Section VI.

II. Literature Review

Recently, machine learning has made it possible for computers to replace people in a variety of tasks. Deep learning, a part of machine learning, accurately describes object characteristics and has been successfully applied in computer vision. Computer-vision-based methods provide the benefits of ongoing observation, objective evaluation, and real-time student behavior interaction in the classroom. However, it encounters difficulties such as backdrop congestion, picture distortion, multiple points of view, and ambiguous student stances. To address these difficulties and monitor such scenarios, many approaches have been put forth. Object detection is the process of determining which object categories are present in a picture, their corresponding confidence levels, and the position and size of each item with rectangle boundaries. This area of computer vision has developed quickly and is still evolving. Convolutional neural networks were used to significantly improve picture classification tasks, as demonstrated by the 2012 ImageNet classification challenge, which marked a major leap in object identification approaches' efficacy. There are two types of deep learning-based object detection algorithms: single-stage algorithms and two-stage algorithms [4] [5]. YOLOV3 [4], YOLOV4 [6], YOLOV7 [7] [8], YOLOV8 [9], and other single-stage target identification algorithms are often utilized.

A. Student Learning Behavior Recognition Based on Deep Learning

The YOLO series has higher detection speed but comparatively lesser accuracy since it can directly forecast the location and category of objects [10] [11]. From version 1 to version 8, the YOLO has steadily improved, which has increased the most recent YOLO-based network design inference speed, flexibility, and deployability. Tian et al. [12] enhanced YOLO to handle low-resolution features and enhanced network performance using fuzzy features and the DenseNet technique. Gomaa et al. [13] increased the detection speed and accuracy by introducing K-means clustering and KLT tracker into YOLOv2 and also proposing a new counting approach. Ren et al. [14] improved the feature extraction of the YOLOv4 network by using a structure with jumping routes for feature extraction and integrating top-down and bottom-up techniques. Yu et al. [15] enhanced the YOLOv4 backbone feature extraction network and later implemented an adaptive picture scaling technique to improve the network's capacity for learning and minimize computation. Tang et al. [16] used the feature pyramid structure of YOLOv5 in conjunction with a weighted bi-directional feature pyramid network (BiFPN) to detect targets. YOLOv6 embraced the RepVGG architecture, augmenting GPU device adaptability and engineering adaptations [17]. Fan Yang et al. [18] proposed the Student Classroom Behavior Detection system based on based on YOLOv7-BRA (YOLOv7 with Bi-level Routing Attention). YOLOv7 incorporated module re-referencing and dynamic tag assignment strategies, bolstering both speed and accuracy, effectively outpacing existing target detectors in the 5 FPS to 160 FPS range [7].

B. convolutional Neural Network

In independent and identically distributed settings, CNNs have demonstrated state-of-the-art performance; yet, they are still quite susceptible to distributional changes, adversarial noise, and common picture corruptions. Conversely, single-stage detection techniques such as SSD [19]. Many SVM classifiers and bounding box regressors must be kept since R-CNN restricts the size of the input picture, and the feature computation of each candidate box produced by the SS technique is too complex. There are several redundancies in the model, which is complicated. Fast R-CNN was progressively suggested by Cir-shick et al. [20]. In contrast, Fast R-CNN significantly increases the detection time by replacing the SVM classifier with the softmax classifier and accepting images of any size as input. Faster R- CNN [21] effectively addresses this issue, however, it still requires a lot of time to create candidate frames because it employs the Selective Search method. RPN is used. Regional Propose Network as an alternative to the Selective. To provide the network model with a better classification capacity, Tang et al. [22] enhanced Faster R-CNN for the issues of occlusion and poor resolution. They

then employed merged ROI pooling (MRP) to fuse feature maps at various levels.

III. Data Collection Strategy and Methodology

This section outlines the approach for collecting, and preparing the dataset, and the methodology approaches used in this study.

A. Datasets

Computer vision has emerged as a pivotal tool in contemporary research endeavors aimed at identifying and evaluating student activities within classroom settings. However, a notable impediment to the substantial potential of this field lies in the scarcity of publicly accessible data about student behavior within educational institutions. This deficiency poses a substantial challenge to the development and generalization of behavior detection methodologies within educational environments. A publicly available dataset was utilized in this approach. This dataset provides valuable insights into students' behavior and classroom dynamics, making it an invaluable resource for Artificial Intelligence (AI) research in the field of education. The dataset comprises 18,400 labels and 4,200 images, with an average of 4.4 individuals annotated per image. The dataset categorizes student behavior into three distinct categories: reading, writing, and hand-raising. These categories are illustrated in Figure 2. The dataset was captured from various viewpoints, including front, side, and back angles. Preprocessing techniques, such as data augmentation through rotation, zooming, and flipping, were employed to enhance dataset diversity and mitigate overfitting. The model's setting in the classrooms isn't always conducive. As seen in Figure 3, there are a lot of photos that include both indistinct and crowded classroom atmospheres. We discovered that in some scenarios, writing and reading also have a lot in common. These photos supply the model with difficult and demanding settings and configurations, enabling it to produce more accurate results.

