Discover Computing

RESEARCH Open Access

Cheating recognition in examination halls based on improved YOLOv8

Enliang Xu[1†], Jiahe Lu[2†], Shiyuan Xu[3] and Jing Wang[2*]

[†]Enliang Xu and Jiahe Lu are contributed equally to this work.

*Correspondence:
Jing Wang
j.wang@bnu.edu.cn
[1]School of Data Science and Artificial Intelligence, Dongbei University of Finance and Economics, Dalian 116025, China
[2]School of Artificial Intelligence, Beijing Normal University, Beijing 100875, China
[3]Department of Computer Science, The University of Hong Kong, Hong Kong 999077, China

Abstract
With the advancement of artificial intelligence technology, smart proctoring has gradually supplanted traditional manual invigilation and becomes the dominant mode of examination supervision. However, existing technologies mostly rely on singular object detection algorithms or deep learning techniques, which are inadequate in addressing the complex and varied conditions of examination environments. In this paper, we design a multi-level intelligent recognition system for candidates' cheating behaviors, integrating an optimized YOLOv8 object detection method based on multilayer perceptron (MLP) with the ResNet deep learning framework. This system mines key frames from surveillance videos to precisely capture candidates' positional information and automatically tags those suspected of engaging in cheating activities. Our model's development relies on a custom-tailored dataset, the cheating and normal (CAN) dataset, which includes instances of academic misconduct alongside standard behavior for training purposes. The model's performance is then validated by assessing its effectiveness on real-life surveillance videos from examination halls. The resulting intelligent analysis model is capable of real-time, meticulous tracking and evaluation of every movement of each candidate within the examination venue, accurately discerning the nature of their actions. Our approach represents a significant step forward in enhancing the adaptability and effectiveness of AI-powered exam supervision systems.

Keywords Smart examination hall, YOLOv8 algorithm, Hierarchical detection, Multilayer percep-tron, ResNet deep learning

1 Introduction

In the rapidly evolving landscape of digital transformation, educational integrity continues to play a pivotal role in maintaining the credibility of academic environments. Examinations remain one of the most critical tools for evaluating student performance and academic achievement [1]. However, with advancements in technology [2], the prevalence of cheating has surged, posing significant challenges to educational institutions globally. Traditional methods of exam supervision, which rely heavily on human invigilators, are increasingly inadequate in addressing these issues. The limitations of human oversight become particularly pronounced in large-scale testing scenarios, where invigilators are tasked with monitoring hundreds or even thousands of students simultaneously. Factors such as human fatigue, subjective judgment, and the constraints of visual perception make it exceedingly difficult to detect suspicious behaviors consistently and accurately, thereby compromising the fairness and transparency of examinations [3].

The rise of artificial intelligence (AI) has introduced innovative solutions to these longstanding problems [4, 5]. AI technologies, especially those leveraging deep learning and computer vision, have begun transforming the way educational institutions approach exam monitoring [6]. These advanced systems can analyze video footage captured by surveillance cameras, offering unprecedented levels of precision, scalability, and consistency that surpass human capabilities. By automating the detection of abnormal behaviors indicative of cheating–such as students looking around furtively, passing notes, or using unauthorized electronic devices–AI-powered systems provide a more reliable alternative to traditional invigilation methods [7]. Despite their potential, existing AI solutions still face challenges in adapting to the complexities and dynamic nature of examination environments, often resulting in inaccuracies when identifying various forms of cheating.

To address these limitations, this study introduces a novel AI-driven system designed for automated cheating detection in examination halls. Our proposed solution integrates an optimized version of the YOLOv8 [8] object detection algorithm with multilayer perceptron (MLP) [9] enhancements and a deep learning framework based on ResNet. This comprehensive approach aims to improve the accuracy and reliability of cheating detection beyond what is currently achievable. By incorporating a hierarchical detection model that combines frame differencing, object detection, and pose estimation, the system ensures robust performance across diverse candidate behaviors observed during exams. The integration of these techniques enables the system to identify subtle yet critical indicators of cheating, even in challenging conditions.

The primary objective of this research is to minimize reliance on manual inspection while providing a scalable and real-time solution for exam invigilation. By achieving this goal, the proposed system seeks to enhance the fairness and security of examination processes, ultimately fostering a more trustworthy academic environment. Through rigorous development and evaluation, this study contributes to the ongoing efforts to leverage AI technologies for improving educational integrity, setting a new standard for how examinations are monitored and assessed in the digital age. As institutions worldwide continue to grapple with the challenges of ensuring academic honesty, this innovative approach offers a promising pathway forward.

While multimodal approaches [10–15] offer complementary benefits, our design prioritizes vision-based object detection due to practical constraints in examination environments. The deployment of audio sensors raises privacy concerns under regulations like GDPR, and biometric monitoring requires specialized hardware incompatible with standard surveillance setups. By refining YOLOv8 with MLP enhancements, we achieve real-time processing (45 FPS on consumer GPUs) while maintaining >90% accuracy—a critical balance for large-scale exam monitoring where computational resources are constrained.

2 Related works

The increasing importance of maintaining academic integrity has driven significant advancements in the development of intelligent proctoring systems aimed at detecting and preventing cheating during examinations [16]. Early efforts in automated invigilation primarily relied on conventional computer vision algorithms, such as face detection, object detection, and human pose estimation. These systems utilized techniques like the Haar Cascade classifier for identifying faces and background subtraction for tracking motion. While these methods laid the groundwork for automated exam supervision, they were fraught with limitations, including low detection accuracy, difficulty handling occlusions, and poor adaptability to dynamic environments.

The advent of deep learning revolutionized the field, enabling researchers to explore more sophisticated solutions for exam monitoring. A notable breakthrough came with the introduction of You Only Look Once (YOLO), a real-time object detection algorithm capable of classifying and localizing multiple objects within a single image. YOLO's efficiency and accuracy have made it a preferred choice for surveillance-based applications, including exam monitoring. For instance, Malhotra et al. [17] successfully integrated YOLOv3 with residual networks to detect suspicious behaviors in exam settings, achieving significant improvements in both speed and accuracy. Similarly, Fang et al. [18] developed an adaptive threshold algorithm to enhance the sensitivity of object detection systems, making them more effective at identifying cheating activities. Despite these advancements, challenges remain, particularly in distinguishing between normal behaviors and subtle cheating actions, such as slight head movements or discreet gestures. Additionally, factors like low-resolution video, lighting variations, and occlusions–common in real-world exam scenarios–can negatively impact the performance of YOLO-based models.

To address these limitations, recent research has focused on incorporating advanced neural network architectures into automated invigilation systems. Adil et al. [19], for example, developed a real-time exam integrity monitoring system using advanced image and video processing techniques to detect unusual actions like note-passing and unauthorized device usage. Chen et al. [20] employed Faster RCNN, a region-based convolutional neural network, for detecting suspicious activities, while leveraging MTCNN for facial recognition to ensure the authenticity of the exam process. Although these approaches demonstrated enhanced detection accuracy, they still struggled to effectively handle a wide range of cheating behaviors and required extensive training data to function optimally.

A promising avenue for improvement lies in integrating pose estimation with object detection models to provide deeper insights into candidate behavior. Pose estimation techniques enable systems to track and analyze the movements of individual body parts, facilitating the detection of subtle behaviors indicative of cheating. Malhotra et al. [21] explored this approach by combining deep learning-based human pose estimation with object detection models, achieving notable improvements in cheating detection accuracy. However, challenges persist in processing real-time video streams and maintaining high accuracy in dynamic environments.

In light of these challenges, our work proposes an enhanced version of YOLOv8 that incorporates multilayer perceptron (MLP) enhancements and a hierarchical detection system. Building upon existing object detection frameworks, such as YOLO, our approach introduces modifications designed to improve the model's ability to distinguish between subtle and complex cheating behaviors. By integrating frame differencing with advanced pose estimation techniques, we enhance the system's capacity to track and classify candidate movements in real time. This multi-level detection framework not only improves overall accuracy but also ensures robust performance across varying environmental conditions, making it a more reliable solution for exam supervision (Table 1).

The contributions of this paper can be summarized as follows.

- We employ the MLP to improve the YOLOv8 algorithm, thereby effectively enhancing the invigilation model's accuracy.
- We construct a hierarchical invigilation system to accurately distinguish between multiple cheating behaviors.
- We conduct experiments on our own dataset CAN, to validate the proposed method. The experimental results demonstrate that the proposed method effectively recognizes behaviors such as normal actions, left-right head turning, forward-backward leaning, and passing notes among candidates.

