IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 42, No. 4, August, 2025, pp. 2323-2332
Journal homepage: http://iieta.org/journals/ts

Image-Based Intelligent Classroom Monitoring and Learning Behavior Analysis via Joint Object Detection and Semantic Segmentation

Kun Du[1], Ling Xu[2*], Yanwu Zhao[3], Lei Wang[3], Xue Wang[1], Xiaoying Chen[1], Rui Ma[1]

[1] Education and Teaching Research Center of Handan University, Handan 056038, China
[2] Affiliated School of Hebei University of Engineering, Handan 056038, China
[3] Education College of Handan University, Handan 056038, China

Corresponding Author Email: 13283188316@163.com

https://doi.org/10.18280/ts.420441

Received: 5 January 2025
Revised: 13 May 2025
Accepted: 26 June 2025
Available online: 14 August 2025

Keywords:
intelligent classroom, image analysis, object detection, semantic segmentation, learning behavior analysis, attention mechanism

ABSTRACT

With the advancement of educational informatization, intelligent classrooms increasingly rely on image analysis technologies to automate environmental monitoring and learning behavior analysis. However, current research faces three key limitations: (1) the disjoint handling of object detection and semantic segmentation leads to suboptimal feature utilization; (2) existing models perform poorly in detecting dim classroom boundaries due to inadequate attention mechanisms; and (3) conventional loss functions struggle to address the pixel imbalance between boundary lines and the background. To address these challenges, this paper proposes a joint object detection and semantic segmentation model tailored for intelligent classroom scenarios. The model employs a shared encoder with dual decoder branches to achieve collaborative reasoning for both environmental object detection and learning behavior region segmentation. A Bi-directional Feature Pyramid Network (BiFPN) is integrated to introduce an attention-like weighted feature fusion mechanism, enhancing the capture of subtle boundary features. Additionally, an improved EFL Focal Loss is introduced to mitigate pixel imbalance issues. The main contributions of this work include: constructing a unified framework to enhance feature synergy between detection and segmentation tasks; designing a targeted attention mechanism to optimize boundary detection; and improving the loss function to balance pixel-wise training. Experimental results demonstrate improved completeness and accuracy in classroom scene analysis.

1. INTRODUCTION

With the deepening development of educational informatization, intelligent classrooms [1-4], as the core carrier integrating information technology with teaching and education, are gradually becoming an important support for the innovation of modern education models. Against this background, how to realize real-time perception of the classroom environment and accurate judgment of learning behaviors through technical means has become a key issue in improving teaching quality and optimizing teaching management. The rapid development of artificial intelligence and computer vision technologies [5-7] provides a feasible technical path for this demand. Image analysis-based monitoring and analysis methods [8-10] can break through the limitations of traditional manual observation and realize the automatic acquisition and interpretation of environmental information such as classroom lighting and equipment status, as well as learning data such as student concentration and interactive behaviors.

At present, research in related fields still has three significant limitations: first, most existing methods treat object detection and semantic segmentation tasks separately [11-13], resulting in a lack of synergy in the feature extraction process of the two tasks. For example, some studies only use a single model to complete either environmental object recognition or behavior region segmentation, making it difficult to meet the dual requirements of "environmental element localization" and "behavior region division" in classroom scenes. Second, for the commonly seen dim boundary lines in classroom scenes, such as the edges of desks and chairs and the borders of blackboards, existing detection models often have weak feature responses and blurred boundaries due to the limitations of attention mechanism design [14-16]. Third, in pixel-level analysis [17-19], problems such as low proportion of boundary pixels and long-tail distribution between foreground and background pixels are common. Traditional loss functions find it difficult to balance the training weights of different pixel categories, resulting in boundary details being easily obscured by background information in the segmentation results.

This paper focuses on the above problems and proposes an intelligent classroom environmental and learning behavior detection model based on object detection and semantic segmentation. The main research contents include: in terms of model structure, a framework composed of a shared encoder and two independent decoder branches is designed, realizing collaborative reasoning of environmental object detection and learning behavior region segmentation through a shared low-

level feature extraction module; in terms of feature enhancement, the BiFPN used introduces a weight allocation mechanism similar to the attention mechanism, which dynamically adjusts feature channel weights and spatial attention to strengthen the feature capture ability for dim boundary areas; in terms of loss function optimization, an improved version of EFL focal loss is adopted, which alleviates the imbalance between boundary pixels and background pixels by adaptively adjusting the loss weights of hard and easy samples. Through multidimensional model optimization, this study effectively improves the completeness of environmental monitoring and the accuracy of learning behavior analysis in intelligent classroom scenarios, providing a technical reference for building a more adaptive intelligent teaching support system.

2. INTELLIGENT CLASSROOM ENVIRONMENTAL MONITORING AND LEARNING BEHAVIOR DETECTION MODEL BASED ON OBJECT DETECTION AND SEMANTIC SEGMENTATION

2.1 Model design

According to the needs in the intelligent classroom environment, this paper needs to detect whether students are studying attentively in their seats, and at the same time determine whether workers have left the learning area, which requires the detection of the boundary lines of the intelligent classroom environment. The former belongs to the object detection task, while the latter can be regarded as a semantic segmentation task. In order to meet real-time requirements, this paper designs an intelligent classroom environmental monitoring and learning behavior detection model. The model adopts a structure with shared encoder and dual decoder branches, mainly based on the dual needs of "collaboration between object detection and semantic segmentation" and "real-time performance assurance" under the intelligent classroom scenario. On the one hand, learning behavior detection belongs to object detection, which requires locating specific objects and judging their states; intelligent classroom environmental monitoring belongs to semantic segmentation, which requires pixel-level boundary division. If two independent models are used for processing respectively, it will not only cause repeated computation in the feature extraction process and increase resource consumption, but also reduce the consistency of the results due to the disconnection of the feature correlation between the two tasks. For example, spatial matching deviations may occur between student position judgment and boundary line segmentation. The shared encoder can generate shared low-level features for both tasks through one-time feature extraction, while the dual decoder branches respectively perform feature refinement according to the task characteristics of object detection and semantic segmentation. This can reduce redundant computation to meet the response speed requirements of real-time monitoring, and also enhance the spatial correlation between “student status” and “region boundary” through feature sharing, thus improving the coordination of overall detection.