Fig. 2. Three Samples Categories in the Dataset.

Fig. 3. Difficulties with the Dense Environment and Unclear Images in the Dataset.

B. Proposed Methodology

To identify student behaviors within classroom images, two distinct models, YOLOv8 Fig.4 and an extended version developed called YOLOv8+CNN(HEAD) Figure 5. YOLOv8 utilizes the state-of-the-art YOLOv8 object detection architecture, and this combines this with a custom CNN header layer designed to extract high-level features specific to student behaviors. Leveraging these two complementary models, aimed to robustly identify key student activities like hand-raising, reading, and writing in diverse classroom settings. Both models were rigorously evaluated on training and validation data for metrics like loss and accuracy. Those will be discussed in the following subsection about the working method in detail. Overall, these methods showcase modern deep-learning techniques for robust student behavior detection in a classroom setting.

1) Model: Yolov8: To advance student behavior detection within classroom images, employed the YOLOv8 object detection framework [23] as depicted in Figure 4. This experimentation centered on the meticulously curated 4.2k HRW YOLO dataset, hosted on Google Drive, featuring images annotated with YOLOv7 format labels. This dataset comprehensively covered student behaviors, including hand-raising, reading, and writing. Preceding model training, standard preprocessing techniques such as random resized cropping and normalization were applied for data consistency and diversity augmentation. The instantiation of the YOLOv8 model was facilitated through the Ultralytics library [24], designed for YOLO models. The training process was conducted on a local machine using Go for efficient execution and orchestration. Google Drive facilitated seamless dataset access, streamlining workflows. Multiple epochs were trained with iterative hyperparameter tuning for optimal model performance. Following training, a comprehensive evaluation was conducted, utilizing training and validation metrics to assess model convergence and generalization. Visualizations, including loss vs epochs and validation accuracy vs epochs, were generated using the matplotlib library [25]. Adhering to ethical standards, privacy, and consent were prioritized throughout the handling of the 4.2k HRW YOLO dataset. To enhance transparency and reproducibility, the complete set of code, dataset, and configurations has been publicly shared on GitHub. This detailed methodology integrates state-of-the-art object detection techniques with ethical considerations. Subsequent sections provide a thorough analysis of the results, offering nuanced

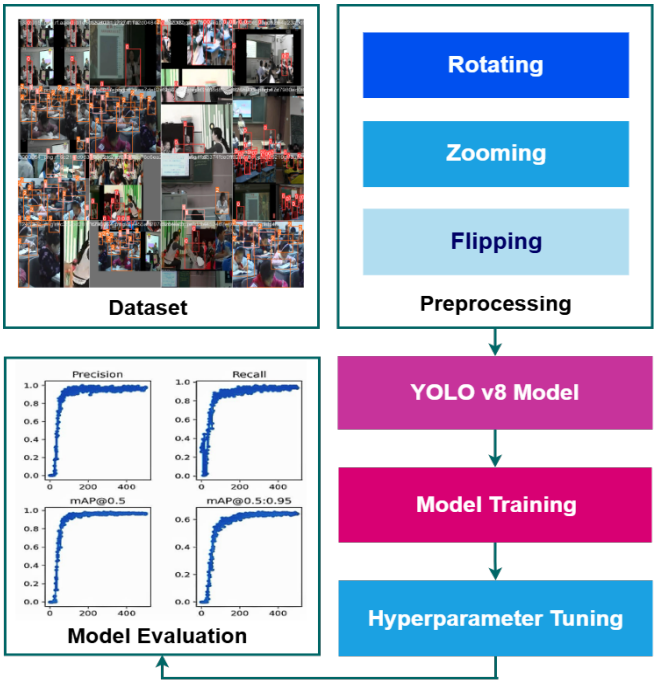

Fig. 4. An Applied Method of Yolov8.