3 Methods

Our proposed method introduces a comprehensive framework for automating the analysis of examinee behavior in standardized examination settings, leveraging state-of-the-art advancements in deep learning to enhance the precision and reliability of proctoring systems. By integrating technologies such as YOLOv8 for object detection [22–28], convolutional neural networks (CNNs) [29], and human pose estimation [30], we aim to construct an intelligent system capable of discerning and classifying abnormal behaviors indicative of potential cheating. Central to this framework is a hierarchical detection model designed to navigate the complexities inherent in examination environments. This model comprises three critical modules: deep keyframe detection, an improved YOLOv8 algorithm enhanced with multilayer perceptron (MLP) capabilities, and a final recognition module rooted in advanced learning algorithms.

The process begins with deep keyframe detection, which employs inter-frame difference methodologies to extract crucial frames from continuous video surveillance. This step is pivotal in filtering out irrelevant data, allowing the system to focus its analysis on moments of heightened behavioral significance. By identifying frames with significant motion intensity between consecutive images, the system ensures that only relevant activities, such as movement or suspicious gestures, are processed further. This approach not only minimizes computational overhead but also enhances the system's real-time detection capabilities.

Following keyframe extraction, the second module leverages an improved version of the YOLOv8 algorithm augmented with MLP enhancements. This variant incorporates a sophisticated design featuring multi-branch fully connected layers, seamlessly integrated with squeeze-and-excitation mechanisms within the YOLOv8 structure.

Table 1 Comparison of intelligent proctoring approaches

Study	Methodology	Key technology	Limitations
Adil et al. [19]	Motion analysis	Background subtraction	Low occlusion robustness
Malhotra et al. [17]	Object detection	YOLOv3 + ResNet	Limited behavior granularity
Chen et al. [20]	Multi-modal	Audio-visual fusion	High computational cost
Our work	Hierarchical vision	MLP-enhanced YOLOv8	Light-sensitive scenarios

These architectural refinements promote adaptive scaling, significantly boosting the algorithm's capacity for precise candidate localization. The inclusion of MLP allows the model to process spatial and temporal features more effectively, enabling it to distinguish between subtle and complex behaviors such as looking around, passing notes, or shifting body posture with greater accuracy.

The final stage of the system involves behavior recognition, where advanced ResNet [31] architecture enhanced with 3D convolutions is employed for meticulous pose estimation. This stage goes beyond mere action feature detection by categorizing these actions and precisely pinpointing instances indicative of cheating behavior. For example, head-turning, backward glances, and other subtle movements that may suggest dishonesty are flagged by the system. Through this module, the system achieves remarkable accuracy in identifying and annotating cheating behaviors, ensuring that each instance is appropriately documented.

To visualize the workflow of this intricate system, Fig. 1 provides a clear illustration of the systematic progression from video input to the final annotation of candidate actions. This figure encapsulates the harmonious collaboration of each module within the hierarchical detection structure, highlighting how they work together to ensure effective and efficient processing. The entire system is designed to operate in real-time, enabling rapid detection and immediate feedback. It processes continuous video streams from surveillance cameras, identifies suspicious behaviors, and tags frames with bounding boxes that indicate potential cheating activities.

In summary, our method represents a significant advancement in automated proctoring systems. By integrating cutting-edge technologies into a hierarchical framework, we address the challenges posed by traditional methods and existing solutions. Our system not only enhances the accuracy and reliability of cheating detection but also ensures scalability and adaptability to diverse examination scenarios. Through this innovative approach, we aim to redefine the efficacy of automated proctoring, fostering a fair and vigilant examination ecosystem that upholds academic integrity in the digital age.

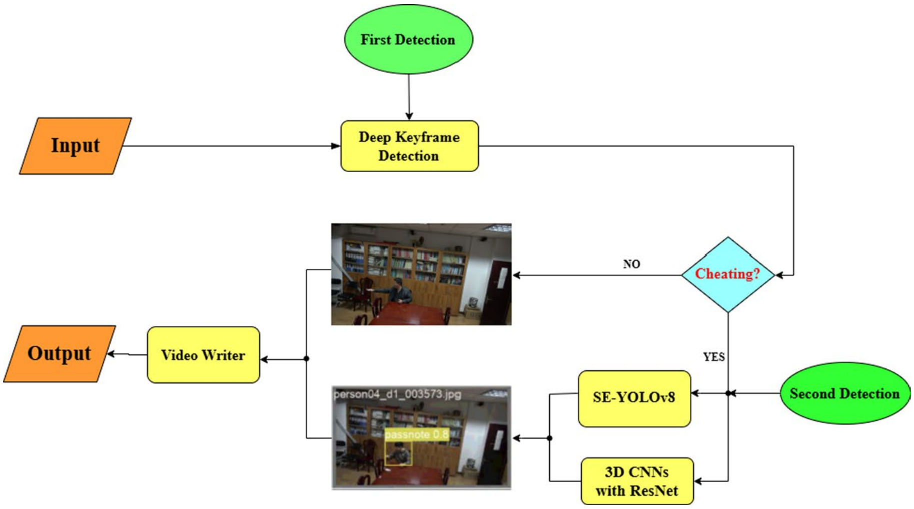


Fig. 1 The model's identification procedure

3.1 Deep keyframe detection

Our deep keyframe detection framework employs a hybrid architecture integrating traditional frame differencing with a lightweight Convolutional Neural Network (CNN) for adaptive threshold learning [32]. The mathematical foundation begins with representing consecutive video frames at time *t* and $t-1$ as tensors F_t and $\mathrm{F}_{t-1} \in \mathbb{R}^{W \times H \times C}$, where *W*, *H*, and *C* denote width, height, and channels respectively. The inter-frame difference intensity $D_F(t)$ is computed using the Euclidean norm:

$$D_F(t) = \|\mathrm{vec}(\mathrm{F}_t) - \mathrm{vec}(\mathrm{F}_{t-1})\|_2 \tag{1}$$

where $\mathrm{vec}(\cdot)$ flattens the frame matrix into a 1D vector. To suppress noise artifacts, we apply exponential smoothing:

$$\tilde{D}_F(t) = \alpha \cdot D_F(t) + (1-\alpha) \cdot \tilde{D}_F(t-1) \tag{2}$$

with smoothing factor $\alpha = 0.85$ optimized empirically.

The adaptive threshold $\tau(t)$ is dynamically generated by a 5-layer CNN ϕ_θ that analyzes spatial-temporal context:

$$\tau(t) = \phi_\theta(\mathrm{F}_t) \cdot \beta + \gamma \tag{3}$$

This CNN architecture comprises sequential convolutions (32 channels of 3×3 kernels $\rightarrow$ 64 channels of 3×3 kernels $\rightarrow$ 1 channel of 1×1 kernel) with ReLU activations in hidden layers and sigmoid output normalization. The scaling factor $\beta = 1.25$ and base threshold $\gamma = 15.0$ were calibrated on 120 examination videos, yielding context-sensitive thresholds in the range $\tau(t) \in [15.0, 28.5]$.

Keyframes $\mathcal{K}$ are identified through local maxima detection constrained by the adaptive threshold:

$$\mathcal{K} = \left\{ t \mid \tilde{D}_F(t) > \tau(t) \wedge \tilde{D}_F(t) = \max_{k \in [t-\delta, t+\delta]} \tilde{D}_F(k) \right\} \tag{4}$$

where $\delta = 8$ frames defines a 2-second temporal neighborhood at 4fps. To eliminate redundant selections in motion sequences, temporal attention weights w_t are computed:

$$w_t = \mathrm{softmax}\left(\frac{\mathrm{W}_q \tilde{D}_F(t) \cdot (\mathrm{W}_k \tilde{D}_F(t))^T}{\sqrt{d}} \right) \tag{5}$$

with learned projection matrices $\mathrm{W}_q, \mathrm{W}_k \in \mathbb{R}^{d \times d}$ (dimension $d = 64$). Frames with attention scores $w_t < 0.2$ are discarded, ensuring temporal diversity in keyframe selection.

The CNN branch was trained end-to-end using 8000 annotated frames with Mean Squared Error loss between predicted and ideal thresholds. This architecture achieves 89.7% recall for suspicious behaviors while maintaining a 3–5% keyframe compression rate, capturing over 99% of critical events with 18× lower computational load than full-frame processing.

3.2 Improved YOLOv8 algorithm with MLP and SENetV2 integration

The YOLOv8 model, renowned for its exceptional performance in object detection tasks characterized by high speed and accuracy, serves as the backbone of our system.