The BiFPN used in the model introduces a weight allocation mechanism similar to the attention mechanism to solve the problem of accuracy in detecting dim boundary lines in intelligent classrooms. In actual classroom scenarios, boundary lines such as the junction between the wall and the floor, and the separation line between the learning area and the non-learning area often appear dim due to uneven lighting, equipment occlusion and other factors. These boundary lines are the core basis for judging whether workers have left the learning area. If boundary line detection is blurry or broken, it may lead to incorrect region division and thus misjudgment. Traditional attention mechanisms often assign weights based on global feature distribution, which tends to suppress features in low-contrast regions such as dim boundary lines. The improved attention mechanism dynamically adjusts the weight allocation of feature channels and spatial positions to enhance feature response to dim regions: strengthening the feature channels related to edge detection in the channel dimension, and focusing on low-brightness but continuous edge regions in the spatial dimension, thereby improving the completeness and clarity of boundary lines and providing reliable spatial basis for region judgment.

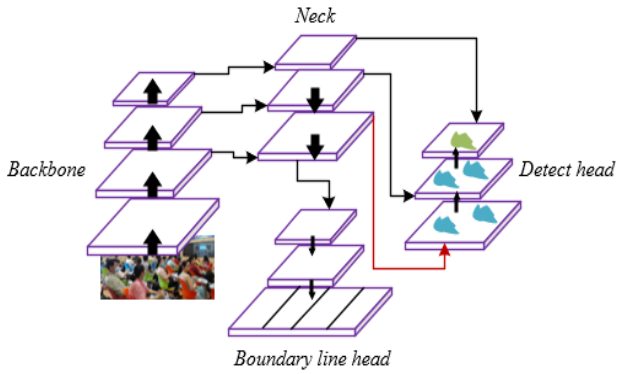

Figure 1. Structure diagram of intelligent classroom environmental monitoring and learning behavior detection model

The model also innovatively adopts an improved version of EFL focal loss, which directly addresses the core contradiction of “pixel imbalance” in intelligent classroom boundary line detection. As slender structures, boundary lines usually account for less than 5% of pixels in images, while background pixels account for an extremely high proportion, forming a typical long-tail distribution. If traditional loss functions are used, the model will overfit background features due to the numerical advantage of background pixels, causing the training weight of boundary line pixels to be diluted, and resulting in broken or missing boundary lines in the final segmentation result. This will directly affect the accurate definition of the learning area, thereby interfering with the judgment of "whether workers have left the learning area." The improved EFL focal loss solves this problem through two key optimizations: first, an adaptive hard sample mining mechanism is introduced to assign higher loss weights to “hard-to-detect pixels” such as boundary lines, ensuring that the model focuses on low-proportion but key regions during training; second, the loss ratio of positive and negative samples is dynamically adjusted to avoid excessive suppression of boundary line loss signals by background pixel loss values. This design enables the model to balance the learning priority of different pixel categories during training, ultimately improving the accuracy and stability of boundary line segmentation. Figure 1 shows the structure diagram of the intelligent classroom environmental monitoring and learning behavior detection model.

2.2 Model structure

The algorithm used in this paper is an improved version based on YOLOP. In order to make the network characteristics highly compatible with the requirements of object detection and semantic segmentation in intelligent classrooms, this paper chooses CSPNet as the backbone network. To ensure that the model can simultaneously perform object detection and semantic segmentation, the backbone network is required to extract sufficiently rich features while meeting the efficiency requirements of real-time monitoring. CSPNet happens to meet both needs: on the one hand, as an efficient structure verified by mainstream detection models such as YOLOv4 and YOLOv5, it inherits CSPNet's advantage of "enhancing feature representation capability", and can extract multi-level features from classroom images, including texture, shape, and edge. These features can support both object detection in judging student positions and states, and semantic segmentation in providing edge features for boundary line recognition; on the other hand, CSPNet achieves "lightweight and efficiency" through gradient flow optimization, with significantly lower computation and parameter count than traditional networks, enabling improved inference speed while ensuring feature richness, which aligns with the scenario requirements of real-time monitoring in intelligent classrooms. Figure 2 shows the structure diagram of the CSPNet used.

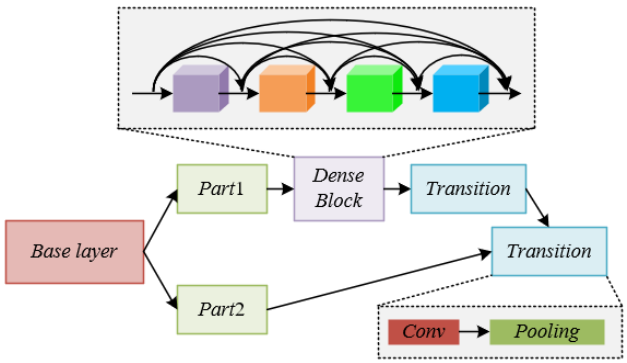

Figure 2. Structure diagram of the CSPNet used

The core principle of CSPNet in supporting object detection and semantic segmentation comes from its "gradient flow optimization" and "feature fusion design". At the gradient flow level, CSPNet divides the basic layer feature map into two parts, and realizes separated fusion through cross-stage layers, which avoids the optimization bottleneck caused by repeated gradient information in traditional networks, allowing gradients to propagate along different paths. This not only retains the richness of features but also reduces redundant computation. At the feature extraction level, CSPNet inherits the advantage of “feature reuse” from DenseNet, while avoiding feature redundancy through truncated gradient flow. The generated feature maps possess both “detail preservation” and “global consistency”, which precisely match the feature requirements of object detection and semantic segmentation in intelligent classrooms: object detection requires accurate local features to locate students, while semantic segmentation requires continuous global features to recognize boundary lines. In addition, CSPDarknet, as a fusion of Darknet53 and CSPNet, further enhances the lightweight property and robustness. Its multiple CSP modules can progressively extract features from low-level to high-level, providing basic features adapted to different tasks for the subsequent dual decoder branches. Figure 3 shows the backbone network structure diagram of the constructed model.

