Sensors and Materials, Vol. 38, No. 4 (2026) 2149–2168
MYU Tokyo

S & M 4428

Detecting Learning Behaviors Using Deep-learning-based Classroom Teaching Quality Analysis System

Chunming Liu,[1] Sijie Qiu,[2] and Chi-Hsin Yang[2*]

[1]School of Marxism, Sanming University, Sanming 365004, Fujian Province, China
[2]School of Mechanical and Electric Engineering, Sanming University, Sanming 365004, Fujian Province, China

(Received November 18, 2025; accepted April 14, 2026)

Keywords: YOLOv8 network, classroom behavior detection, classroom teaching quality analysis system, CTQA-YOLOv8-NET

The application of the deep-learning-based You Only Look Once (YOLO) network model for identifying teaching and learning behaviors in classrooms has attracted attention in recent studies. However, there is limited scholarly focus on using these models to evaluate instructional quality. In this study, we propose a system to analyze instructional quality through advanced deep learning technology, aiming to assess and enhance university classroom activities. It comprises two main components. First, the YOLOv8 architecture supports the classroom teaching quality analysis network (CTQA-YOLOv8-NET), which detects and categorizes student behaviors into six types during class. The second component is a visual platform for analyzing teaching quality based on data obtained through CTQA-YOLOv8-NET. By extracting one image frame per minute from a 45-minute course video, this method visually represents the distribution of student behaviors across categories and compiles data into trend graphs showing effective versus ineffective learning ratios over time. These visual tools provide scientific evidence for evaluating instructional effectiveness, allowing teachers to objectively assess outcomes and develop targeted improvement strategies in areas such as content adjustments and pedagogical innovations. In this study, these methods were specifically applied to the university general educational class, Ideology, Morality, and Rule of Law, suggesting optimization options while examining the strengths and weaknesses of the course, ultimately confirming the effectiveness of interventions in enhancing instructional quality.

1. Introduction

As computer technology has advanced, machine-vision-based object defect detection has gained popularity in real-world applications, such as identifying flaws in solar panels,(1) helmets,(2) and roads.(3) Deep-learning techniques for object detection have advanced significantly in recent years as a result of the quick development of computer network algorithms. Deep-learning object detection techniques are divided into two groups in the literature currently available by two-stage and single-stage approaches. Faster region-based convolutional neural

*Corresponding author: e-mail: 20190207@fjsmu.edu.cn
https://doi.org/10.18494/SAM6063

ISSN 0914-4935 © MYU K.K.

networks (F-RCNNs)[4] and enhanced F-RCNNs[5] are examples of two-stage approaches. The You Only Look Once (YOLO) series[6–16] and the single-shot multibox detector (SSD)[17,18] are the two main types of single-stage approaches.

Because of its quick and precise identification results, the YOLO series object detection technique has become widely used in many different sectors. An enhanced YOLOv8 network is suggested for real-time object identification in compliance with Ref. 6. Additionally, a number of improved YOLOv8 networks have been created for the purposes of detecting weeds in rice fields,[7] detecting metal surface defects,[8] identifying fish underwater,[9] and detecting smoking behavior in elderly people.[10] Furthermore, enhanced YOLOv10 and YOLOv11 networks are designed for detection under low-lighting conditions,[11] fire scenes,[12] welding faults,[13] small traffic signs,[14] and lightweight tomatoes.[15] In particular, in Ref. 16, a YOLOv11-based object detection system for deployment in maritime settings was presented. The work achieved robust performance in identifying ship categories and estimating heading angles.

Using several series of YOLO algorithms, we thoroughly investigated the problem of behavior recognition in the classroom. With a bidirectional weighted feature pyramid network, the use of an enhanced YOLOv5 network was suggested in Ref. 19 for an end-to-end single-stage classroom behavior detection algorithm. In Refs. 20 to 24, the use of multiple YOLOv7, YOLOv8, and YOLOv10-based networks with various modified submodules for monitoring classroom behavior was presented. The modified YOLOv7 network was created in Ref. 20 to teach behavior recognition in the classroom. The enhanced feature capture network based on YOLOv8 was introduced in Ref. 21 to address the existing networks' lack of detail feature capture capability. The YOLOv10-based classroom behavior detection model was created by Ma *et al.* to improve teaching strategies and raise the caliber of instruction in packed classrooms.[24]

However, the majority of methods used in earlier research[19–24] fail to take into account the spatial interactions that occur in the classroom, which can successfully connect local and global feature information. A low-complexity, high-accuracy dual-residual spatial interaction network for multiperson pose estimation (MPPE) is proposed in Ref. 25. The research area for detecting classroom behavior is expanded by this work. Additionally, in Ref. 26, multi-object behavior identification based on object detection cascaded picture classification was used to handle a large number of students and similarity of student behaviors.

From the results in the aforementioned references,[19–24] it is evident that most research efforts have concentrated on the accurate and efficient detections of classroom behaviors. However, the outcomes derived from such behavioral detection and evaluation remain underexplored, particularly in terms of their potential application to assessing improvements in teaching quality and formulating data-driven optimization strategies. For example, in Ref. 21, the researchers classified facial expressions and seated postures as behavioral indicators of positive versus negative learning statuses in a basic manner, but did not further apply them in real-world educational interventions.

In fact, within the context of enhancing university course instruction, improvement strategies proposed by researchers often lack robust empirical or scientific support without any data-driven approach. Consequently, these strategies are frequently criticized as being speculative, leading to inconclusive or vague academic discussions.

Motivated by previous research, we propose a deep-learning-based system for analyzing classroom teaching quality. The system comprises two main components. The first component is the classroom teaching quality analysis YOLOv8 network (CTQA-YOLOv8-NET), which was developed on the basis of the YOLOv8 architecture. This network is designed to detect and evaluate student behaviors during class sessions and classify them into six distinct behavioral categories.

The second component is a visualization platform built upon CTQA-YOLOv8-NET. It enables the display of statistical data derived from detecting and identifying one image frame per minute throughout a 45-minute video lecture. These statistics are then summarized and analyzed. As the lesson progresses, the system generates graphical representations illustrating the changes in the number of students exhibiting effective and ineffective learning behaviors. Referring to these trend curves, the system provides scientifically grounded data that can assist educators in evaluating the effectiveness of specific instructional activities and support initiatives aimed at improving course instruction, including enhancements to teaching content and pedagogical methods, as well as optimization strategies.

The novelty of this study lies in three key advancements

(1) Architectural Enhancement.
Under the base network, CTQA-YOLOv8-NET adds a dedicated detection branch optimized for fine-grained behavioral cues from sensed images. The network boosts localization and classification accuracy for subtle, low-resolution actions in crowded classrooms, and overcomes the poor sensitivity of standard detectors to small objects.

(2) Holistic Assessment Infrastructure
By applying cameras to obtain the images, we propose a unified pipeline integrating behavioral analysis, pedagogical evaluation, and interpretive visualization. The pipeline converts fragmented behavioral events into continuous, time-resolved quality metrics to replace subjective impressions with objective, data-driven teaching assessments.