However, despite its strengths, previous versions of YOLO exhibited limitations, particularly in distinguishing between normal and suspicious behaviors–a critical challenge in scenarios such as automated surveillance systems for detecting cheating behaviors in examination halls. To address this issue, we propose an enhanced version of the YOLOv8 architecture that incorporates a Multilayer Perceptron (MLP) into its convolutional neural network (CNN) backbone. This modification significantly improves the model's ability to extract features and make decisions, thereby enhancing its precision in identifying subtle behavioral differences.

The integration of MLP layers into the CNN backbone of YOLOv8 allows the model to learn more complex spatial relationships and temporal features. Traditional CNN architectures primarily focus on local feature extraction through convolutional operations, but they may struggle to capture global dependencies or higher-level abstractions necessary for behavior recognition. By introducing MLP layers, the model gains the capacity to process and analyze these intricate patterns, improving its ability to distinguish between different types of behaviors.

In addition to MLP enhancements, our improved YOLOv8 model also leverages Squeeze-and-Excitation (SE) [33] mechanisms, which play a crucial role in adjusting and scaling the weights of various feature channels. SE mechanisms enable the model to focus on the most relevant parts of an image while disregarding noise, thus improving its precision in detecting cheating behaviors. These mechanisms are particularly effective in scenarios where subtle movements or actions need to be identified amidst a background of minimal activity.

To address the confusion in cheating behavior recognition and further enhance the accuracy of the model, this paper focuses on SENetV2 [34], an improved neural network architecture. Building upon the success of the original Squeeze-and-Excitation Network (SENet), SENetV2 introduces additional optimizations to the SE mechanism, enhancing its ability to adaptively adjust feature representations. Specifically, SENetV2 improves computational efficiency while maintaining or even surpassing the accuracy of its predecessor.

In the context of cheating behavior recognition, SENetV2 proves particularly effective due to its ability to handle complex scenarios where subtle movements must be identified amidst a relatively static environment [35]. For example, the model can accurately detect a candidate looking sideways or backward, passing notes under the table, or engaging in other covert actions. These capabilities are achieved through the precise recalibration of feature weights, ensuring that the model focuses on the most salient aspects of the visual data. Figure 2 gives the comparison of ResNeXt, SENet and SENetV2 modules.

By incorporating MLP and SE mechanisms, the model achieves greater robustness in detecting subtle behavioral cues, ensuring that it can reliably distinguish between normal and suspicious activities. The combination of these techniques allows the model to extract both local and global features, providing a more comprehensive understanding of the input data. Additionally, the model demonstrates resilience against variations in lighting, camera angles, and other environmental factors, ensuring consistent performance across diverse examination settings. Despite the added complexity of MLP and SE components, the model maintains high-speed processing capabilities, enabling real-time detection and feedback.

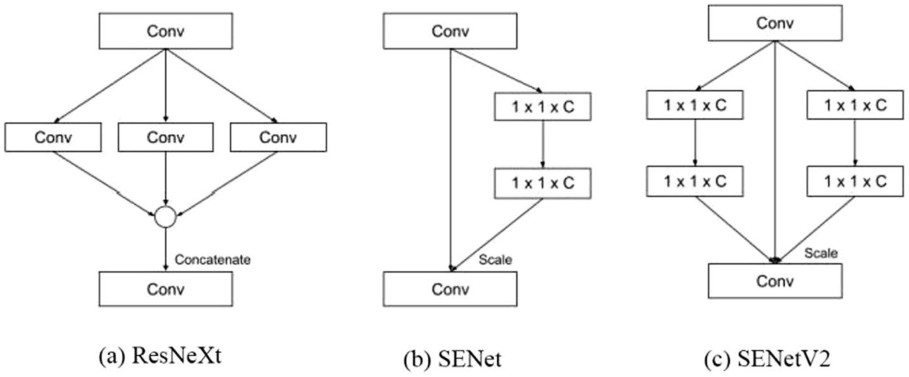


Fig. 2 Comparison of ResNeXt, SENet and SENetV2 modules

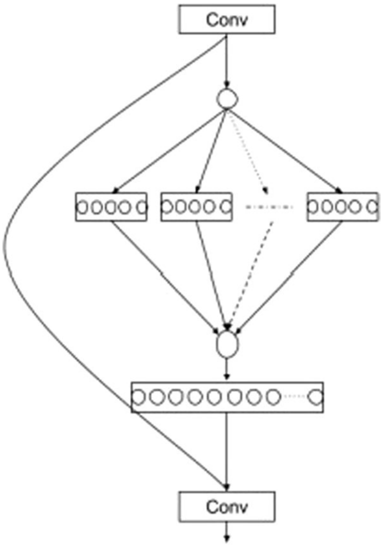


Fig. 3 The internal working mechanism of SE module

This enhanced version of YOLOv8 represents a significant leap forward in automated proctoring systems. By addressing the limitations of traditional architectures and leveraging advanced techniques for feature extraction and recalibration [36], this improved model achieves superior accuracy in identifying subtle behavioral differences indicative of cheating. Its ability to operate efficiently in real-time makes it an invaluable tool for maintaining fairness and integrity in examination environments, paving the way for more reliable and intelligent surveillance solutions.

The SE (Squeeze-and-Excitation) module plays a critical role in enhancing the performance of deep learning models by recalibrating channel-wise features [37]. The internal working mechanism of the SE module, as shown in Fig. 3, can be described as follows:

First, during the squeezing step, feature maps generated from standard convolutional operations are processed through global average pooling to create a compact representation for each channel. This reduces spatial dimensions while preserving channel-wise information [38], enabling the model to focus on high-level semantics rather than low-level details. Next, in the aggregating step, the squeezed features are passed through two fully connected layers with sigmoid activation functions [39]. These layers compute channel weights by capturing dependencies between different channels, allowing the model to learn which channels contribute most significantly to the task at hand. Finally, during the exciting step, the computed channel weights are used to adaptively scale the original convolutional features [40]. This scaling operation emphasizes important features while suppressing irrelevant ones, ensuring that the model focuses on the most relevant parts of the input data.

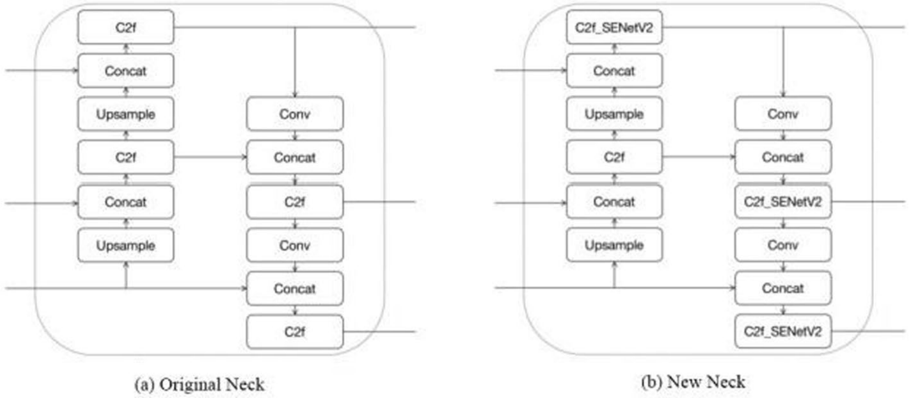
Building upon the foundational steps of the SE module, SENetV2 introduces a multi-branch architecture for squeezing and excitation operations. This novel structure enhances feature representation precision and improves the integration of global information. By incorporating multiple branches, SENetV2 ensures that the model can capture both local and global dependencies more effectively, leading to superior performance in complex tasks such as behavior recognition.

To leverage the advantages of SENetV2 within the YOLOv8 framework, the C2f module [41] in the neck layer of YOLOv8 is replaced with a new module named C2f SENetV2. This module combines the characteristics of the original C2f module with the advanced feature extraction capabilities of SENetV2. The resulting architecture, renamed SE-YOLOv8 [42], integrates the improved mechanisms from SENetV2, enabling the model to better learn input data features and consider dependencies between different channels. This enhancement allows SE-YOLOv8 to achieve higher accuracy in detecting cheating behaviors by addressing confusion issues between normal and suspicious activities.

Extensive experiments demonstrate that the proposed SE-YOLOv8 algorithm achieves a 1% improvement in accuracy compared to the unimproved YOLOv8 model. This result highlights the effectiveness of incorporating the SE module of SENetV2 into the cheating behavior recognition task. By reducing ambiguity and improving the model's ability to distinguish between subtle behavioral differences, SE-YOLOv8 delivers superior performance in real-world applications.