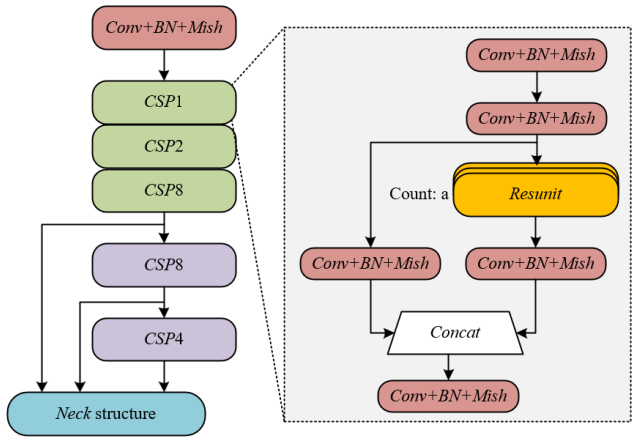

Figure 3. Backbone network structure diagram of the constructed model

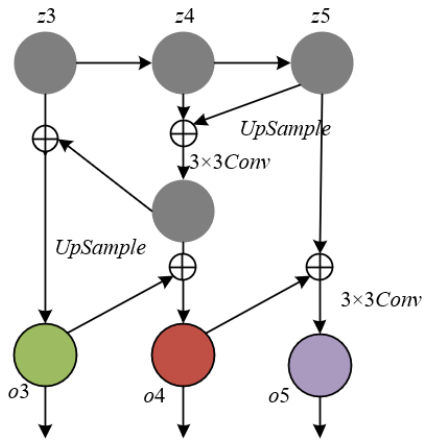

Figure 4. Structure diagram of the BiFPN used

From the task characteristics perspective, object detection for “whether students are studying attentively in their seats” needs to deal with targets of different scales, such as clear postures of front-row students and small-scale silhouettes of back-row students, while semantic segmentation for “classroom boundary line detection” needs to consider both fine-grained edges at the junction of desks and floors, and global region division of learning and non-learning areas. Both types of tasks rely on the effective fusion of multi-scale features. Therefore, this paper chooses BiFPN as the network Neck. BiFPN, as an improved version of PANet, happens to meet this requirement: its bidirectional cross-scale connection design can aggregate features of different levels output from the backbone network, including low-level texture features, mid-level shape features, and high-level semantic features, generating fusion features with both detail and global information. This provides multi-scale basis for judging student postures in object detection, and supplements continuous edge features of boundary lines for semantic segmentation. Meanwhile, BiFPN's weighted feature fusion mechanism and lightweight structure can control time cost while ensuring fusion effect, avoiding the decline of real-time performance due to complex feature processing, which fits the real-time monitoring scenario of intelligent classrooms, such as rapid response to dynamic changes in student behavior and

environmental status during class. Figure 4 shows the structure diagram of the BiFPN used.

The basic principle of BiFPN in supporting object detection and semantic segmentation originates from the synergy of its “bidirectional feature flow” and “weighted fusion mechanism”. At the feature flow level, BiFPN realizes bidirectional complementarity of cross-scale features through repeated top-down connections from high-level semantic features to low-level detail features and bottom-up connections from low-level detail features to high-level semantic features: the top-down path can inject semantic information such as “learning area” and “student identity” into low-level features, improving the regional correlation of boundary line segmentation and better distinguishing between “student seat area boundary” and “non-learning area boundary”; the bottom-up path can integrate detail features such as “edge texture” and “local contour” into high-level features, enhancing the detection ability of object detection for small-scale students or ambiguous postures, including downward head movements of back-row students. At the fusion mechanism level, the weight allocation mechanism similar to attention mechanism introduced by BiFPN can dynamically adjust the contribution of features from different sources. For example, in boundary line detection, it enhances the weight of low-level edge features, and in student posture judgment, it increases the proportion of mid-level shape features, making the fused features more adapted to the requirements of specific tasks.

Specifically, from the scenario requirements of the intelligent classroom, the detection branch needs to perform object localization and classification of "whether students are attentively studying in their seats", which belongs to sparse object detection and requires attention to the position and category of discrete targets. The segmentation branch needs to realize pixel-level classification of "boundary line and background", which belongs to dense pixel classification and requires attention to the pixel attribution of continuous regions. If a single branch is used to handle both types of tasks, the conflict of optimization objectives will lead to a decline in accuracy for both tasks. Based on the essential differences in requirements between object detection and semantic segmentation tasks, this paper adopts a separated design with two independent decoder branches to ensure that each branch optimizes the task characteristics in a targeted manner and avoids "task interference" during the feature processing.

The design principle of the detection branch focuses on the demand of "sparse object multi-scale localization", realizing accurate detection through the PAN structure and anchor mechanism. In the intelligent classroom, students may appear in different positions such as the front row, back row, or corners, with significant differences in object scale. This requires the detection branch to have multi-scale adaptation capability. The bottom-up feature migration property of the PAN structure can transmit low-level precise positional features to the high level, compensating for the deficiency of traditional feature pyramids that have strong semantics but weak localization, and providing accurate spatial reference for locating targets at different scales. The multi-scale feature maps with 8×, 16×, and 32× downsampling can match small, medium, and large-scale objects respectively, and combined with the grid allocation strategy of three types of anchor boxes, can cover the common scale range of students in classroom scenarios. In addition, the design of prediction parameters such as position offset, scale transformation, and class probability directly corresponds to the core requirements of "locating student position" and "judging whether the student is studying attentively", making the detection results directly usable for the subsequent judgment of "whether the student is within the learning area". Specifically, assume that the predicted bounding box is represented by (*ya*, *yb*, *yq*, *yg*), the coordinates of the top-left corner of the grid are represented by (*za*, *zb*), and the width and height of the anchor box assigned to the grid are represented by (*oq*, *og*). The transformation formulas are as follows:

$$ya = \left(2 * SIGMOID\left(sa\right) - 0.5\right) + z_a \tag{1}$$

$$yb = \left(2 * SIGMOID\left(sb\right) - 0.5\right) + z_b \tag{2}$$

$$yq = O_Q * e^{sq} \tag{3}$$

$$yg = O_g * e^{sg} \tag{4}$$

$$SIGMOID\left(a\right) = \frac{1}{1+e^{-a}} \tag{5}$$

The design principle of the segmentation branch is centered on "lightweight and efficient pixel-level classification", adapting to boundary line detection needs through a simplified process and targeted upsampling strategy. The core of boundary line detection in intelligent classrooms is to distinguish "boundary line pixels" from "background pixels", without requiring complex semantic category classification. Therefore, the segmentation branch adopts a lightweight design of "3 times upsampling + nearest neighbor interpolation": selecting the lowest 8× downsampled 32×32 feature map from the Neck as input, which retains basic edge features of boundary lines, avoids detail loss caused by excessively high downsampling rates, and controls the size of the initial feature map to reduce computation. Three times of upsampling restore the feature map to the size of the input image, ensuring that the output can correspond one-to-one with the pixels of the original image, meeting the accuracy requirements of "pixel-level segmentation". Nearest neighbor interpolation is chosen instead of deconvolution, significantly reducing computational complexity at the cost of a small amount of detail accuracy, avoiding delay in overall model real-time performance due to the segmentation process. The final output feature map of dimension (*W*, *H*, 2) directly provides the probability distribution of “boundary line/background”, which can provide clear pixel-level basis for “learning area division”, forming spatial correlation with the student position information from the detection branch.