(3) Practice-oriented Implementation Framework
In ideological and political education, we establish a self-sustaining improvement loop that is driven by sensed behavioral image data that are used to identify instructional gaps via pattern mining, designs evidence-informed, context-aware pedagogical adjustments, and evaluates their impact through longitudinal behavioral measurement. The loop makes instructional refinement measurable, transferable, and adaptable across diverse teaching contexts.

The primary contributions of this study are summarized as follows.

(1) Building upon the YOLOv8 network, the CTQA-YOLOv8-NET model has been developed to detect and assess student behaviors during classroom instruction, enabling classification and labeling across six distinct behavioral categories.

(2) A visual CTQA system based on CTQA-YOLOv8-NET has been established. This system generates statistical curves that illustrate changes in the number of students exhibiting effective versus ineffective learning behaviors over the course of a class. These dynamic data representations provide scientific evidence to support educators in formulating strategies for enhancing teaching quality.

(3) Using the university general educational course, named Ideology, Morality, and Rule of Law, as a case study, a comprehensive teaching evaluation was conducted. The advantages and limitations are analyzed and discussed in depth, leading to the proposal of targeted optimization strategies for improving the course's teaching quality. The effectiveness of these strategies in enhancing teaching outcomes is subsequently validated.

In addition, the developed technology can be modified and applied to other fields, for example, monitoring worker behavior in plants and factories, traffic behavior in heavily congested areas, or people's behavior in crowded situations.

The structure of this paper is as follows. Brief descriptions of the original YOLOv8 network and the architecture of CTQA-YOLOv8-NET are given in Sect. 2. We outline the experimental configuration and dataset tagging protocols for network training in Sect. 3. Section 4 contains a discussion of the training and efficiency verification of CTQA-YOLOv8-NET. The CTQA system is presented in Sect. 5. The teaching evaluation of course execution using the university general education course, Ideology, Morality, and Rule of Law, as an example, is presented in Sect. 6. The benefits and drawbacks should be further examined and discussed, and optimization techniques should be suggested for the project, aiming at raising the standard of instruction in this course. A summary is given in Sect. 7.

2. Architectures of YOLOv8 Network and CQTA-YOLOv8-NET

The YOLOv8 network, developed by Ultralytics, represents a leading-edge real-time object detection architecture operating in a single-stage paradigm.[6,21] Beyond fundamental detection capabilities, it provides a scalable suite of pretrained configurations ranging from ultra-lightweight (nano) to high capacity (xlarge) to accommodate various hardware resources and task-specific requirements. Key architectural and training defaults of YOLOv8 are outlined below.[6,21]

(1) The standard input dimension is fixed at 640 × 640 pixels. Preprocessing is automatically performed, including intensity normalization, mosaic augmentation, random horizontal mirroring, and additional conventional augmentation.
(2) Feature extraction is realized through a hierarchical backbone composed of repeated CBS units, which is a combination of convolution, batch normalization (BN), and sigmoid linear unit (SiLU) activation, alongside C2f modules. The implement enhances residual learning for facilitating the progressive abstraction of discriminative spatial features.[6]
(3) Feature enhancement and cross-scale integration are achieved via a hybrid neck structure, which is a synergistic fusion of a path aggregation network (PAN) and feature pyramid network (FPN), termed PAN-FPN. The structure supports concurrent top-down semantic refinement and bottom-up spatial detail recovery across feature levels.
(4) Detection outputs are generated through a disentangled head architecture, where class prediction and bounding-box coordinate estimation are handled by separate subnetworks. Three parallel heads process feature maps sized 80 × 80, 40 × 40, and 20 × 20 pixels, the picture elements of an image. Each head is tuned for detecting objects at coarse, medium, and fine granularities, respectively.

(5) Model optimization relies on a composite objective function comprising a regression loss based on complete intersection over union (CIoU), a classification loss incorporating focal weighting to mitigate class imbalance, and the distribution focal loss (DFL) to refine distribution-aware localization.
(6) Optimization employs either Adam or SGD with momentum and complementary training protocols, which include gradual learning rate ramp-up, root-mean-squared regularization on weights, and convergence-triggered early termination. The default settings are integrated to ensure robust and efficient model adaptation.

2.1 Architecture of YOLOv8 network

Because of its stable detection accuracy and straightforward structure, we apply the YOLOv8 network as the baseline network in this study. The architectural layout of the YOLOv8 network is depicted in Fig. 1, and Ref. 6 provides thorough definitions and operational recommendations. Each key portion serves the following purposes.

(1) Input. Preprocessing is performed on the input image to make it 640 × 640 × 3 pixels and to standardize the image data.
(2) Backbone. An essential component for extracting characteristics from the input image is the Backbone. The CBS and C2f modules make up the majority of YOLOv8's backbone. The CBS module is responsible for down-sampling, whereas the C2f module is used for feature extraction.
(3) Neck. The PAN-FPN structure, which adds a bottom-up path based on an FPN, is used by the neck component of YOLOv8. This route makes it possible to re-fuse low-level and high-level characteristics, which improves object detection accuracy and helps capture targets of different sizes.

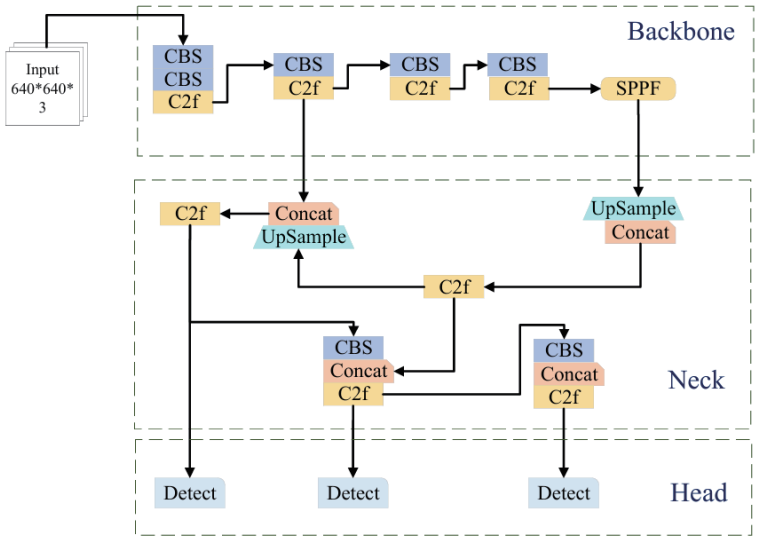

Fig. 1. (Color online) Architectural layout of YOLOv8.

(4) Head. The head uses a decoupled head structure, with several decoupled heads processing the various feature map levels—80×80×256, 40×40×512, and 20×20×1024—that are produced by the neck. The head module generates the final predictions, which comprise bounding box coordinates and dimensions, along with class-specific confidence scores for each predicted bounding box, thereby enabling precise object classification and localization.