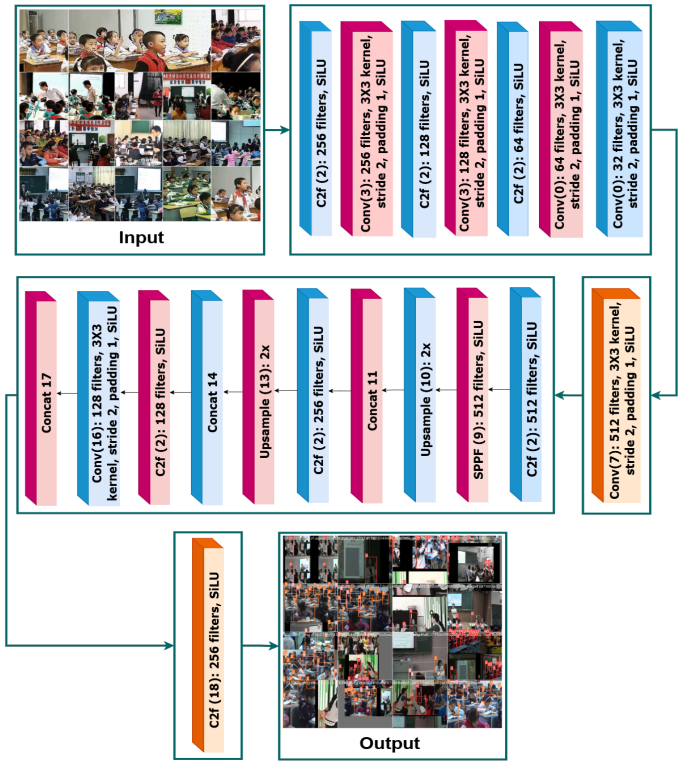

Fig. 5. An Applied Method of YOLO v8 + CNN Layer (Head).

insights into the efficacy and robustness of our approach.

2) Model: Yolov8+CNN(head): This approach proposed a unique way to identify student behavior inside a classroom, combining YOLOv8 for object recognition with a specialized CNN layer for better feature extraction. The YOLOv8+CNN architecture tries to capture detailed patterns linked with diverse student activities in classrooms. The methodology 5 combines YOLOv8 with a custom CNN head layer for enhanced student behavior detection in classrooms. Input images are processed through convolutional (Conv) and Cross-Stage Partial (C2f) layers, extracting hierarchical features with SiLU activation. Multi-scale features are refined using Spatial Pyramid Pooling (SPP) and up-sampling layers. The custom CNN head focuses on high-level behavior-specific features, producing precise detection outputs for activities like hand-raising and writing. The Ultralytics package made it easier to instantiate this hybrid model.

IV. Result Analysis

This section evaluates the performance of the proposed YOLOv8-based and YOLOv8+CNN models for detecting student behaviors in classroom environments.

A. Experimental Setup

To improve the detection of student behavior in classroom images, experiments were conducted using a local machine with a MAC Unified GPU. The first model employed the YOLOv8 object detection framework and a YOLO dataset containing 4.2k high-resolution images, meticulously annotated to capture diverse student behaviors such as hand-raising, reading, and writing. The YOLOv8 model instantiated through the Ultralytics library, underwent training for 50 epochs, each comprising 16 batches. Hyperparameters were tuned iteratively, with a learning rate of 0.001 and the Adam optimizer with weight decay, ensuring optimal convergence. Post-training, model performance was evaluated using training and validation metrics, and results were visualized using Matplotlib. The second model utilized a hybrid architecture combining the YOLOv8 framework with a custom CNN layer as the head. The same 4.2k YOLO dataset was employed, and accessed via local storage. Training for this model spanned 60 epochs, with 16 batch sizes per epoch. Hyperparameter optimization involved setting a learning rate of 0.001429 and leveraging the Adam optimizer with precise weight adjustments for optimal outcomes. After training, model accuracy was assessed using training and validation metrics, with Matplotlib employed for result visualization.

B. Experimental Results

The performance of the YOLOv8+CNN(HEAD) model has been compared with the existing models I for classroom behavior detection. The YOLOv8 model has demonstrated a precision of 0.4 and a recall of 0.5, which is competitive with YOLOv7 in Fan Yang et al.'s work, where precision and recall are approximately 0.6 and 0.55, respectively. However, YOLOv8+CNN(HEAD) has significantly improved these metrics, achieving a precision of 0.7 and a recall of 0.65, to the integration of CNN for better feature extraction. In contrast, models like ResNet50 and VGG16 have been slower and lacked clear performance metrics. 3D CNNs, while improving