2.3 Model loss function

The constructed model loss consists of the detection head loss and the segmentation head loss. The loss of the detection head mainly comes from classification loss ($loss_{CL}$), confidence loss ($loss_{OBJ}$), and bounding box regression loss ($loss_{BOX}$), while the boundary line segmentation head uses weighted cross-entropy loss. The expression of the loss function of the detection head is as follows:

$$loss_{DET} = \beta_1 loss_{CL} + \beta_2 loss_{OBJ} + \beta_3 loss_{BOX} \tag{6}$$

In intelligent classroom scenarios, samples of students "studying attentively" account for a very high proportion in the collected data, while samples of "not studying attentively" account for a very low proportion, forming a typical long-tail distribution. If traditional loss functions are used, the model will be dominated by the losses of majority class samples during training, leading to weak recognition ability for minority class samples. In order to solve the problem of "long-tail distribution of sample classes" in intelligent classroom learning behavior detection and achieve accurate balance of class loss, the classification and confidence loss adopt the design principle of EFL loss. EFL, as an improved version of Focal Loss, dynamically adjusts the loss weights of samples from different classes, which can suppress overfitting of majority classes and strengthen the loss signals of minority classes, making the model pay more attention to rare but important abnormal behavior samples during training. At the same time, using EFL for confidence loss can improve the reliability of judgment on "whether the student is in the seat": for ambiguous samples such as students half-standing or occluded, EFL can force the model to focus on hard samples' feature learning by reducing the weight of easy samples, so that classification and confidence results better match the core requirement of "recognizing learning behavior status". Let x_s represent the balance between positive and negative samples, o_s represent the confidence score of the predicted target. The Focusing Factor in balanced data scenarios that controls the basic behavior of the classifier is represented by parameter ε_y. The cumulative gradient ratio of the positive and negative samples of class k is represented by parameter h^k, and the scaling factor is represented by hyperparameter t. The expression of the EFL loss function is as follows:

$$EFL(o_s) = -\sum_{k=1}^{z} \beta_s \left(\frac{\varepsilon_y + \varepsilon_n^k}{\varepsilon_y} \right) (1-o_s)^{\varepsilon_y + \varepsilon_n^k} \log(o_s) \tag{7}$$

$$\varepsilon_n^k = t\left(1-h^k\right) \tag{8}$$

The bounding box regression loss chooses Expected Intersection over Union (EIOU) and integrates the design principle of Focal Loss, aiming to improve the accuracy and convergence efficiency of student position localization, adapting to the detection needs of dynamic targets in intelligent classrooms. The key premise of learning behavior detection is accurate localization of students in their seats. The predicted bounding box must highly match the actual student position, which requires the regression loss to comprehensively measure the difference between the predicted box and the target box. EIOU is designed with three components: "overlap loss, center distance loss, width-height loss". Compared with Complete Intersection over Union (CIOU), it more directly optimizes box shape and position. In intelligent classrooms, this can quickly reduce the gap between the predicted box and the actual position of the student and accelerate model convergence. After integrating Focal Loss, the loss function can reduce the optimization weight of low-overlap anchor boxes and focus training on high-overlap anchor boxes, further improving localization accuracy, which is crucial for "judging whether the student is in the seat". Suppose the width and height of the minimum enclosing box covering both the predicted and ground truth boxes are denoted as z_q and z_g, the centers of the predicted and ground truth boxes are denoted by y and y^{hs}, ϑ calculates the Euclidean distance between the two centers, and the width and height of the predicted and ground truth boxes are represented by q, q^{hs} and g, g^{hs} respectively. The calculation formula for CIOU is as follows:

$$\begin{aligned} loss_{RUPI} &= loss_{UPI} + loss_{DIC} + loss_{ASP} \\ &= 1 - IoU + \frac{\vartheta^2\left(y, y^{hs}\right)}{\left(z_q\right)^2 + \left(z_g\right)^2} \\ &+ \frac{\vartheta^2\left(q, q^{hs}\right)}{\left(z_q\right)^2} + \frac{\vartheta^2\left(g, g^{hs}\right)}{\left(z_g\right)^2} \end{aligned} \tag{9}$$

By integrating EIOU Loss and Focal Loss, and letting λ be the hyperparameter that controls the curvature of the loss curve, the final EIOU Loss expression is:

$$nloss_{F-E} = IoU^{\lambda} loss_{RUPI} \tag{10}$$

The overall loss function design of the detection head serves the core goal of "accurate detection of learning behavior" through the synergy of "class balance" and "localization optimization." The EFL design of classification and confidence loss ensures that the model can effectively distinguish between "serious learning" and "non-serious learning" in a long-tailed sample distribution, avoiding missed detections of abnormal behaviors due to class imbalance. The EIOU+FocalLoss design for bounding box regression ensures the accuracy and stability of student position localization, providing reliable spatial reference for determining whether students are seated. The combination of both enables the detection head output to accurately reflect the core features of learning behavior and to link with the subsequent boundary line segmentation results, ultimately achieving the dual-precision monitoring of "learning behavior status + spatial position" in smart classrooms.

The segmentation head for boundary line detection adopts weighted cross-entropy loss, whose core principle is to specifically address the extreme imbalance between "boundary line pixels and background pixels" in the smart classroom scenario by forcing the model to focus on critical boundary line pixels through a weighting mechanism. In smart classroom images, boundary lines usually appear as thin elongated lines, accounting for less than 5% of the total pixels, while background pixels occupy an overwhelmingly large proportion. If the original cross-entropy loss is used, the model will overfit the background features due to the numerical advantage of background pixels, causing the classification errors of boundary line pixels to be diluted. This manifests as broken, blurred, or "swallowed" boundary lines in the segmentation results, directly affecting the accurate division of learning regions. The weighted cross-entropy loss assigns higher loss weights to boundary line pixels, significantly increasing their share in the total loss: when the model misclassifies boundary line pixels, it incurs higher loss values, forcing the model to enhance its learning of low-proportion boundary line features during training. This ensures the integrity and clarity of boundary line segmentation and provides reliable pixel-level support for distinguishing between "learning areas and non-learning areas." Assuming the balance factor is represented by q_Z, the total number of pixels in the image is denoted as V, and the total number of pixels in the Z-th category of the foreground is denoted as V_Z,