2.2 Architecture of CQTA-YOLOv8-NET

In actual teaching scenarios, object detection within classroom environments presents distinct challenges. When analyzing the specific context of a classroom, it becomes evident that student-related target information is often small in scale and densely distributed. Within a spacious and well-lit classroom, dozens of students occupy individual seats, with each figure appearing small relative to the overall scene. Furthermore, limited spatial separation between individuals results in closely adjacent positioning, causing student targets captured by surveillance systems or teaching aids to exhibit characteristics of compactness and low resolution. Traditional object detection algorithms, such as the standard YOLOv8 architecture, frequently struggle with accurately detecting small objects within such images. Because of the limited pixel area occupied by these small targets, their distinguishing features may be overshadowed by surrounding environmental elements and neighboring objects, thereby reducing the model's ability to precisely identify and localize each student.

Although the YOLOv8 network is recognized as an efficient object detection framework, its effectiveness in identifying small-scale objects may not reach optimal levels under default settings. To address this limitation and improve the model's performance in detecting small targets, particularly students within a classroom setting, we have restructured the detection head module on the basis of the original architecture. Specifically, we introduced an additional detection head especially designed for small objects to enhance the model's sensitivity and precision in capturing such targets. The overall structure of CTQA-YOLOv8-NET is presented in Fig. 2.

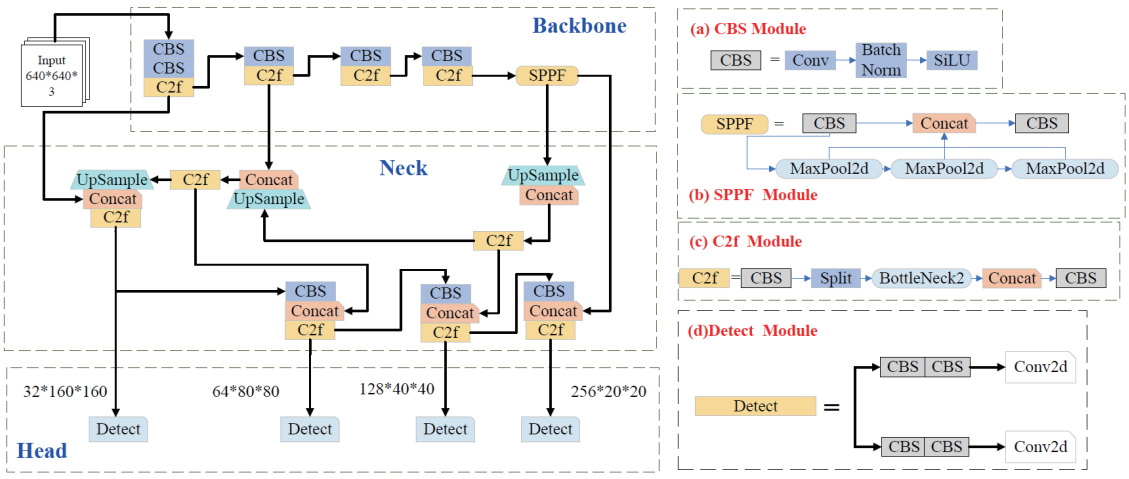

Fig. 2. (Color online) Architecture of CQTA-YOLOv8-NET.

Small objects typically occupy only a limited number of pixels in images and are prone to being masked by dominant semantic features in high-level feature maps. By integrating an extra detection head at the shallow feature map level, more fine-grained spatial details can be preserved and utilized effectively, thereby significantly improving detection accuracy for small targets. This newly added detection head operates on higher-resolution feature maps, enabling more precise localization and recognition of small-scale objects.

3. Dataset Labeling for Network Training

3.1 Experimental configuration

The Windows 10 operating system with a 16 GB memory capacity was used in the experimental setup for this investigation. The 11th generation i7-11800H, 16 core, 2.30 GHz CPU, 512 G SSD, and NVIDIA GeForce RTX3060 GPU with 16 GB of VRAM made up the system configuration. The software is implemented using CUDA version 11.1 to speed up computations and PyTorch version 1.13.1+cu116, a deep-learning framework. In this study project, Python 3.10 was the main programming language.

3.2 Collection and labeling of network training datasets

During the data collection phase, we sought to enable the precise and efficient analysis and evaluation of classroom teaching quality by gathering a wide variety of students' classroom behavior data. To uphold ethical standards, rigorous privacy protection measures were applied to the collected data. Considering the limited diversity of the initial data sources, publicly available classroom behavior datasets were also incorporated to enrich the overall dataset. These external datasets included classroom scenarios from diverse regions and educational institutions, thereby addressing potential limitations in the self-collected data. Furthermore, to enhance the dataset's variability and general applicability, data was captured from multiple angles within the classroom environment, such as the front, back, and sides, as depicted in Fig. 3. Photo recordings were conducted from different perspectives to comprehensively capture student behavior. During the photo collection process, suitable resolution and frame rate settings were chosen to ensure that detailed student actions can be clearly observed.

Data image labeling plays a crucial role in identifying students' behaviors in the classroom, and thus directly influencing the accuracy and effectiveness of CTQA-YOLOv8-NET training. In this study, professional labeling software was employed to annotate the collected image data.[7] A total of approximately 35000 images depicting students' classroom behaviors were collected and categorized into six distinct behavioral classes, that is, hand-raising, reading, writing, using a phone, bowing the head, and leaning over the table. Figure 4 presents the statistical distribution of these labeled categories used during network training and validation. Among them, the "reading" category contains the largest number of images, approaching 14000. The "hand-raising" category includes approximately 9000 images. The "writing" and "using a phone" categories contain similar volumes of around 5000 images each. The categories of

Fig. 3. (Color online) Student classroom behavior dataset.

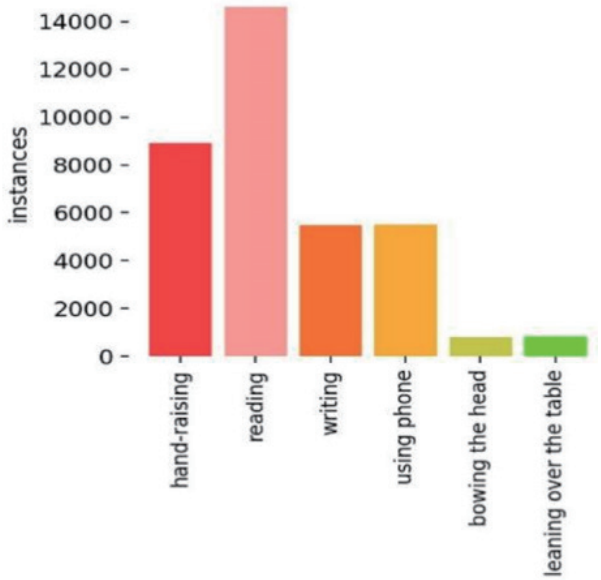

Fig. 4. (Color online) Statistics of category labels in the training dataset.