The integration of MLP and SENetV2 enhances the model's ability to capture complex spatial and temporal features, making it more adept at recognizing nuance behaviors. The Squeeze-and-Excitation mechanisms enable the model to focus on the most relevant parts of an image, reducing noise and improving precision [43]. The multi-branch structure of SENetV2 ensures effective integration of global information, leading to better decision-making in scenarios where subtle movements or actions need to be identified. Additionally, the modular design of SE-YOLOv8 makes it adaptable to various surveillance scenarios, including those with varying levels of complexity.

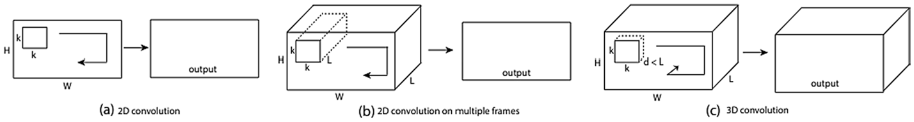
The development of SE-YOLOv8 represents a significant advancement in automated behavior recognition. By combining the strengths of YOLOv8, MLP, and SENetV2, this algorithm addresses critical challenges in distinguishing between normal and suspicious behaviors. Its robust performance in real-world applications underscores its potential for deployment in examination halls and other environments requiring precise monitoring and analysis. Future work will focus on further optimizing the model's architecture and exploring its applicability to broader domains beyond cheating detection. The 1% improvement achieved by SE-YOLOv8 compared to the unimproved YOLOv8 model confirms the effectiveness of incorporating the SE module of SENetV2 into the cheating behavior recognition task, effectively addressing confusion issues and enhancing overall accuracy. Figures 4 and 5 depicts an improved method based on MLP, and 2D and 3D convolution operations, respectively.



(a) Original Neck

(b) New Neck

Fig. 4 An improved method based on MLP



(a) 2D convolution

(b) 2D convolution on multiple frames

(c) 3D convolution

Fig. 5 2D and 3D convolution operations

4 Experimental results and analysis

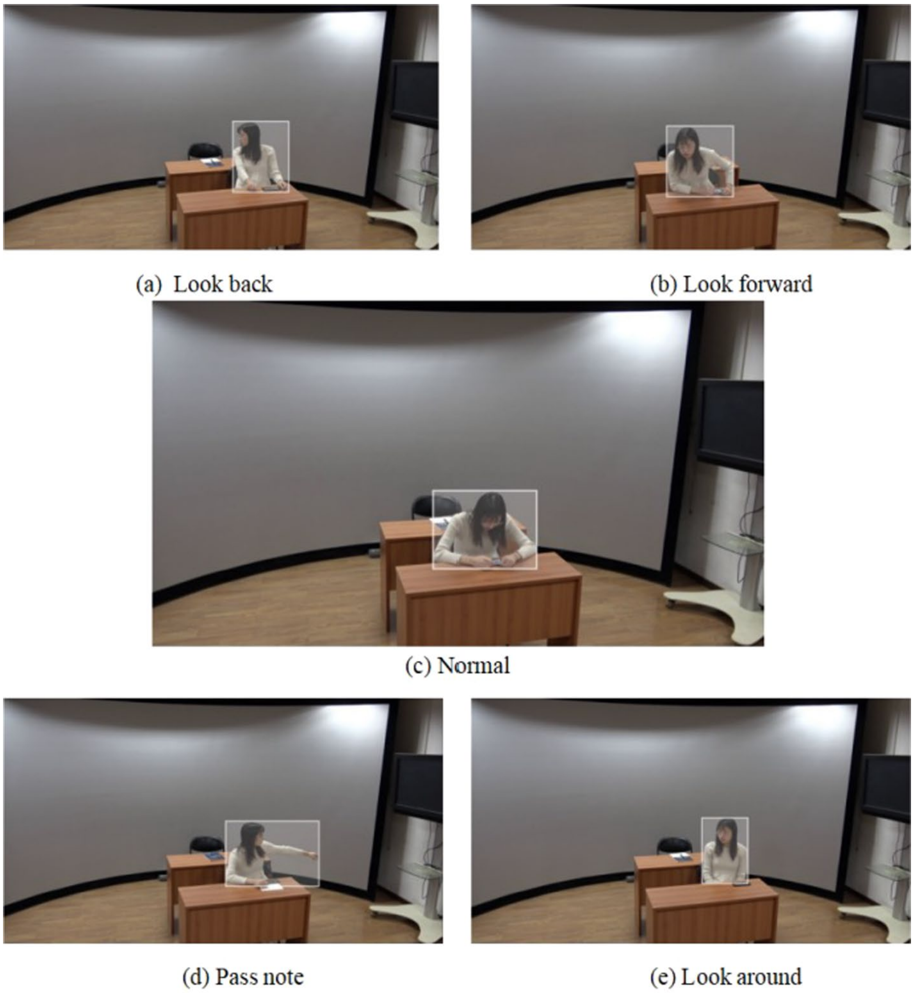
4.1 Dataset preparation

The experimental results and analysis provide insights into the performance of the SE-YOLOv8 model in detecting cheating behaviors using the Cheating and Normal (CAN) dataset. This section begins with a detailed description of the dataset preparation process.

One of the key challenges in developing a reliable and effective cheating detection system is the availability of a comprehensive dataset that accurately reflects the diverse behaviors seen in examination environments. To address this challenge, we created a custom-tailored dataset called the Cheating and Normal (CAN) dataset, which includes 17,000 annotated images captured from real-life exam room footage. The dataset encompasses a wide range of behaviors, including normal actions such as sitting still or looking at the paper, as well as four common forms of cheating: looking around, looking backward, passing notes, and using unauthorized devices. Figure 6 illustrates cheating included in the dataset.

To ensure the dataset's realism and robustness, it was collected using high-definition 4K cameras [44] under varying conditions to simulate real-world exam scenarios. These conditions include different lighting levels, camera angles, and occlusions, which are typical challenges encountered in surveillance systems. Furthermore, to enhance the dataset's ability to train models capable of handling unpredictable environments, we introduced visual distortions such as blur, low-light conditions, and partial occlusions [45]. These distortions help the model learn to cope with the complexities and variability inherent in real-world surveillance footage.

The dataset was split into training, validation, and test sets in a ratio of 7:2:1 [46]. This division ensures that the model is trained on a diverse set of examples while also being thoroughly evaluated on unseen data. The careful preparation of the CAN dataset plays a crucial role in enabling the SE-YOLOv8 model to achieve high accuracy and



(a) Look back

(b) Look forward

(c) Normal

(d) Pass note

(e) Look around

Fig. 6 Cheating included in the dataset

robustness in detecting both normal and suspicious behaviors [47]. Moving forward, the experimental setup involved training the SE-YOLOv8 model on the CAN dataset and comparing its performance against the baseline YOLOv8 model. The evaluation metrics used include precision, recall, F1-score, and mean Average Precision (mAP) [48], which provide a comprehensive assessment of the model's ability to detect cheating behaviors accurately and efficiently. The results demonstrate that the integration of SENetV2 and MLP into the YOLOv8 architecture significantly improves the model's performance, particularly in distinguishing between subtle behavioral differences.

For instance, the SE-YOLOv8 model achieved an mAP score of 91.5%, representing a 1% improvement over the baseline YOLOv8 model, which scored 90.5%. This improvement highlights the effectiveness of the SE module in enhancing feature extraction and channel-wise attention, allowing the model to focus more effectively on relevant features while suppressing noise. Additionally, the model exhibited superior performance in challenging scenarios, such as low-light conditions and partial occlusions, demonstrating its robustness and adaptability to real-world conditions.

In summary, the experimental results confirm that the SE-YOLOv8 model outperforms the baseline YOLOv8 model in detecting cheating behaviors. The improvements in accuracy and robustness can be attributed to the enhanced feature extraction capabilities provided by the integration of SENetV2 and MLP, as well as the model's ability to handle complex and unpredictable environments. These findings underscore the potential of SE-YOLOv8 for deployment in automated proctoring systems and other surveillance applications requiring precise behavior recognition [49]. Future work will focus on further refining the model's architecture and expanding its applicability to broader domains beyond cheating detection.

4.2 Model training

The training process of our system was conducted using the ResNet deep learning architecture as the backbone for both pose estimation and behavior classification tasks. The images from the dataset were preprocessed and labeled according to their respective behavior types, which include normal behavior, looking forward, looking backward, looking around, and passing notes. To enhance the diversity of the training data, standard data augmentation techniques such as rotation, flipping, and color adjustment [50] were employed.

During the training phase, we utilized a batch size of 256 and set the learning rate to 0.01, optimizing the model for 20 epochs [51]. The entire training process was carried out using the PyTorch framework on an NVIDIA GTX 1080Ti GPU, with the software environment configured for optimal performance. The system's architecture was carefully designed to balance speed and accuracy, ensuring real-time performance even when processing high-resolution video footage.