the loss function is expressed as:

$$loss_{mm-SEG} = -\sum_{z=1}^{L} q_Z b_z \log(o_z) \tag{11}$$

$$q_z = \frac{V - V_Z}{V} \tag{12}$$

The design of this loss function further serves the overall research goal of "linking environmental monitoring with learning behavior analysis" by improving segmentation precision as a foundation for spatial relationship judgment. One of the core requirements in smart classrooms is to perform spatial inference by combining the "boundary line segmentation result" with the "object detection result", which demands high spatial accuracy in boundary line segmentation. If boundary line positions deviate or are missing, it directly leads to region segmentation errors, resulting in behavioral judgment deviations. Weighted cross-entropy loss, while balancing pixel loss, retains the natural suitability of cross-entropy loss for pixel-level classification: by minimizing the error between "network output probabilities of boundary line/background" and "ground truth pixel labels", it ensures each pixel's classification result closely reflects the actual scene.

Assuming the parameters used to balance the detection head and segmentation head losses are represented by δ_1 and δ_2 respectively, the overall loss function of the model is expressed as:

$$loss_{ALL} = \delta_1 loss_{DET} + \delta_2 loss_{mm-SEG} \tag{13}$$

Figure 5 shows the flowchart for judging the smart classroom environment and student learning status.

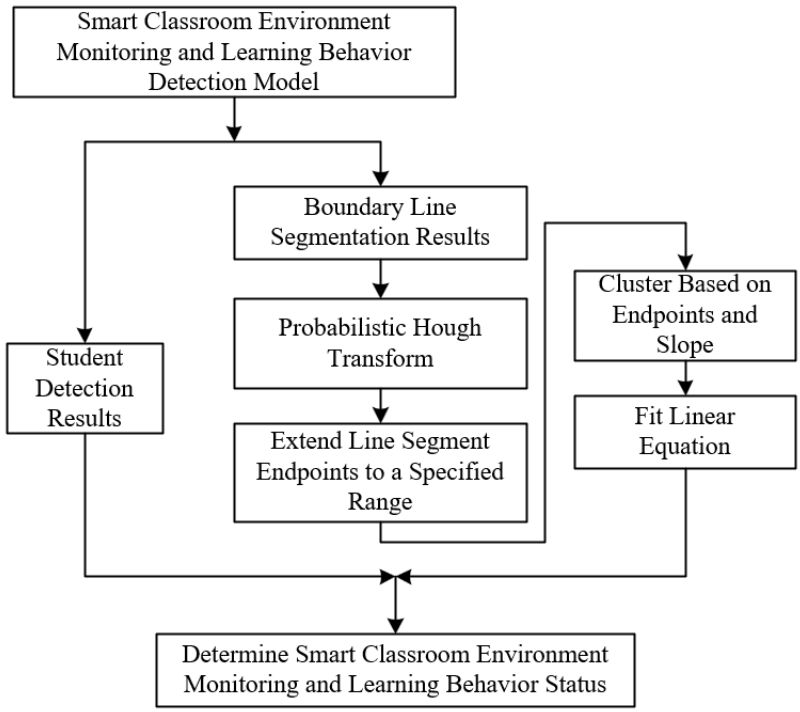

Figure 5. Flowchart for judging smart classroom environment and student learning status

3. EXPERIMENTAL RESULTS AND ANALYSIS

From the experimental results in Table 1, it is clearly observed that the proposed "shared Encoder + dual independent Decoder branches" collaborative model exhibits significant advantages: In terms of the detection task, the Recall of the dual-branch model is 91.2%, which is 6 percentage points higher than the single detection branch's 85.2%; the mAP50 is 77.2%, 3.8 percentage points higher than the 73.4% of the detection-only branch. This indicates that under the dual-branch collaboration, the environmental boundary constraints from the segmentation branch complement the target features of the detection branch, significantly reducing missed detections and localization errors. In terms of the segmentation task, the Accuracy of the dual-branch model is 71.2%, an increase of 5.8 percentage points compared to the 65.4% of the segmentation-only branch; the IOU is 34.5%, 5.9 percentage points higher than the 28.6% of the single segmentation branch. This confirms that the target location information from the detection branch provides guidance for boundary segmentation, enhancing pixel classification accuracy for dim boundary lines through prior knowledge of object and background regions. At the same time, BiFPN's dynamic weight allocation mechanism enhances the fusion of multi-scale features, and the improved EFL loss alleviates the sample imbalance between boundary pixels and background, further amplifying the synergistic gain of the dual-task design. Although the dual-branch model has lower speed due to increased computation, its comprehensive accuracy improvements fully verify the rationality of the “shared feature extraction + dual-task collaborative inference” architecture and the effectiveness of feature enhancement and loss function optimization strategies.

Table 1. Training results of the model under different branch configurations

Training Method	*Recall* (%)	*mAP*50 (%)	*Accuracy* (%)	*IOU* (%)	*Speed*(*fps*)
Detection Branch	85.2	73.4	-	-	34.8
Segmentation Branch	-	-	65.4	28.6	37.2
Dual-Branch Model	91.2	77.2	71.2	34.5	24.6

Table 2. Ablation experiment comparison

Algorithm Model	Parameters	Precision	Recall	mAP	Model Size (MB)
Replace BiFPN with standard FPN Neck	7124536	0.889	0.914	0.935	13.6
Use standard Focal Loss for classification/confidence loss, EIoU for regression loss	4758265	0.875	0.928	0.948	11.4
Independent Encoder + dual Decoder	5862452	0.912	0.916	0.952	13.8
Complete model	4896523	0.915	0.938	0.956	11.2