"bowing the head" and "leaning over the table" have the smallest datasets, with approximately 1000 images each.

4. Training of Network

4.1 Training strategy and evaluation index

The training approach used in this study was the same for every experiment. The input image size stays fixed at 640 by 640 pixels, and a batch size of 16 is employed. Two hundred epochs make up the training procedure, which begins with a learning rate of 0.001. In this study, the Adam optimizer[27] is used as the optimizer to optimize the proposed model.

The Adam optimizer[27] uses three key hyperparameters for stable and efficient convergence, namely, the exponential decay rate for first-moment estimates, β_1, the second-moment estimates, β_2, and a small numerical stability constant, ε. Specifically, β_1 controls how much weight past gradients receive in the running average of the first moment. It is typically set close to 1 to smooth updates by incorporating longer-term gradient history. Similarly, β_2 governs the decay of the running average of squared gradients (the second moment). It is set near 1 to improve stability by emphasizing historical gradient magnitudes. Finally, ε is a small positive constant added to the denominator of the update rule to avoid division by zero and ensure numerical robustness. In the following training process, a coefficient set of $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 10^{-4}$ is applied. Additionally, to guarantee consistency between experimental results and the distribution of the real dataset while attaining full model convergence, an IoU threshold of 0.5 is used for assessment reasons, and data augmentation[28] is turned off for the final 30 epochs.

To improve the training effectiveness of CTQA-YOLOv8-Net, the model's initial weights were initialized on the basis of the pretrained YOLOv8 network parameters. This initialization method facilitates faster convergence and enhances the overall performance of the model. Moreover, a learning rate warm-up approach[27] was incorporated, in which the learning rate progressively rises from a low starting point to the intended initial value during the initial training phase. This technique helps maintain training stability by avoiding excessively high learning rates at the outset. In particular, the warm-up procedure was carried out over the first three epochs.

In addition, to reduce the likelihood of overfitting, a combination of weight decay regularization[27] and norm-based constraint techniques was applied within the parameter space. The introduction of a preset regularization intensity factor enabled CTQA-YOLOv8-Net to develop compression mapping properties in the hypothesis space, leading to a decrease in the upper limit of generalization error. As a result, the potential for overfitting due to mismatches between training and testing data distributions is significantly diminished.

One crucial activity that necessitates the network's speed and accuracy is the identification of different student behavior patterns in the classroom. Consequently, detection accuracy is a crucial measure in the experiment. In object detection, mean average precision (*mAP*) is frequently used to represent the model's overall accuracy. The performance of the relevant network is assessed in this study using a mAP with an IoU threshold of 0.5 (*mAP*@0.5). The integration of precision (*P*) at various recall levels (*R*) is represented by the average precision (*AP*). While *mAP* is the mean *AP* for all target classes, *AP* for each class is an amalgam of the network's detection accuracy for that particular class. The formulas for computing these metrics are shown in Eqs. (1) through (4).

$$P = \frac{TP}{TP + FP} \tag{1}$$

$$R = \frac{TP}{TP + FN} \tag{2}$$

$$AP = \int_0^1 P(R)dR \tag{3}$$

$$mAP = \frac{1}{N}\sum_{i=0}^{n} AP(i) \tag{4}$$

The network's accuracy in predicting positive samples is shown as *TP* (true positive), and its error in predicting positive samples by *FP* (false positive). The number of negative samples that the network incorrectly predicted is shown by *FN* (false negative). *N* is the total number of categories that the dataset includes.

4.2 Network training

Figure 5 shows the training results for CTQA-YOLOv8-NET trained by 200 epochs. It shows promising performance in the task of identifying students' classroom behavior with low computing cost and good accuracy. The following important conclusions have been drawn from the results of thorough testing and analysis.

The bounding box regression loss function, train/box_loss, decreased sharply from approximately 1.4 to around 0.5, indicating a continuous improvement in the network's ability to predict target bounding box positions.

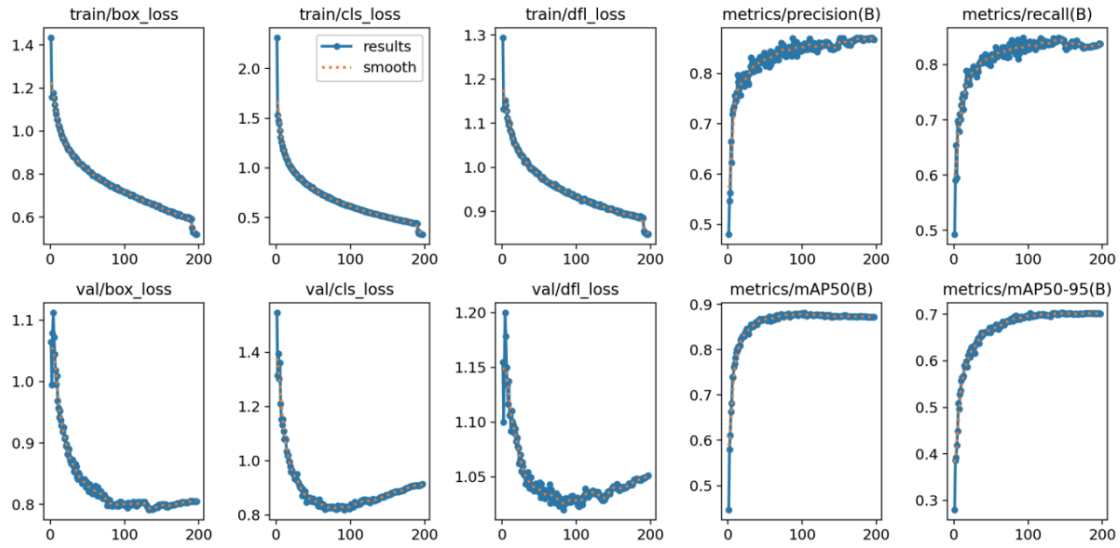

Fig. 5. (Color online) Results of the network training process.

(1) The classification loss function, train/cls_loss, dropped significantly from about 2.0, reflecting a progressive enhancement in the network's capacity to classify target categories during training.
(2) The distributed focus loss function, train/dfl_loss, related to the positional distribution of bounding boxes, declined from 1.3 to below 0.9, suggesting an ongoing optimization in the network's performance in predicting positional distributions.
(3) During validation, the bounding box loss function, val/box_loss, showed minor fluctuations while decreasing from 1.1 to approximately 0.8, remaining generally stable, which indicates relatively reliable predictions on the validation dataset.
(4) The classification loss function in the validation phase, val/cls_loss, experienced slight fluctuations after a rapid decline but remained at a consistently low level, demonstrating stable classification performance on the validation dataset.
(5) The distributed focus loss function during validation, val/dfl_loss, exhibited some fluctuation following the initial decrease, implying slightly reduced stability in predicting bounding box distributions; however, the variation remains within acceptable limits.
(6) The precision rate, metrics/precision(B), increased rapidly from 0.5 to nearly 0.9 and subsequently stabilized, indicating the network's high accuracy in identifying positive samples.
(7) The recall rate, metrics/recall(B), rose from approximately 0.5 to 0.9, showing that the network is effective in detecting a large proportion of actual targets.
(8) At an IoU threshold of 0.5, the mean average precision, metrics/mAP50(B), quickly increased to around 0.9 and remained steady, demonstrating strong detection performance under standard evaluation criteria.
(9) The mean average precision across IoU thresholds from 0.5 to 0.95, metrics/mAP50-95(B), stabilized at approximately 0.7, indicating that the network maintains good generalization performance even under more rigorous evaluation conditions.