behavior recognition over time and space, are computationally expensive and less suited for real-time applications. In experiments, YOLOv8 has shown promise but with room for improvement. It has successfully identified half of all positive examples, but with only 40% accuracy in positive predictions, leading to a moderate mAP@50 of 0.5 and a lower mAP@50-95 of 0.3. YOLOv8+CNN(HEAD), with a recall of 65% and precision of 70%, has been more effective at identifying positive examples, achieving a better mAP@50 of 0.72, though its performance dips under stricter criteria (mAP@50-95 of 0.55). Both models suffer from false positives, but YOLOv8+CNN(HEAD) improves recall and ranking, making it a more robust solution for real-time classroom behavior detection. To provide a more comprehensive evaluation of Precision and Recall, introduced the metrics of Average Precision (AP, Eq. 1) and mean Average Precision (mAP, Eq. 2). These types of evaluation metrics calculate the average Precision over a range of Recall values, thereby providing a more holistic assessment of the model's performance. Precision values at all Recall levels, while mAP calculates the mean AP value averaged across various categories or classes.

TABLE I
COMPARATIVE PERFORMANCE OF VARIOUS STUDENT BEHAVIOR MONITORING MODELS

Method	Precision	Recall	mAP@50	mAP@50-95
YOLOv8 (Pre-trained)	0.4	0.5	0.5	0.3
YOLOv8 + CNN (This work)	0.7	0.65	0.72	0.55
YOLOv7 (Yang et al.)	0.6	0.55	0.6	0.4

$$mAP = \frac{1}{N}\sum_{i=1}^{N} APi \tag{1}$$

$$AP = \frac{1}{11}\sum_{Recall=0}^{1} p(Recall) \tag{2}$$

The consistent decline in training and validation losses for YOLOv8 and YOLOv8+CNN(HEAD) in Figures 6 and 7 shows effective learning from training data. Our YOLOv8+CNN(HEAD) model outperformed the original YOLOv8, especially on bounding box and classification loss metrics. Though both models had moderate precision and recall, YOLOv8+CNN(HEAD) outperformed YOLOv8, indicating improved object detection capabilities due to its additions. Both models have shown promising results, and further training could improve them. Our YOLOv8+CNN(HEAD) model outperforms YOLOv8 on several key metrics.

V. DISCUSSION

The YOLOv8+CNN hybrid model outperforms YOLOv8 and YOLOv7 in real-time student behavior detection. Combining YOLOv8's rapid detection with a CNN layer improves precision to 70% and recall to 65%. This allows it to capture subtle behaviors like hand-raising and reading, which traditional YOLO models struggle with. Unlike computationally heavy models like 3D CNNs, the YOLOv8+CNN model maintains moderate complexity, making it suitable for real-time deployment across classrooms of varying sizes. Scalability tests confirm consistent performance across small and large environments, without significant delays. However, false positives remain an issue, particularly in crowded classrooms. Future work will focus on reducing these errors and optimizing the model for more dynamic scenarios.

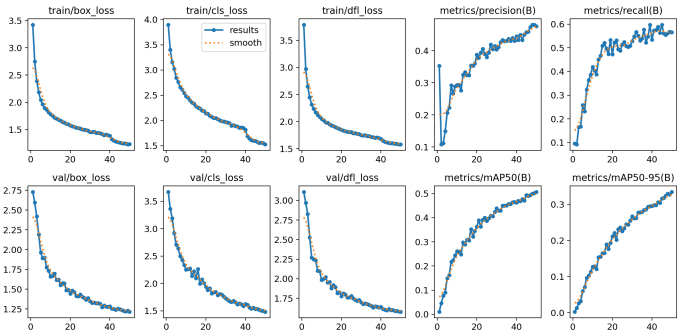

Fig. 6. YOLOv8 Results.

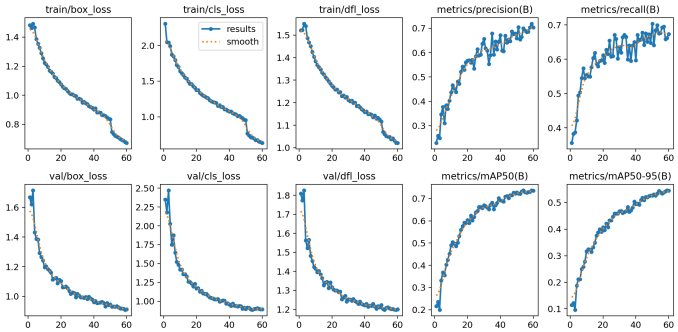

Fig. 7. YOLOv8 + CNN (HEAD) Results.