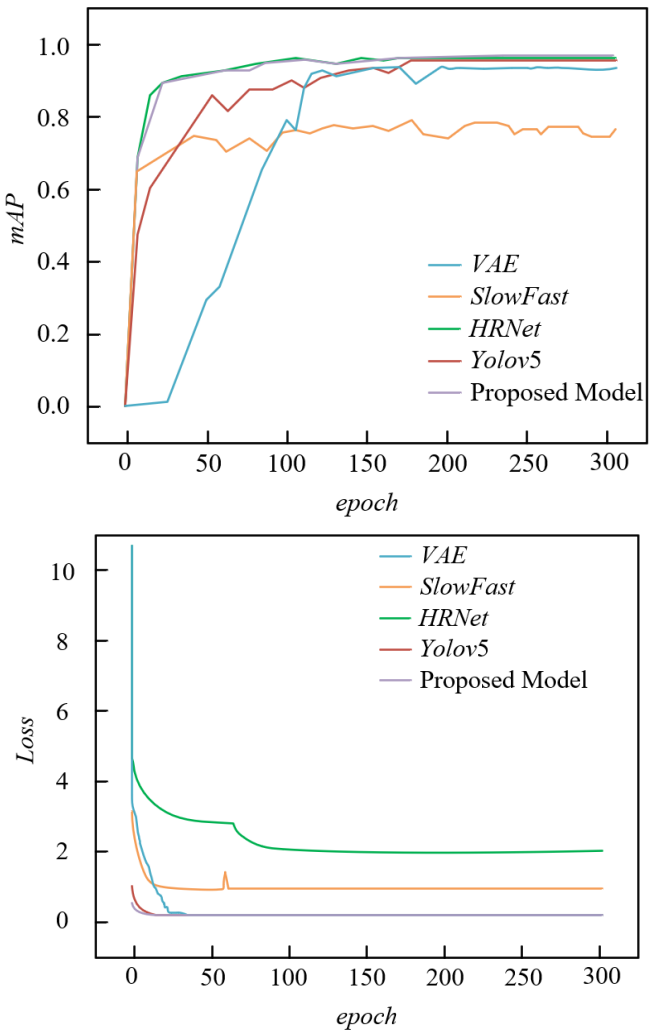

Figure 6. Comparison of mAP and loss values during training for different models

From the results of the ablation experiments in Table 2, it can be seen that the proposed complete model using shared Encoder + BiFPN + improved EFL loss shows significant advantages across multiple dimensions. The "Independent Encoder + dual Decoder" variant has 5,862,452 parameters and a size of 13.8MB, while the complete model reduces parameters by 16.5% and model size by 18.8%, and achieves better precision, recall, and mAP. This verifies the design value of the shared bottom-level Encoder: by reusing common features of environmental boundaries and learning behaviors, it avoids redundant parameters in dual-task independent encoding, and leverages the synergy of "spatial constraints from segmentation assisting detection" and "object prior from detection optimizing segmentation" to improve dual-task accuracy. When BiFPN is replaced with a standard FPN, parameters increase to 7,124,536 and size to 13.6MB, but precision, recall, and mAP drop significantly. BiFPN's dynamic weight allocation mechanism enables more efficient multi-scale feature fusion. For difficult-to-distinguish features such as "dim boundaries" and "small-scale learning behaviors" in smart classrooms, it enhances the weights of key channels and spatial regions, boosting feature capture while reducing model complexity. Replacing classification loss with standard Focal Loss alone results in precision dropping from 0.915 to 0.875, and mAP from 0.956 to 0.948. Although recall slightly increases, overall accuracy declines. This indicates that the improved EFL focal loss, with its "adaptive difficult-sample weight adjustment" mechanism, effectively alleviates the sample imbalance issue between sparse boundary pixels and numerous background pixels. By assigning higher loss weights to sparse boundary pixels, it prevents them from being overwhelmed by the background, improving classification confidence and overall accuracy.

From the training curve comparison in Figure 6, the proposed model exhibits core advantages of fast convergence, high accuracy ceiling, and strong training stability, which can be deeply interpreted from the model design. In terms of convergence speed, the proposed model rapidly exceeds 0.8 mAP within the first 50 epochs and continues to rise, far surpassing VAE and Yolov5. This is due to the dual-task collaborative mechanism of the shared Encoder: spatial boundaries from segmentation and object features from detection share bottom-level encoding, allowing the model to quickly learn effective features early on via complementary learning of "segmentation constraining detection" and "detection guiding segmentation". Regarding accuracy ceiling, the proposed model's final mAP approaches 1.0, significantly better than SlowFast and HRNet. This benefits from BiFPN's dynamic weight enhancement: for difficult scenarios such as "dim boundaries" and "small-scale behaviors" in smart classrooms, BiFPN dynamically adjusts channel and spatial weights via an attention-like mechanism, enhancing multi-scale feature fusion accuracy and breaking the performance ceiling of single-task models. For the loss curve, the proposed model's loss drops rapidly to near 0 within 50 epochs. Compared to VAE and HRNet, this reflects the synergy of improved EFL loss and shared structure: improved EFL adaptively adjusts loss weights for hard and easy samples, solving the sample imbalance of "few boundary pixels and many background pixels", preventing gradient

explosion/vanishing; the shared Encoder reduces parameter redundancy and training complexity, accelerating loss convergence. In terms of stability, the proposed model's loss exhibits almost no fluctuation in the later stages, whereas Yolov5 and SlowFast show oscillations. This is due to the task decoupling design of the dual Decoder branches: environmental segmentation and behavior detection have independent Decoders after the shared Encoder, each optimizing task-specific features, reducing gradient interference between the two tasks. Meanwhile, BiFPN's dynamic weights stabilize the feature fusion process, making training smoother. In summary, from three dimensions—dynamic convergence, performance ceiling, and training robustness—the training curves verify the scientificity of the proposed model design.

Table 3. Experimental results on different datasets

Dataset	*Precision*	*Recall*	*mAP*	*FPS/s*
Classroom Video Dataset	0.78	0.72	0.77	12.5
COCO Dataset	0.81	0.91	0.81	13.2
STU-HCI Dataset	0.81	0.76	0.75	11.4
SUN RGB-D Dataset	0.93	0.93	0.95	22.8

From the experimental results on different datasets in Table 3, the performance advantages of the proposed model can be deeply analyzed: On the dedicated classroom scene Classroom Video Dataset, the model achieved Precision of 0.78, Recall of 0.72, and mAP of 0.77. Despite challenges such as uneven lighting and dim boundaries, BiFPN's dynamic weight allocation strengthened feature capture in boundary regions. Combined with shared Encoder's dual-task collaborative reasoning, effective detection was still achieved. On the multi-scale object COCO Dataset, Recall reached 0.91 and mAP 0.81, thanks to BiFPN's efficient fusion of multi-scale features and improved EFL loss's adaptive weighting of "small objects and blurry boundaries", demonstrating precise detection of multi-scale targets in classrooms. For fine-grained behavior scenarios on the STU-HCI Dataset, Precision was 0.81 and Recall 0.76, reflecting the value of the shared Encoder in assisting behavior region localization through environmental segmentation. Although fine-grained behavior still has room for improvement, it meets the core needs of behavior analysis in smart classrooms. On the generalized indoor scene SUN RGB-D Dataset, both Precision and Recall reached 0.93, with mAP 0.95 and FPS 22.8, verifying the shared Encoder's generalization ability for common indoor features and the efficiency gains from model lightweighting. In summary, from dedicated classrooms to generalized indoor scenes, from complex lighting to fine-grained behaviors, the proposed model achieves a balance between accuracy and efficiency through threefold design: shared Encoder collaborative reasoning, BiFPN feature enhancement, and improved EFL loss optimization, fully demonstrating the method's scientificity and effectiveness in smart classrooms and extended scenarios.