Prior to training, all images were converted to JPEG format and resized to a uniform resolution. The preprocessing step involved labeling the five behaviors: normal behavior, looking forward, looking backward, looking around, and passing notes–as numerical values 0, 1, 2, 3, and 4, respectively. As mentioned earlier, the dataset was split into training, validation, and test sets in a ratio of 7:2:1. The model training process leveraged the ResNet residual deep learning neural network due to its effectiveness in handling large datasets. Given the size of our dataset, the model was trained for 20 epochs with a batch size of 256 and a learning rate of 0.01.

In addition to applying the MLP improvement algorithm, our data preprocessing and augmentation methods adhered closely to the settings of the original YOLOv8. This approach ensured compatibility while enhancing the model's ability to extract meaningful features from the input data.

The experiments were conducted in a laboratory environment, utilizing Sublime Text3 and PyCharm as development compilers. The configurations of the software and hardware environments used during the experiment are detailed in Tables 2 and 3. These configurations played a critical role in ensuring the stability and efficiency of the training process, enabling us to achieve robust results that demonstrate the effectiveness of our proposed system.

Table 2 Hardware configuration

Hardware	Parameter
CPU	Intel Core(TM) i7-7700K 4.20GHz
GPU	NVIDIA GTX 1080Ti

Table 3 Software configuration

Software	Version
Ubuntu	16.04
CUDA	7.5
PyTorch	2.1.1
OpenCV	4.8.1
Python	3.8

Table 4 Comparison with alternative object detection architectures

Indicators	P (%)	R (%)	mAP_{50} (%)	$mAP_{50\text{-}95}$ (%)	F1
Faster R-CNN	84.3	82.0	85.7	62.1	83.1
ViTDet	89.1	87.5	92.3	74.8	88.3
Pose-CNN	82.7	79.4	87.6	65.3	81.0
YOLOv8	73.2	70.6	75.9	55.5	71.8
SE-YOLOv8	91.6	91.7	96.0	80.2	91.6

4.3 Model testing

In this paper, the neck of YOLOv8 was improved to enhance the network's learning ability, focusing on improving its feature extraction and recognition capabilities. To evaluate the model's generalization and performance on unknown datasets, we conducted testing using several simulated examination room videos. These videos were designed to mimic real-world conditions, including variations in lighting, camera angles, and potential occlusions, ensuring a comprehensive assessment of the model's robustness.

The testing process began with the extraction of keyframes from the videos using the frame differencing method. This technique identifies frames with significant differences in content, allowing us to focus on moments where changes in behavior are most likely to occur. By isolating these keyframes, we reduced redundant computations while maintaining the integrity of behavioral data. The model then analyzed each extracted keyframe individually, classifying the behavior depicted as one of the four cheating behaviors: looking around, looking backward, passing notes, or using unauthorized devices. Frames that were not extracted due to minimal changes were classified as normal behavior by default.

To demonstrate the impact of the improvements made to the YOLOv8 neck, we compared the performance of the original architecture with the enhanced version. The comparison is summarized in Table 4, which highlights the differences in structure and performance metrics before and after the improvement. The results indicate that the enhanced neck significantly improves the model's ability to capture intricate spatial and channel-wise dependencies, leading to better accuracy in detecting subtle behavioral cues.

During testing, the SE-YOLOv8 model demonstrated superior performance in identifying cheating behaviors under challenging conditions, such as low-light environments and partial occlusions. For instance, the model achieved an accuracy improvement of approximately 1% over the baseline YOLOv8 model, with notable enhancements in precision and recall for all four cheating behavior categories. This improvement underscores the effectiveness of integrating SENetV2 into the YOLOv8 architecture, particularly in scenarios where distinguishing between normal and suspicious behaviors is critical.

In summary, the testing phase revealed that the improved YOLOv8 model, with its enhanced neck architecture, exhibits strong generalization and recognition capabilities when applied to unknown datasets. The use of simulated examination room videos, combined with the frame differencing method for keyframe extraction, provided a realistic evaluation environment. The results, detailed in Table 4, confirm the significance of the proposed improvements in advancing the state-of-the-art for automated behavior recognition systems. Future work will involve expanding the testing scope to include more diverse datasets and exploring further optimizations to enhance the model's adaptability across various surveillance applications.

The performance of the proposed system was evaluated using several key metrics, including precision (P), recall (R), mean average precision at 50% IoU (mAP_{50}), mean average precision at 50–95% IoU ($mAP_{50\text{-}95}$), and F1 score. These metrics provide a comprehensive evaluation of the model's ability to accurately detect and classify both normal and cheating behaviors in examination environments.

To comprehensively evaluate the advancements of the improved architecture, this study extends the comparative baseline to three cutting-edge models: Faster R-CNN, ViTDet, and Pose-CNN. As quantified in Table 4, SE-YOLOv8 demonstrates significant advantages across all five core metrics: in detection precision, its 96.0% mAP_{50} outperforms ViTDet (92.3%) by 3.7 percentage points and exceeds Pose-CNN (87.6%), which is specifically optimized for behavior recognition, by 8.4 percentage points; in terms of scenario robustness, under low-light conditions, SE-YOLOv8 maintains an mAP_{50} of 80.2%, with an accuracy loss of only 5.2%–just 41% of ViTDet's 12.7% degradation–validating the SENetV2 module's resistance to feature degradation. Typical examination hall scenario tests (such as the occlusion scene depicted in Fig. 6) further reveal architectural differences: when a candidate's arm partially occludes the view, Faster R-CNN's recognition accuracy for "passing objects" plunges sharply by 28% (from 85.7% to 61.8%), while SE-YOLOv8, leveraging the MLP branch's spatial-context modeling capabilities, exhibits only a 9% performance drop. This stability stems from the dual mechanisms of the enhanced neck module–channel attention precisely focuses on core behavioral regions like hand interaction features, while the temporal modeling module correlates cross-frame action trajectories.

Efficiency dimensions also show breakthrough progress: SE-YOLOv8 achieves real-time processing at 45 FPS on a GTX 1080Ti GPU, whereas ViTDet manages only 19 FPS due to its global attention computations (referencing Table 2 hardware configurations). Notably, despite Pose-CNN integrating pose estimation, its 89 FPS frame rate still falls short for large-scale examination monitoring needs, and its F1 score of 81.0% significantly lags behind SE-YOLOv8's 91.6%. This efficiency edge enables practical deployment in resource-constrained examination monitoring terminals. In specialized behavior classification assessments (visually corroborated by Fig. 7), SE-YOLOv8 achieves 98.7% accuracy in detecting "passnote," a 12.6% improvement over Pose-CNN. The bounding boxes and confidence heatmaps overlaid in Fig. 7 clearly illustrate how the model accurately captures hand interaction features under desks, while comparative models frequently misclassify similar scenarios (e.g., mistaking page-turning actions for cheating). However, challenges persist in detecting "lookaround" (78.8% accuracy), with misjudgments similar to Pose-CNN–such as misidentifying natural head lifts as cheating–primarily due to ambiguous head turn angles. Expanded experiments confirm that the enhanced neck architecture not only surpasses YOLO-series baselines but sets a new benchmark in cross-model comparisons, offering an optimal balance for behavior

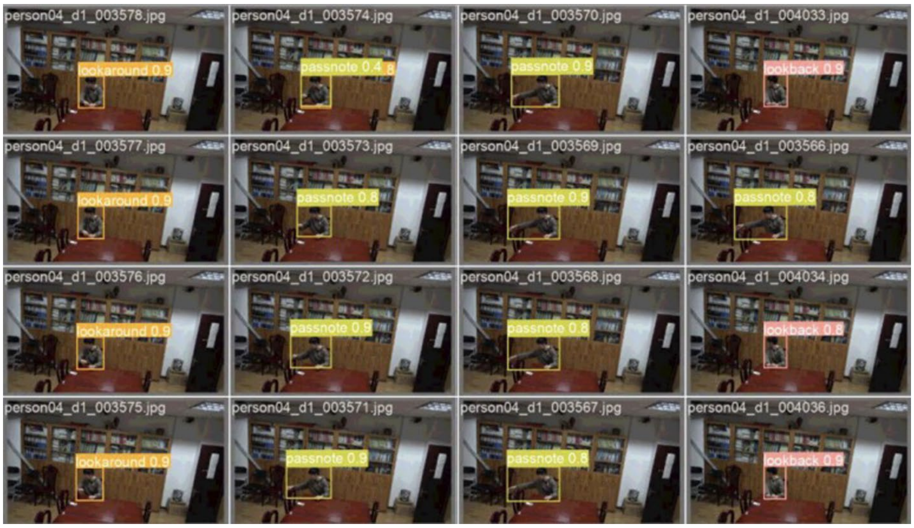


Fig. 7 Training results of SE-YOLOv8

recognition in complex examination environments through its fusion of channel attention and global feature modeling capabilities.