Furthermore, as demonstrated in Figs. 6 and 7, the training finished early at 200 epochs and the convergence was smooth, demonstrating the effectiveness of the hyperparameters applied.

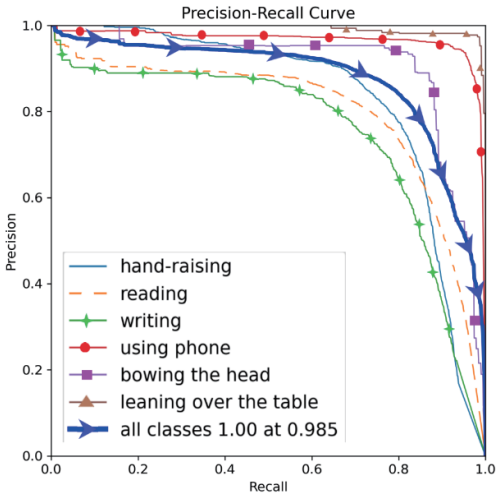

Fig. 6. (Color online) Curves of precision versus recall.

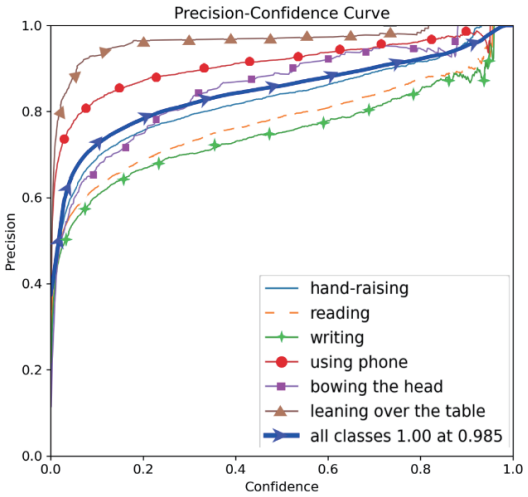

Fig. 7. (Color online) Curves of precision versus confidence.

Figure 6 shows the precision–recall curves, which reflect the performance of object detection across various categories. The overall precision versus recall is represented by the blue curve, with a corresponding mAP value of 0.874, indicating a balanced assessment of the network's effectiveness across all categories. A precision–recall curve that approaches the upper-right corner signifies higher recall while maintaining high precision, thereby demonstrating superior capability in identifying students' behaviors in classroom settings.

Among the specific detected behaviors, leaning over the table, using a phone, and bowing the head exhibit consistently strong precision exceeding 0.9 across different recall levels. This result indicates that a highly effective detection performance is achieved. The occurrence frequency of these three types of behavior in the classroom is suitably utilized as key performance indicators for evaluating the reduction in learning effectiveness within this study.

Figure 7 illustrates the precision performance across different categories under varying confidence thresholds. As the confidence level increases progressively from 0 to 1, the precision rates associated with each category generally exhibit an upward trend. This suggests that a higher confidence level, indicating stronger network certainty in prediction outcomes, is typically correlated with a higher precision rate. Notably, the blue curve shows that a precision value of 1.00 is reached at a confidence threshold of 0.985, indicating that when the confidence level is set to 0.985, the overall precision across all categories achieves perfect accuracy. This observation implies that under conditions of very high confidence, the network's predictive performance remains exceptionally accurate.

5. Classroom Behavior Detection System Design and Result Visualization

5.1 Design of student classroom behavior detection

Upon completion of the training and validation of CTQA-YOLOv8-NET, the subsequent stage focuses on analyzing and detecting student behaviors in the classroom setting. A high-resolution surveillance digital camera was employed to record the classroom setting throughout the entire instructional session. It was fixed to the ceiling near the front-right section of the room, providing a top-down view that encompassed all student desks and facilitated holistic monitoring of classroom activities.

The procedure for detecting classroom behaviors starts with extracting video frames from recorded classroom sessions, followed by a series of sequential data processing phases. These steps collectively facilitate the precise recognition of student actions and the production of relevant output outcomes. A schematic diagram of this workflow is provided in Fig. 8. The detailed procedures for classroom behavior detection are summarized below.

Step 1. Acquire the video stream and segment it into individual frames

The video stream is captured from the classroom camera and transmitted to the management system through video capture cards or network interfaces. Within the management system, the video stream is segmented into frame-by-frame images at one-minute intervals. As a result, a 45-minute course video will produce 45 distinct frame images, which will serve as the input data for subsequent classroom behavior detection.

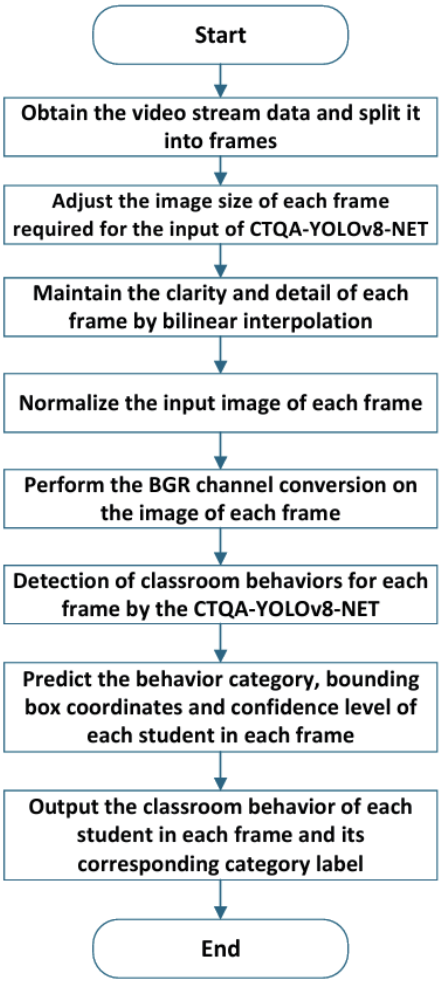

Fig. 8. (Color online) Flowchart for detecting students' classroom behavior.

Step 2. Adjust the image size of each frame for CTQA-YOLOv8-NET
All extracted frame images are scaled to a standardized resolution of 640×640 pixels, in accordance with the model's defined input size

Step 3. Ensure the preservation of frame clarity and detail by bilinear interpolation.
Throughout the resizing procedure, CTQA-YOLOv8-NET is simultaneously employed to maintain the visual sharpness and intricate details of each individual frame.