Table 4. P, R, mAP values of each learning behavior in different smart classroom environments

Smart Classroom Environment	Metric	Independent Learning	Interactive Discussion	Temporary Leaving Seat	Staying at Fixed Position	Movement Within Learning Area
Natural Light-Dominated	*Precision*	0.84	0.81	0.91	0.91	0.97
	Recall	0.88	0.97	0.82	0.84	0.91
	mAP	0.91	0.95	0.91	0.92	0.94
High-Density Interaction	*Precision*	0.82	0.66	0.87	0.87	0.95
	Recall	0.81	0.96	0.71	0.84	0.83
	mAP	0.88	0.91	0.83	0.92	0.92
Multimedia-Intensive	*Precision*	0.82	0.66	0.87	0.88	0.95
	Recall	0.87	0.97	0.74	0.82	0.85
	mAP	0.93	0.95	0.85	0.92	0.93
Dynamic Work Type	*Precision*	0.77	0.63	0.85	0.88	0.95
	Recall	0.78	0.95	0.71	0.77	0.81
	mAP	0.85	0.91	0.81	0.87	0.91

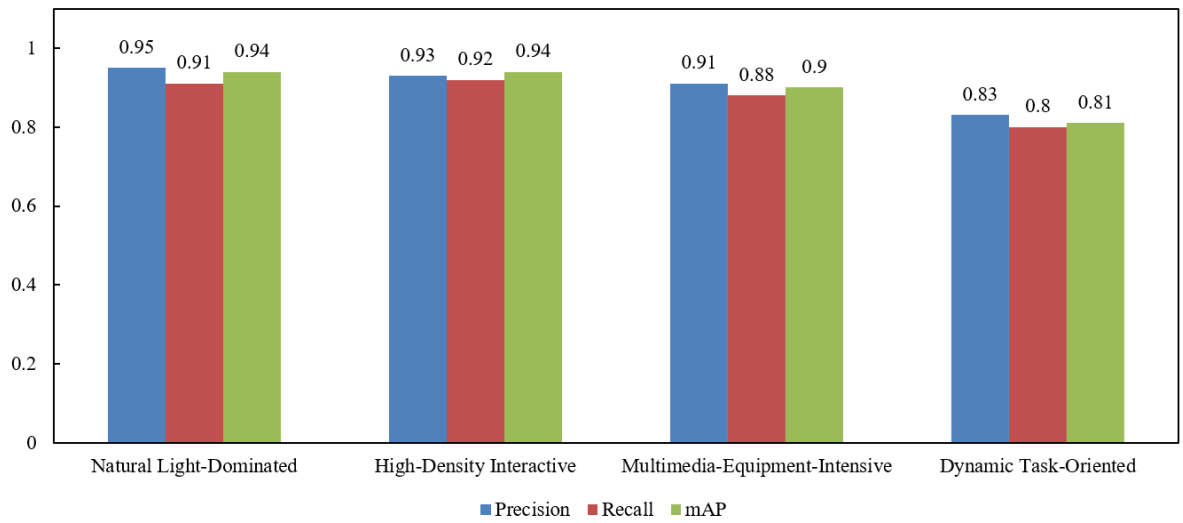

Figure 7. Comparison results under different thresholds in various smart classroom environments

From the detection results of multi-scenario learning behaviors in Table 4, the advantages of the proposed model in complex environment adaptability, behavioral feature distinguishability, and task collaborative robustness can be deeply analyzed based on model design logic. In natural light-dominated classrooms, the interactive discussion behavior achieved Recall of 0.97 and mAP of 0.95, far surpassing other behaviors. This is due to BiFPN's dynamic spatial attention adjustment: for dim boundaries caused by window glare and shadowed corners, BiFPN strengthens feature channel weights in low-light areas. Combined with shared Encoder's collaborative reasoning between environmental segmentation and behavior detection, it accurately captures posture associations during student interaction. Temporary leaving seat behavior achieved Precision of 0.91, relying on the seat area segmentation output by the shared Encoder, which assists the detection module in determining spatial boundaries of "leaving the seat", reducing misjudgments. In high-density interaction classrooms with crowded students and overlapping postures, independent learning behavior maintained stable mAP of 0.88. This benefits from the dual-task decoupling design of the shared Encoder: the segmentation Decoder outputs "seat regions", and the detection Decoder identifies "human targets", collaborating spatially to distinguish individual learning states among adjacent students. Although the Recall of interactive discussion behavior is only 0.66, its mAP of 0.91 reflects the improved EFL loss's ability to optimize hard samples: assigning higher loss weights to occluded interaction postures ensures overall classification accuracy. In multimedia-intensive classrooms facing screen glare and device occlusion, interactive discussion behavior still achieved Recall of 0.97 and mAP of 0.95, thanks to BiFPN's channel weight dynamic allocation: it automatically enhances feature channels highly correlated with interaction, such as "gestures" and "face orientation", while the shared Encoder's segmentation module accurately extracts "device contours" to define "non-interference zones" for behavior detection. The mAP of 0.92 for staying at a fixed position behavior depends on the segmentation module's accurate segmentation of fixed zones like "book corners" and "question areas", combined with the detection module's position tracking, achieving high-stability judgments. In dynamic work-type classrooms with intervention by workers or tools, movement within learning area behavior achieved excellent performance with Precision of 0.95 and mAP of 0.91. This is due to the shared Encoder's dual-task collaboration, quickly distinguishing "students" from "workers", and improved EFL loss adaptively increasing weights for fast-moving student targets, avoiding missed detections due to motion blur. Although the Precision (0.77) and Recall (0.78) of independent learning behavior are slightly lower, its mAP of 0.85 is still well above random guessing, proving that under dynamic interference, the model can still maintain core behavior detection robustness through the threefold collaboration of "spatial segmentation + feature enhancement + loss optimization".