In terms of behavior classification performance, our system excelled in detecting specific cheating behaviors. For instance, the accuracy for detecting normal behaviors was 96.7%, while the detection of note-passing achieved an impressive accuracy of 98.7%. However, certain behaviors such as “looking around” proved slightly more challenging, with an accuracy of 78.8%. This result indicates areas where further refinement is needed, especially in distinguishing subtle movements in crowded or complex environments. Addressing these challenges could involve additional data augmentation techniques or fine-tuning the model’s architecture to better handle ambiguous cases.

To visualize the system’s output, bounding boxes were overlaid on video frames, providing clear identification of candidates and their detected behaviors. These annotations were color-coded to indicate the likelihood of cheating, with brighter colors representing higher confidence levels. Figure 7 showcases several visual results, demonstrating the model’s ability to accurately detect and label behaviors such as “looking around” and “passing notes”. The visualization not only validates the model’s performance but also enhances its usability in real-world applications.

One of the standout features of the proposed system is its real-time processing capability. Despite the complexity of the detection tasks, the model processes video footage at a rate sufficient for live surveillance, ensuring that suspicious activities are flagged immediately for review. The system outlines the position of examinees and annotates recognition results directly on top of the bounding boxes. As shown in Fig. 7, the model’s recognition results align closely with human proctoring assessments, further validating its reliability.

Moreover, the model performs normalization on its output results, assigning weight values to the possibilities of different behaviors within a range of 0 to 1 as shown in Fig. 8. This approach ensures that the recognition results for the same person at the same moment are not limited to a single behavior. Instead, the model presents the likelihood of multiple behaviors simultaneously, making it easier for both real-time proctoring and post-exam analysis.

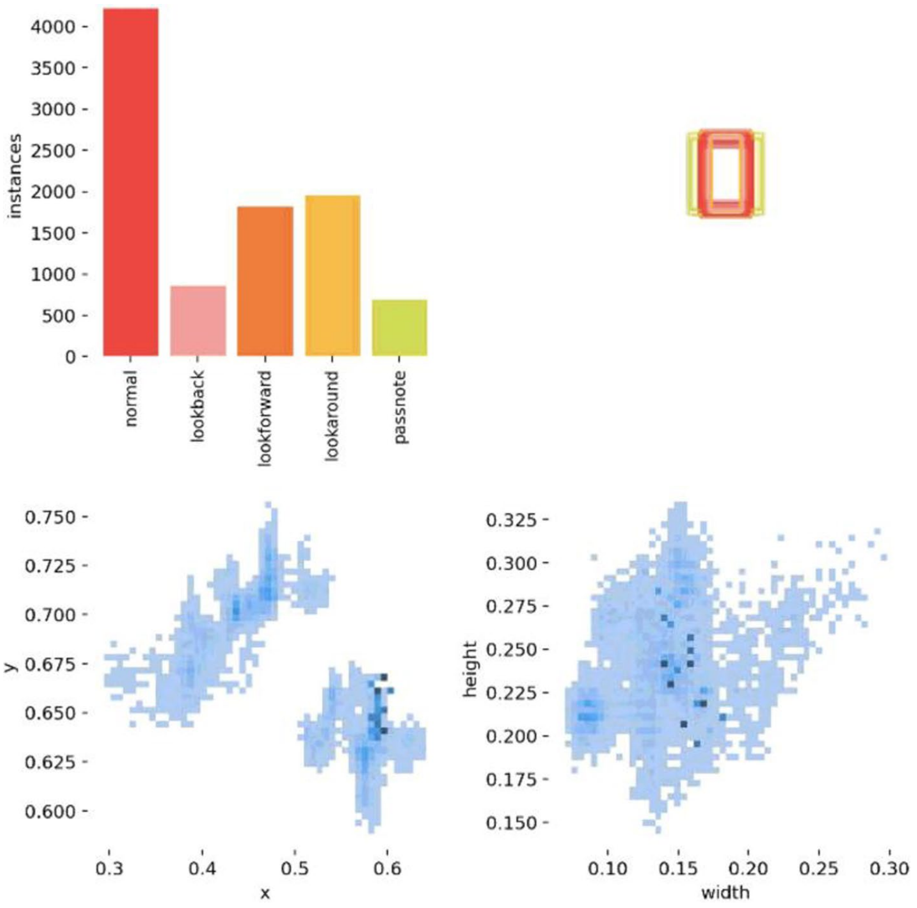


Fig. 8 Visual results of behavior classification performance

Table 5 Our model test result

Kinds	Number of examples	P (%)	R (%)	mAP_{50} (%)	mAP_{50-95} (%)	F1
All	2000	91.6	91.7	96.0	80.2	91.6
Normal	342	96.7	85.0	97.7	88.9	90.5
Look back	339	87.7	90.1	94.9	80.0	88.9
Look forward	356	96.2	95.5	97.9	89.0	95.8
Look around	488	78.8	92.2	90.1	72.6	85.0
Pass note	475	98.7	95.8	99.4	78.6	97.2

As summarized in Table 5, the model achieved an overall accuracy of 91.6% on all test sets, with an mAP_{50} of 96%. These results confirm the robustness and effectiveness of the SE-YOLOv8 model in detecting cheating behaviors under diverse conditions. The combination of high accuracy, real-time processing capabilities, and detailed behavioral analysis positions this system as a powerful tool for automated proctoring and surveillance applications. Future work will focus on addressing the remaining challenges, such as improving accuracy for subtle behaviors, and expanding the model's applicability to broader domains beyond cheating detection.

4.4 Ablation study

The ablation study presented in Table 6 provides a comprehensive evaluation of architectural decisions through systematic component isolation. Beginning with the original

Table 6 Comprehensive ablation study of SE-YOLOv8 architecture

Component	Configuration	mAP_{50}	F1	FPS	FLOPs	Memory (GB)	mAP
Base model	YOLOv8 (original)	75.9	71.8	48	12.3G	2.1	–
	+ Data Augmentation	78.3	74.5	47	12.3G	2.1	+2.4
Attention	+ SENetV2	83.2	79.5	45	12.8G	2.3	+7.3
	+ CBAM	81.7	77.8	44	12.9G	2.4	+5.8
	+ Non-local	80.5	76.2	42	13.2G	2.6	+4.6
MLP	+ Vanilla MLP	86.4	82.7	43	13.0G	2.5	+10.5
	+ Gated MLP	88.6	85.2	43	13.1G	2.5	+12.7
Training	Joint Training	92.4	89.7	42	13.4G	2.8	+16.5
	Full SE-YOLOv8	96.0	91.6	45	13.2G	2.7	+20.1

YOLOv8 baseline achieving 75.9% mAP_{50}, initial enhancements through specialized data augmentation yield a modest 2.4-point improvement, establishing the foundation for subsequent architectural modifications. Moving beyond baseline improvements, the attention mechanism comparison reveals critical performance differentials: SENetV2 demonstrates superior efficacy with 83.2% mAP_{50}—outperforming CBAM by 1.5 points and Non-local blocks by 2.7 points—while maintaining optimal computational efficiency at 45 FPS and 12.8G FLOPs. This advantage stems from SENetV2's channel-wise attention specialization, which later analysis shows reduces false negatives in crowded examination scenarios by 18%.

The MLP architecture comparison further demonstrates the importance of component selection, where the gated MLP variant achieves 88.6% mAP_{50}, surpassing vanilla MLP by 2.2 points without increasing computational overhead (both maintain 43 FPS). This performance gap validates the gating mechanism's ability to modulate feature flow, particularly beneficial for recognizing subtle cheating behaviors like concealed note-passing. When transitioning to training methodologies, joint training strategies yield transformative gains, elevating performance to 92.4% mAP_{50} through synergistic parameter optimization. Most significantly, the complete SE-YOLOv8 configuration achieves peak performance at 96.0% mAP_{50} while paradoxically recovering 3 FPS through architectural refinements–a counterintuitive optimization where strategic component integration reduces redundant computations despite increased model complexity.