A. Future Work and Limitations

The model's main limitation is the occurrence of false positives in crowded settings. Future improvements will aim to enhance detection accuracy with advanced filtering techniques. While the model has been tested in various classroom sizes, further validation in more diverse environments, including virtual classrooms, is necessary. Additional training data and optimization for faster inference times will also be explored to improve scalability and accessibility in real-world applications.

VI. CONCLUSION

The YOLOv8+CNN(head) hybrid model has demonstrated exceptional efficacy in real-time student behavior detection within educational settings. This model merges the object detection capabilities of YOLOv8 with a modified CNN layer for feature extraction, yielding precision and recall values of 70% and 65%, respectively. These outcomes are notably

impressive. higher than those of conventional YOLO models and those of the other existing models like the YOLOv7, or ResNet50, for instance. The incorporation of a CNN header effectively addresses the limitations, enabling more accurate recognition of behaviors like hand-raising, reading, and writing. In conclusion, the YOLOv8+CNN model offers a scalable and efficient solution for real-time student behavior monitoring, balancing speed, and accuracy to enhance classroom management and provide educators with valuable insights into student engagement.

REFERENCES

[1] L. Su and W. Zheng, "Classroom behavior recognition and teaching quality evaluation system based on deep learning," in *International Conference on Cloud Computing, Performance Computing, and Deep Learning (CCPCDL 2023)*, vol. 12712. SPIE, 2023, pp. 373–379.

[2] Y. Wei, D. Qin, J. Hu, H. YAO, and Y.-f. SHI, "Recognition of students' classroom behavior based on deep learning [j]," *Modern Educational Technology*, vol. 29, no. 7, pp. 87–91, 2019.

[3] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

[4] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[5] S. Wu, Y. Xu, and D. Zhao, "Survey of object detection based on deep convolutional network," *Pattern recognition and artificial intelligence*, vol. 31, no. 4, pp. 335–346, 2018.

[6] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[7] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7464–7475.

[8] K. O. Monnikhof, P. Areerob, Z. Wu, T. Tanasnitikul, W. Kumwilaisak *et al.*, "Novel personal protective equipment detection technique with attention-based yolov7 and human pose estimation," *APSIPA Transactions on Signal and Information Processing*, vol. 12, no. 1, 2023.

[9] H. Lou, X. Duan, J. Guo, H. Liu, J. Gu, L. Bi, and H. Chen, "Dc-yolov8: Small-size object detection algorithm based on camera sensor," *Electronics*, vol. 12, no. 10, p. 2323, 2023.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[11] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[12] Y. Tian, G. Yang, Z. Wang, H. Wang, E. Li, and Z. Liang, "Apple detection during different growth stages in orchards using the improved yolo-v3 model," *Computers and electronics in agriculture*, vol. 157, pp. 417–426, 2019.

[13] A. Gomaa, T. Minematsu, M. M. Abdelwahab, M. Abo-Zahhad, and R.-i. Taniguchi, "Faster cnn-based vehicle detection and counting strategy for fixed camera scenes," *Multimedia Tools and Applications*, vol. 81, no. 18, pp. 25 443–25 471, 2022.

[14] L. Tang, T. Xie, Y. Yang, and H. Wang, "Classroom behavior detection based on improved yolov5 algorithm combining multi-scale feature fusion and attention mechanism," *Applied Sciences*, vol. 12, no. 13, p. 6790, 2022.

[15] J. Yu and W. Zhang, "Face mask wearing detection algorithm based on improved yolo-v4," *Sensors*, vol. 21, no. 9, p. 3263, 2021.

[16] M. Hu, Y. Wei, M. Li, H. Yao, W. Deng, M. Tong, and Q. Liu, "Bimodal learning engagement recognition from videos in the classroom," *Sensors*, vol. 22, no. 16, p. 5932, 2022.

[17] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie *et al.*, "Yolov6: A single-stage object detection framework for industrial applications," *arXiv preprint arXiv:2209.02976*, 2022.

[18] F. Yang, T. Wang, and X. Wang, "Student classroom behavior detection based on yolov7-bra and multi-model fusion," *arXiv preprint arXiv:2305.07825*, 2023.

[19] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 21–37.

[20] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[22] L. Tang, C. Gao, X. Chen, and Y. Zhao, "Pose detection in complex classroom environment based on improved faster r-cnn," *IET Image Processing*, vol. 13, no. 3, pp. 451–457, 2019.

[23] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *arXiv preprint arXiv:1506.02640*, 2016.

[24] Ultralytics, "Ultralytics," https://github.com/ultralytics/yolov5.

[25] J. D. Hunter, "Matplotlib: A 2d plotting library," https://matplotlib.org/, 2007.