Step 4. Normalize the input image of each frame
The pixel intensity values, which originally ranged from 0 to 255, are scaled to a normalized range of 0 to 1. This preprocessing stage helps to speed up the training convergence of the CTQA-YOLOv8-NET model.

Step 5. Perform the BGR channel conversion
This step involves conducting BGR channel conversion on the image data of each frame. Furthermore, a transformation process was implemented to convert the image from RGB to BGR format, ensuring compliance with the input specifications of the neural network model.

Step 6. Detection of classroom behaviors

The processed images are then fed into the trained CTQA-YOLOv8-NET system to identify and recognize various learning behaviors in the classroom.

Step 7. Predict the behavior category

For each student present in every frame, the model forecasts their behavioral classification, bounding box positions, and associated confidence scores. The network first detects all potential student objects in each frame, and then proceeds to estimate their respective behavioral types, spatial coordinates, and reliability measures.

Step 8. Output behaviors in the classroom

For each frame, generate the classroom behavior of every student together with the associated category label. By utilizing the features and patterns acquired during the training process, the model recognizes various student behaviors in the images, including actions such as hand-raising, reading, writing, using a phone, bowing the head, or leaning over the table. The output comprises the respective category label, the positional coordinates of the bounding box's top-left and bottom-right corners, and a confidence score that reflects the probability of the detected object being correctly classified into the predicted category.

5.2 Visualized CTQA system

The output detection results from CTQA-YOLOv8-NET must undergo postprocessing to derive the final classroom behavior identification outcomes and support the visual display of student target detection. In this study, a confidence threshold of 0.55 was established; when the confidence level of a detected behavior surpasses this value, the behavior classification is regarded as reliable, which significantly minimizes false positive detections. Furthermore, the non-maximum suppression (NMS) technique is employed to remove bounding boxes with excessive spatial overlap among the filtered candidate targets. This method functions by computing the IoU metric between detection boxes. More precisely, if the IoU of two overlapping boxes exceeds the set threshold of 0.55, the system keeps the box with the higher confidence score while suppressing those with lower scores, thereby avoiding multiple detections of the same behavioral event.

The postprocessed test results will be presented in a visual format to facilitate viewing and analysis. Within the detection results visualization system interface, student classroom behaviors are annotated on the images using rectangular bounding boxes with distinct colors. Adjacent to each box, corresponding behavior category labels and confidence scores are displayed. An illustration of the CTQA system's visualization interface is provided in Fig. 9.

As shown in Fig. 9, the frequency of six categorized classroom behaviors is aggregated to evaluate students' academic engagement during the lesson. For each frame captured, the valid detection outcomes related to three types of conduct, hand-raising, reading, and writing, are combined to reflect the count of students' effective learning behaviors. In contrast, the detected occurrences of the other three behaviors, such as using a phone, bowing the head, and leaning over the table, are totaled to signify the number of invalid learning behaviors. Given that one

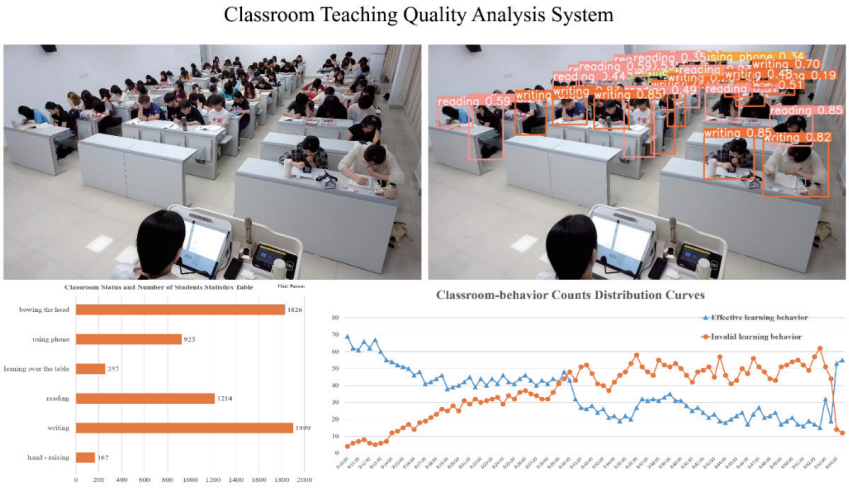

Fig. 9. (Color online) Interface design for classroom behavior visualization system.

frame is sampled every minute from the video feed, the counts of both effective and invalid learning students are documented per frame. These records are then utilized to construct a variation trend curve, as demonstrated in Fig. 9, which illustrates the real-time fluctuations in student engagement throughout the entire class period. Such a trend curve enables further investigation into the suitability of instructional content and teaching approaches, offering empirical data support for enhancing educational effectiveness.

6. Case Study in Course of Ideology, Morality, and Rule of Law

In this section, we empirically evaluate classroom instruction quality for the course, Ideology, Morality, and Rule of Law, by applying the introduced CTQA-YOLOv8-NET. In the course, each 45-minute class session involves 75 students. The instructional procedure is structured as follows.

(1) Introduction to new topics and foundational knowledge (five minutes)
Instructors use multimedia, such as images, texts, videos, and music, to present real-life scenarios, illustrative stories, or thematic elements, aiming to engage students cognitively and build initial understanding.

(2) Delivery of core content (35 minutes)
Content is organized into clear and task-based modules. Each module uses a question-centered and inquiry-driven approach. This builds a coherent and hierarchical knowledge structure, helping students internalize core concepts and apply them to real-world problems.

(3) Recap and consolidation (five minutes)
Instructors summarize key points to reinforce comprehension and long-term retention.

6.1 Analysis and assessment of teaching quality

The CTQA-YOLOv8-NET developed in this study detects and classifies student behaviors. It quantifies six behavior types, grouped into two categories: (1) effective learning behaviors,

including hand-raising, reading, and writing, whose frequencies are summed to yield a total engagement count, and (2) invalid learning behaviors, involving phone use, head-bowing, and leaning over the table, whose occurrences are likewise summed. The temporal distribution of students exhibiting behavior of each category is tracked and summarized across class time, as shown in Fig. 10.

As shown in Fig. 10, during the first five minutes, students exhibit more effective than invalid learning behaviors, which indicates an initially favorable learning environment. However, effective learning behaviors gradually decline and invalid ones rise as the lesson proceeds. The two trends cross at the 21st minute, which is a critical turning point signaling deteriorating classroom conditions. Beyond this moment, the count of effective learning behaviors continues to fall and that of invalid ones rises, which reflects a progressive decline in learning efficacy.

On the basis of observed fluctuations in students' effective versus invalid learning behaviors in class, the following issues have been identified.

(1) Student concentration declines over time

Focus is relatively high during the first 20 minutes, but is fatigued from minute 21 onward, which is evidenced by the decrease in the counts of effective learning behaviors. Invalid learning behaviors persist across all students until roughly three minutes before the class ends.