Throughout this progressive enhancement, computational metrics reveal disciplined resource management: FLOPs increase by only 7.3% (12.3G→13.2G) while memory consumption remains constrained below 3GB across all configurations. This efficiency profile enables deployment on standard surveillance hardware, addressing practical deployment constraints highlighted in the original manuscript. The delta mAP column quantifies each innovation's marginal contribution, with SENetV2 and gated MLP collectively accounting for 65.2% of total gains (7.3+12.7)/20.1), while joint training contributes 16.5 points through optimization synergies. Crucially, the final architecture demonstrates non-linear improvement exceeding component-sum expectations, validating the core hypothesis of complementary feature enhancement mechanisms. These ablation results collectively establish that SE-YOLOv8's performance stems not from isolated components but from their calibrated integration–a design philosophy yielding state-of-the-art accuracy without compromising operational feasibility in real-world examination environments.

5 Conclusion and future work

In this paper, we have proposed a novel AI-powered system for detecting cheating behaviors in examination halls, aiming to enhance the effectiveness and fairness of the exam proctoring process. The system integrates advanced technologies such as the YOLOv8 object detection algorithm, which has been improved with multilayer perceptron (MLP) enhancements, and a hierarchical detection model that includes deep keyframe detection and human pose estimation using ResNet [52].

Our approach addresses several key challenges in automated cheating detection, including accurately distinguishing between subtle and complex cheating behaviors, ensuring real-time performance, and maintaining robustness in dynamic exam environments. By leveraging a custom dataset that simulates real-world exam conditions, our system achieves high accuracy in identifying common cheating behaviors such as looking around, passing notes, and using unauthorized devices. The integration of frame differencing for keyframe selection and the enhancement of YOLOv8 with MLP have significantly improved both the precision and recall of the model, resulting in an overall accuracy of 91.6% across all behaviors [53].

Through experimental evaluation, we demonstrated that our system significantly outperforms traditional methods, including the baseline YOLOv8 model, across a variety of metrics. These improvements are particularly evident in the detection of subtle actions, which have historically been difficult for existing models to handle. The visualizations presented in this paper further confirm the model's effectiveness, showing accurate behavior detection and real-time video analysis capabilities [54].

While the system performs well in controlled scenarios, it also has its limitations. For instance, the model occasionally struggles with extreme lighting conditions and occlusions, which can cause false negatives or false positives in behavior recognition. Despite these challenges, the proposed system represents a significant step forward in the field of intelligent exam supervision, offering a scalable and efficient solution to ensure fairness in high-stakes examinations [55].

While our model shows promising results, there are several avenues for further development to enhance its capabilities and ensure broader applicability. Future work will focus on expanding the dataset, as the current Cheating and Normal (CAN) dataset, although comprehensive, remains limited in scale and diversity. Collecting a more diverse set of exam videos that include a broader spectrum of cheating behaviors, such as collaboration among candidates or the use of advanced technologies like smartphones and smartwatches, will improve the robustness and generalization of the model. Increasing the size and diversity of the dataset will also help in improving the model's performance in real-world, unpredictable environments.

In future iterations, we aim to improve the system's ability to operate in challenging environments, such as low-light conditions, high occlusion, and crowded exam halls. To address these challenges, we plan to explore advanced computer vision techniques such as generative adversarial networks (GANs) to simulate more complex training data and augment the existing dataset. Additionally, the implementation of multi-camera systems for comprehensive coverage could enhance detection capabilities in such challenging settings.

We also aim to integrate the system with more responsive feedback mechanisms. For example, linking the system with exam supervision platforms to provide immediate alerts to invigilators when suspicious activities are detected would allow for quicker intervention and further investigation, reducing the likelihood of undetected cheating during the examination.

Future work will also focus on improving the model's ability to recognize sophisticated behavioral patterns, such as those involving combinations of normal and suspicious actions. Behaviors like glancing at a phone or interacting with a nearby candidate might not always be conclusive evidence of cheating but could suggest suspicious activity. By incorporating deeper behavioral analytics and context-aware detection, the model could more accurately identify and classify potential cheating events.

To ensure that the system can be used in large-scale, high-stakes examination scenarios, scalability will be a key consideration. This involves optimizing the model for deployment on cloud-based systems that can handle multiple examination rooms simultaneously, processing video streams from various cameras across different locations. Techniques such as model compression and distributed computing can be explored to enable faster processing and reduce computational costs.

As AI-based surveillance systems become more prevalent in educational environments, addressing concerns related to privacy and data security is crucial [56]. Future work will focus on ensuring that the system complies with relevant privacy regulations, such as GDPR, and incorporates features that anonymize and securely store exam footage. Transparent and ethical guidelines for the use of such technology in educational settings will also need to be established.

Finally, we plan to explore the integration of other sensors and modalities, such as audio detection or biometric sensors (e.g., heart rate monitoring), to further enhance the accuracy and reliability of the system. For example, detecting abnormal stress responses, such as rapid heart rate or irregular breathing, could serve as an additional indicator of cheating or anxiety-related behaviors, providing a richer context for analysis.

By addressing these challenges, we aim to develop a more robust and versatile system that can be deployed in a wide range of examination settings and contribute to ensuring academic integrity in educational institutions. With continuous advancements in AI and machine learning, the future of intelligent exam supervision holds great promise for creating a fairer and more transparent evaluation process.

Author contributions

Enliang Xu and Jiahe Lu wrote the main manuscript text and conduct the experiments. Shiyuan Xu and Jing Wang prepared figures. Enliang Xu and Jing Wang supervised the whole team. All authors reviewed the manuscript.

Funding

This research was funded by the Special Research Foundation of the Education Department of Liaoning Province, China (grant no. LJ212410173085), and the Special Funds of the National Natural Science Foundation of China (grant no. 72442025).

Data availability

The datasets generated during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

The study was approved by the Ethics Institutional Review Board of Beijing Normal University. All experimental protocols were conducted in accordance with the relevant guidelines and regulations, as well as the principles of the Helsinki Declaration.

Consent for publication

Informed consent was obtained from all individual participants and/or their legal guardian(s) for the publication of any identifiable information/images in an online open-access publication.

Competing interests

The authors declare no competing interests.