(2) Teaching content and delivery need improvement

Instructional methods are perceived as monotonous and content lacks engagement, leading to a steady decline in the number of students engaged in effective learning behaviors.

6.2 Review and enhance teaching content design

From the results of the prior analysis of challenges in delivering the course, Ideology, Morality, and Rule of Law, the following enhancements are recommended.

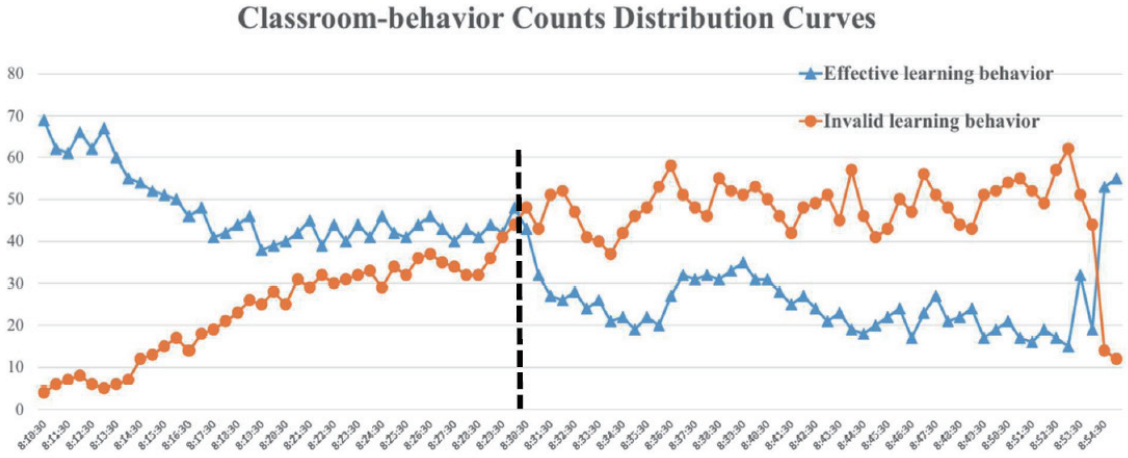

Fig. 10. (Color online) Classroom-behavior counts distribution curves of the original course.

(1) A phased teaching approach

In the first 20 minutes, focus, reasoning, and memory peak of students are ideal for teaching core theories or complex content. As attention typically declines after 20 minutes, interactive re-engagements of students, such as Q&A, multimedia, group discussion, or teacher–student dialogue are carried out. In the final 5 minutes, consolidate learning with quick in-class checks, instructor-led questions, or reinforce key takeaways by providing a visual concept map. Thus, the 45-minute class is structured into three clear segments. That is, 20 minutes of focused instruction, 20 minutes of interactive participation, and 5 minutes of reflection and summary.

(2) Instructional content must be carefully designed

The first step is administering a learning questionnaire to gauge students' baseline knowledge. Results are analyzed to identify and address common misconceptions and difficulties. During instruction, the teacher prioritizes key concepts students typically find challenging. Class materials are streamlined and enriched with engaging narratives or relevant case studies, such as real-life issues drawn from university students' daily experiences or current societal topics, to boost engagement, deepen understanding, and improve knowledge acquisition and retention.

(3) Implement clear reward and accountability measures

In classroom communication activities, incentives, such as bonus points, assignment exemptions, internship recommendations, and certificates, boost intrinsic motivation and professional growth. For low engagement, constructive interventions, e.g., point deductions, supplemental assignments, temporary device restrictions, or reflective self-assessments, promote consistent participation.

6.3 Analysis of enhanced teaching activities based on improved content design

Per the previous section, the redesigned Ideology, Morality, and Rule of Law course is delivered in alternating sessions. CTQA-YOLOv8-NET is reapplied to empirically assess its effectiveness. Figure 11 shows the statistical curve of effective versus invalid learning behavior frequencies observed in these revised sessions.

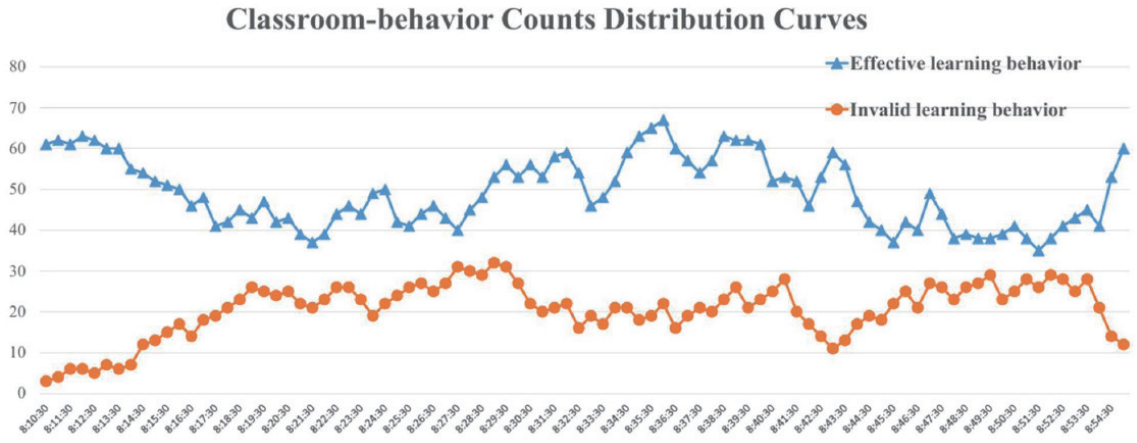

Fig. 11. (Color online) Classroom-behavior counts distribution curves of the improved course.

According to Fig. 11, teaching activities and effectiveness assessment of the enhanced course are interpreted as follows.

(1) First 20 minutes (key concepts)
Effective learning behaviors rise steadily and invalid behaviors increase only slightly, indicating well-structured content sustains attention.

(2) Interactive segment (20th–40th minute)
Counts of effective and invalid learning behaviors remain nearly stable and parallel. It is indicated that multimodal strategies, which are combined with clear incentives and consequences, effectively sustain learner engagement.

(3) Final stage (40th minute to the end)
Effective-learning behavior rises steadily while invalid behavior declines. It is shown that learner-centered review methods successfully re-engage learners and reinforce understanding and retention.

6.4 Discussion on strategies to improve classroom instruction

A student-centered educational paradigm is essential. Traditional didactic instruction no longer meets today's students' need for autonomy. Classrooms must prioritize active engagement over passive reception and rote learning. Content should be reorganized and delivered through questioning, group debate, discussion, and case analysis to boost motivation and improve outcomes.

The teaching structure can be optimized using an evidence-based segmentation approach. To address declining student attention, which is identified via real-time learning behavior analysis, a standard 45-minute lesson is split into three purposeful segments. Across all segments, instructional materials are dynamically integrated using diverse modalities to boost engagement and support structured, deep understanding.