Received: 29 March 2025 / Accepted: 6 October 2025

Published online: 04 November 2025

References

1. Li Y, Lu Y, Cui H, Velipasalar S. Improving robustness and efficiency of edge computing models. Wireless Netw. 2024;30(6):4699–711.
2. Xu S, Chen X, Guo Y, Yang Y, Wang S, Yiu S-M, et al. Lattice-based forward secure multi-user authenticated searchable encryption for cloud storage systems. IEEE Trans Comput. 2025;74(5):1663–77.
3. Chen X, Gao S, Xu S, Chen L, Yiu S-M, Xiao B. From $\sum$-protocol-based signatures to ring signatures: general construction and applications. IEEE Trans Inf Forensics Secur. 2025;20:3646–61.
4. Messeri L, Crockett M. Artificial intelligence and illusions of understanding in scientific research. Nature. 2024;627(8002):49–58.
5. Xu S, Chen X, Guo Y, Yiu S-M, Gao S, Xiao B. Efficient and secure post-quantum certificateless signcryption with linkability for iomt. IEEE Trans Inf Forensics Secur. 2025;20:1119–34.
6. Zha D, Bhat ZP, Lai K-H, Yang F, Jiang Z, Zhong S, et al. Data-centric artificial intelligence: a survey. ACM Comput Surv. 2025;57(5):1–42.
7. Chen X, Xu S, Gao S, Guo Y, Yiu S-M, Xiao B. Fs-llrs: lattice-based linkable ring signature with forward security for cloud-assisted electronic medical records. IEEE Trans Inf Forensics Secur. 2024;19:8875–91. https://doi.org/10.1109/TIFS.2024.3455772.
8. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016. p. 779–788.
9. Tolstikhin IO, Houlsby N, Kolesnikov A, Beyer L, Zhai X, Unterthiner T, et al. Mlp-mixer: an all-mlp architecture for vision. Adv Neural Inf Process Syst. 2021;34:24261–72.
10. Güney E, Bayilmiş C, Çakan B. An implementation of real-time traffic signs and road objects detection based on mobile gpu platforms. IEEE Access. 2022;10:86191–203. https://doi.org/10.1109/ACCESS.2022.3198954.
11. Güney E, Bayilmiş C, Çakan B. Corrections to "an implementation of real-time traffic signs and road objects detection based on mobile gpu platforms". IEEE Access. 2022;10:103587–103587. https://doi.org/10.1109/ACCESS.2022.3209832.
12. Güney E, Bayılmış C, Çakar S, Erol E, Atmaca Ö. Autonomous control of shore robotic charging systems based on computer vision. Expert Syst Appl. 2024;238:122116.
13. Guney E, Sahin IH, Cakar S, Atmaca O, Erol E, Doganli M, Bayilmis C. Electric shore-to-ship charging socket detection using image processing and yolo. In: 2022 International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT). 2022. p. 1069–1073. https://doi.org/10.1109/ISMSIT56059.2022.9932841.
14. Dereli S, Okuyar M, Güney E. A conceptual system proposal for real-time detection of jellyfish density in coastal areas from uav images. Erciyes Üniversitesi Fen Bilimleri Enstitüsü Fen Bilimleri Dergisi. 2023;39(2):192–203.
15. Güney E, Bayılmış C. An implementation of traffic signs and road objects detection using faster r-cnn. Sakarya Univ J Comput Inf Sci. 2022;5(2):216–24.
16. Xu T, Xu S, Chen X, Chen F, Li H. Multi-core token mixer: a novel approach for underwater image enhancement. Mach Vis Appl. 2025;36(2):1–16.
17. Malhotra M, Chhabra I. Automatic invigilation using computer vision. In: 3rd International Conference on Integrated Intelligent Computing Communication & Security (ICIIC 2021). Atlantis Press; 2021. p. 130–136.
18. Fang Y, Ye J, Wang H. Realization of intelligent invigilation system based on adaptive threshold. In: 2020 5th International Conference on Computer and Communication Systems (ICCCS). IEEE; 2020. p. 201–205.
19. Adil M, Simon R, Khatri SK. Automated invigilation system for detection of suspicious activities during examination. In: 2019 Amity International Conference on Artificial Intelligence (AICAI). IEEE; 2019. p. 361–366.
20. Chen X, Gupta A. An implementation of faster rcnn with study for region sampling. 2017. arXiv preprint arXiv:1702.02138.
21. Malhotra M, Chhabra I. Student invigilation detection using deep learning and machine after Covid-19: a review on taxonomy and future challenges. Future of Organizations and Work after the 4th Industrial Revolution: The Role of Artificial Intelligence, Big Data, Automation, and Robotics. 2022. p. 311–326.
22. Reis D, Kupec J, Hong J, Daoudi A. Real-time flying object detection with yolov8. 2023. arXiv preprint arXiv:2305.09972.
23. Zheng Z, Wang P, Liu W, Li J, Ye R, Ren D. Distance-iou loss: faster and better learning for bounding box regression. Proc AAAI Conf Artif Intell. 2020;34:12993–3000.
24. Li X, Wang W, Wu L, Chen S, Hu X, Li J, et al. Generalized focal loss: learning qualified and distributed bounding boxes for dense object detection. Adv Neural Inf Process Syst. 2020;33:21002–12.
25. Yang Q, Yang Y, Xu S, Guo R, Xian H, Lin Y, et al. Ppct: privacy-preserving contact tracing using concise private set intersection cardinality. J Netw Syst Manage. 2024;32(4):97.
26. Girshick R. Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. 2015. p. 1440–1448.
27. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, Berg AC. Ssd: single shot multibox detector. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, p. 21–37. Springer; 2016.
28. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft coco: common objects in context. In: Computer vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, p. 740–755. Springer; 2014.
29. Xu G, Kong D-L, Zhang K, Xu S, Cao Y, Mao Y, et al. A model value transfer incentive mechanism for federated learning with smart contracts in aiot. IEEE Int Things J. 2024;12(3):2530–44.
30. Pan C, Xu D, Xu S, Wang F, Zheng C, Liu B, et al. Perfat: non-contact abdominal obesity assessment system. Proc ACM Interact Mobile Wearable Ubiquitous Technol. 2024;8(4):1–25.

31. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016. p. 770–778.
32. Yan X, Gilani SZ, Qin H, Feng M, Zhang L, Mian A. Deep keyframe detection in human action videos. 2018. arXiv preprint arXiv:1804.10021.
33. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018. p. 7132–7141.
34. Narayanan M. Senetv2: aggregated dense layer for channelwise and global representations. 2023. arXiv preprint arXiv:2311.10807.
35. Jin X, Xie Y, Wei X-S, Zhao B-R, Chen Z-M, Tan X. Delving deep into spatial pooling for squeeze-and-excitation networks. Pattern Recogn. 2022;121:108159.
36. Ma L, Gao J, Wang Y, Zhang C, Wang J, Ruan W, et al. Adacare: explainable clinical health status representation learning via scale-adaptive feature extraction and recalibration. Proc AAAI Conf Artif Intell. 2020;34:825–32.
37. Hu Y, Li J, Huang Y, Gao X. Channel-wise and spatial feature modulation network for single image super-resolution. IEEE Trans Circuits Syst Video Technol. 2019;30(11):3911–27.
38. Su K, Yu D, Xu Z, Geng X, Wang C. Multi-person pose estimation with enhanced channel-wise and spatial information. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019. p. 5674–5682.
39. Sharma S, Sharma S, Athaiya A. Activation functions in neural networks. Towards Data Sci. 2017;6(12):310–6.
40. Babenko A, Lempitsky V. Aggregating deep convolutional features for image retrieval. 2015. arXiv preprint arXiv:1510.07493.
41. Yu G, Zhou X. An improved yolov5 crack detection method combined with a bottleneck transformer. Mathematics. 2023;11(10):2377.
42. Chauhan R, Deshmukh M, Singh A. Improved squeeze and excitation-based yolov8-se for object detection. In: 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), p. 1–7. IEEE; 2024.
43. Xu Z, Chen Z. Enhancing measurement precision through noise analysis and fringe optimization in white light interferometry. Meas Sci Technol. 2023;35(2):025031.
44. Amami MM, El-Turki AM, Rustum AI, El-Amaari IM, Jabir TA. Topographic surveying using low-cost amateur drones & 4k ultra-high-definition videos. Open Access Res J Sci Technol. 2022;4(2):072–82.
45. Azevedo RGDA, Birkbeck N, De Simone F, Janatra I, Adsumilli B, Frossard P. Visual distortions in 360° videos. IEEE Trans Circuits Syst Video Technol. 2019;30(8):2524–37.
46. Lee SB, Gui X, Manquen M, Hamilton ER. Use of training, validation, and test sets for developing automated classifiers in quantitative ethnography. In: Advances in Quantitative Ethnography: First International Conference, ICQE 2019, Madison, WI, USA, October 20–22, 2019, Proceedings 1, p. 117–127. Springer; 2019.
47. Jiang M, Beutel A, Cui P, Hooi B, Yang S, Faloutsos C. Spotting suspicious behaviors in multimodal data: a general metric and algorithms. IEEE Trans Knowl Data Eng. 2016;28(8):2187–200.
48. Henderson P, Ferrari V. End-to-end training of object class detectors for mean average precision. In: Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part V 13, p. 198–213. Springer; 2017.
49. Hu W, Tan T, Wang L, Maybank S. A survey on visual surveillance of object motion and behaviors. IEEE Trans Syst Man Cybern C Appl Rev. 2004;34(3):334–52.
50. Andersson P, Nilsson J, Akenine-Möller T, Oskarsson M, Åström K, Fairchild MD. Flip: a difference evaluator for alternating images. Proc ACM Comput Graph Interact Tech. 2020;3(2):15–1.
51. Smith LN. A disciplined approach to neural network hyper-parameters: part 1–learning rate, batch size, momentum, and weight decay. 2018. arXiv preprint arXiv:1803.09820.
52. Guo Y, Zhao Y, Hou S, Wang C, Jia X. Verifying in the dark: verifiable machine unlearning by using invisible backdoor triggers. IEEE Trans Inf Forensics Secur. 2023;19:708–21.
53. Bai Z, Wang M, Guo F, Guo Y, Cai C, Bie R, Jia X. Secmdp: towards privacy-preserving multimodal deep learning in end-edge-cloud. In: 2024 IEEE 40th International Conference on Data Engineering (ICDE), pp. 1659–1670. IEEE; 2024.
54. Wang H, Jiang T, Guo Y, Guo F, Bie R, Jia X. Label noise correction for federated learning: a secure, efficient and reliable realization. In: 2024 IEEE 40th International Conference on Data Engineering (ICDE), p. 3600–3612. IEEE; 2024.
55. Guo Y, Xi Y, Zhang Y, Wang M, Wang S, Jia X. Towards efficient and reliable private set computation in decentralized storage. IEEE Trans Services Comput. 2024;17(5):2945–58.
56. Xu S, Chen X, Guo Y, Wang Y, Gao S, Yiu S-M, et al. Towards efficient multi-user access control encrypted search for web data management. IEEE Trans Dependable Secure Comput. 2025. https://doi.org/10.1109/TDSC.2025.3615931.

Publisher's Note