Within the interactive segment, collaborative discussions, scaffolded debates, and context-anchored simulations foster meaningful educator–learner interaction. In the concluding phase, concept maps and data visualizations help students build coherent, well-organized knowledge frameworks, which reinforce achievement and sustain intrinsic motivation.

A clear reward-and-discipline system is implemented and a well-structured, logical design enhances teaching effectiveness and learning outcomes through greater efficiency. Surveys and interviews help teachers understand students' real needs, enabling tailored incentives and interventions that foster proactive learning and analytical reasoning.

7. Conclusions

In this study, we presented the development of a deep learning system aimed at assessing classroom teaching quality. Distinct from prior studies, we proposed CTQA-YOLOv8-NET, which is an innovative network capable of detecting and recognizing students' behaviors during class. On the basis of this network, we further developed an integrated visual analytical framework to evaluate teaching performance within higher education settings.

The proposed system will enable precise enhancements in both instructional materials and pedagogical strategies. To demonstrate its practical applicability, a case study was carried out on the course, Ideology, Morality, and Rule of Law, confirming the system's effectiveness. Moreover, the technology developed here can be easily applied to other fields, such as monitoring worker behavior in plants and factories, traffic behavior in heavily congested areas, or people's behavior in crowded situations.

Acknowledgments

This work was carried out as part of the Joint Innovation Project of Industry–University–Research Collaboration of the Department of Science and Technology of Fujian Province (Grant no. 2023H6036), the 2024 National Social Science Foundation of China General Project, 'Research on the Guidance and Dissemination Strategies of Socialist Core Values under the Condition of Artificial Intelligence' (Grant no. 24BKS118), and the Operational Funding of the Advanced Talents for Scientific Research (Grant no. 19YG04) of Sanming University. The authors also acknowledge the support from the School of Marxism and the School of Mechanical and Electric Engineering, Sanming University.

References

1 R. Khanam, T. Asghar, and M. Hussain: Solar **5** (2025) 6. https://doi.org/10.3390/solar5010006
2 L. Zhao, T. Tohti, and A. Hamdulla: Signal, Image Video Process. **17** (2023) 4435. https://doi.org/10.1007/s11760-023-02677-x
3 Z. Diao, X. Huang, H. Liu, and Z. Liu: Int. J. Intell. Syst. **2023** (2023) 8879622. https://doi.org/10.1155/2023/8879622
4 A. Magdy, M. S. Moustafa, H. M. Ebied, and M. F. Tolba: Sci. Rep. **15** (2025) 16163. https://doi.org/10.1038/s41598-025-99242-y
5 J. Z. Pan, C. H. Yang, L. Wu, W. H. Tang, and K. C. Wang: Sens. Mater. **35** (2023) 4653. https://doi.org/10.18494/SAM4589
6 Q. Zhou and L. Chen: IEEE Trans. Aerosp. Electron. Syst. **61** (2022) 1120. https://doi.org/10.1109/TAES.2022.3181237
7 Z. Chen, B. Chen, Y. Huang, and Z. Zhou: Appl. Sci. **15** (2025) 2823. https://doi.org/10.3390/app15052823
8 Y. Liu, Y. Liu, X. Guo, X. Ling, and Q. Geng: Sci. Rep. **15** (2025) 11105. https://doi.org/10.1038/s41598-025-94936-9
9 C. Shah, M. M. Nabi, S. Y. Alaba, I. A. Ebu, J. Prior, M. D. Campbell, R. Caillouet, M. D. Grossi, T. Rowell, F. Wallace, J. E. Ball, and R. Moorhead: Sensors **25** (2025) 1846. https://doi.org/10.3390/s25061846
10 X. Cao, C. Li, and H. Zhai: Comput. Mater. Continua **83** (2025) 5487. https://doi.org/10.32604/cmc.2025.061823
11 S. Zhuo, H. Bai, L. Jiang, X. Zhou, X. Duan, Y. Ma, and Z. Zhou: IEEE Access **13** (2025) 47653. https://doi.org/10.1109/ACCESS.2025.3550947
12 T. T. Huynh, H. T. Nguyen, and D. T. Phu: Comput. Mater. Continua **81** (2024) 2281. https://doi.org/10.32604/cmc.2024.057954
13 W. Wu, H. Cheng, J. Pan, L. Zhong, and Q. Zhang: Appl. Sci. **15** (2025) 4586. https://doi.org/10.3390/app15084586
14 H. Liu, K. Wang, Y. Wang, M. Zhang, Q. Liu, and W. Li: Electron. **14** (2025) 955. https://doi.org/10.3390/electronics14050955
15 A. Li, C. Wang, T. Ji, Q. Wang, and T. Zhang: Agriculture **14** (2024) 2268. https://doi.org/10.3390/agriculture14122268
16 Y. Wang, Y. Jiang, H. Xu, C. Xiao, and K. Zhao: Processes **13** (2025) 201. https://doi.org/10.3390/pr13010201
17 K. Yan and Z. Zhang: IEEE Access **9** (2021) 150925. https://doi.org/10.1109/ACCESS.2021.3125703
18 Z. B. Yin, F. Y. Liu, H. Geng, Y. J. Xi, D. B. Zeng, C. J. Si, and M. D. Shi: PLoS One **19** (2024) e0296314. https://doi.org/10.1371/journal.pone.0296314

19 L. Tang, T. Xie, Y. Yang, and H. Wang: Appl. Sci. **12** (2022) 6790. https://doi.org/10.3390/app12136790
20 L. Ma, T. Zhou, B. Yu, Z. Li, R. Fang, and X. Liu: Electronics **13** (2024) 3726. https://doi.org/10.3390/electronics13183726
21 Y. Cao, Q. Cao, C. Qian, and D. Chen: YOLO-AMM: Sensors **25** (2025) 1142. https://doi.org/10.3390/s25041142
22 L. Zhou, X. Liu, X. Guan, and Y. Cheng: Sensors **25** (2025) 3132. https://doi.org/10.3390/s25103132
23 M. Li, C. Yuan, Q. Wang, C. Zhao, J. Chen, and L. Liu: Comput. Eng. **51** (2025) 287. https://doi.org/10.19678/j.issn.1000-3428.0068656
24 S. Peng, X. Zhang, L. Zhou, and P. Wang: Sensors **25** (2025) 3073. https://doi.org/10.3390/s25103073
25 M. Dang, G. Liu, H. Li, Q. Xu, X. Wang, and R. Pan: Appl. Intel. **54** (2024) 4935. https://doi.org/10.1007/s10489-024-05409-x
26 S. Wu and B. Wang: Knowledge-Based Syst. **295** (2024) 111836. https://doi.org/10.1016/j.knosys.2024.111836
27 K. M. Elgamily, M. A. Mohamed, A. M. Abou-Taleb, and M. M. Ata: Sci. Rep. **15** (2025) 7226. https://doi.org/10.1038/s41598-025-89124-8
28 W. Xue, G. Xu, N. Yang, and J. Liu: J. Supercomput. **81** (2025) 417. https://doi.org/10.1007/s11227-024-06910-3