In summary, the multi-scenario and multi-behavior data in Table 4 fully verify that the proposed model achieves dual improvements in precision and robustness for core behaviors such as "individual learning" and "interactive discussion" in complex smart classroom environments including natural light fluctuation, target density, device occlusion, and dynamic interference. This is accomplished through the dual-task collaboration of the shared Encoder, the dynamic feature enhancement of BiFPN, and the hard sample optimization of the improved EFL.

From the multi-environment metric comparison in Figure 7, the model's adaptability can be deeply analyzed in relation to its design. In natural light-dominated classrooms, the model achieves outstanding results with a Precision of 0.95, Recall of 0.91, and mAP of 0.94. This is mainly attributed to BiFPN's dynamic spatial attention adjustment, which enhances the feature channel weights in low-light boundary regions, such as window-side strong light and corner shadows. Combined with the collaborative inference of environmental segmentation and behavior detection by the shared Encoder, the spatial correlation of student postures is accurately captured, reducing both missed detections and false positives. In high-density interactive classrooms, the balanced performance of Precision 0.93, Recall 0.92, and mAP 0.94 stems from the decoupled dual-task design of the shared Encoder: the segmentation Decoder outputs the "seating area" to provide spatial constraints for the detection Decoder, effectively distinguishing independent behaviors of adjacent students. The improved EFL assigns higher weights to partially occluded interactive postures, ensuring a balance between classification accuracy and recall. In classrooms densely equipped with multimedia devices, although Recall decreases to 0.88, the Precision of 0.91 and mAP of 0.9 remain high. This is primarily due to BiFPN's dynamic channel weight allocation, which automatically enhances behavior-related channels such as "gesture actions" and "facial orientation", suppressing interference signals from screen reflections. The environmental segmentation module of the shared Encoder accurately marks "device-occupied areas", excluding non-learning targets from false detections and ensuring the reliability of core behavior classification. In dynamic-task classrooms, despite having the lowest results among the four scenarios—Precision 0.83, Recall 0.8, and mAP 0.81—these metrics are still significantly better than random guessing. The shared Encoder rapidly delineates "task areas" and "learning areas" through environmental segmentation, aiding the detection module in distinguishing between students and workers. The improved EFL addresses motion-blurred student targets, maintaining robustness in core behavior detection under dynamic interference. To summarize, through dual-task collaboration of the shared Encoder, dynamic feature enhancement by BiFPN, and hard sample optimization by improved EFL, the model achieves collaborative breakthroughs in precision, recall, and mAP across complex scenarios such as lighting fluctuation, target density, device occlusion, and dynamic interference. This fully verifies the scientific and practical value of the proposed method in diverse smart classroom environments.

4. CONCLUSION

This study, targeting the core demands of intelligent classroom environmental monitoring and learning behavior analysis, proposed a dual-branch collaborative model based on object detection and semantic segmentation. It achieved joint extraction of environmental and behavioral features through a shared Encoder, enhanced dim boundary line detection via BiFPN's dynamic weight mechanism, and alleviated pixel distribution imbalance using an improved EFL loss. Together, these formed a complete technical path of "feature sharing–enhancement–optimization". Experimental results

demonstrated that the model achieves high-precision environmental boundary segmentation and learning behavior detection in complex classroom scenarios, including natural light fluctuation, high-density interaction, dense device presence, and dynamic operations. It performed particularly well in dim boundary recognition and imbalanced sample scenarios. The research contributes in two main aspects: In technical aspect, it verified the effectiveness of a dual-task collaborative framework in classroom scenes, providing a reference for multi-task visual model design; In application aspect, it can output real-time joint analysis results of "environmental boundaries + behavior states", providing quantitative support for teachers in adjusting teaching strategies and for administrators in optimizing classroom resource allocation, thereby promoting the transition of intelligent education from "passive monitoring" to "active support".

However, the study still has three limitations: (1) Behavior analysis depends on location and region information, resulting in insufficient accuracy for fine-grained behavior recognition. (2) Model inference delay slightly increases in dynamic scenarios. (3) Generalization relies on annotated classroom scene data, limiting adaptability to atypical classrooms. Future research can proceed in the following directions: Integrate pose estimation with object detection to refine behavior classification via keypoint features; Introduce lightweight network structures to optimize inference speed for edge-end real-time deployment; Adopt semi-supervised learning to reduce annotation dependence, and utilize multi-modal data to enhance robustness in complex scenes, further expanding application boundaries in smart education scenarios.

REFERENCES

[1] Kwet, M., Prinsloo, P. (2020). The 'smart'classroom: A new frontier in the age of the smart university. Teaching in Higher Education, 25(4): 510-526. https://doi.org/10.1080/13562517.2020.1734922

[2] Huang, L.S., Su, J.Y., Pao, T.L. (2019). A context aware smart classroom architecture for smart campuses. Applied Sciences, 9(9): 1837. https://doi.org/10.3390/app9091837

[3] Saini, M.K., Goel, N. (2019). How smart are smart classrooms? A review of smart classroom technologies. ACM Computing Surveys (CSUR), 52(6): 130. https://doi.org/10.1145/3365757

[4] Selim, H.M., Eid, R., Agag, G. (2020). Understanding the role of technological factors and external pressures in smart classroom adoption. Education+ Training, 62(6): 631-644. https://doi.org/10.1108/ET-03-2020-0049

[5] Idrees, H., Shah, M., Surette, R. (2018). Enhancing camera surveillance using computer vision: A research note. Policing: An International Journal, 41(2): 292-307. https://doi.org/10.1108/PIJPSM-11-2016-0158

[6] Tribley, J., McClain, S., Karbasi, A., Kaldenberg, J. (2011). Tips for computer vision syndrome relief and prevention. Work, 39(1): 85-87. https://doi.org/10.3233/WOR-2011-1183

[7] Martynenko, A. (2017). Computer vision for real-time control in drying. Food Engineering Reviews, 9(2): 91-111. https://doi.org/10.1007/s12393-017-9159-5

[8] Chella, A., Frixione, M., Gaglio, S. (2001). Conceptual spaces for computer vision representations. Artificial Intelligence Review, 16(2): 137-152. https://doi.org/10.1023/A:1011658027344

[9] Gurevich, I.B., Yashina, V.V. (2022). On modeling descriptive image analysis procedures on a specialized Turing machine. Pattern Recognition and Image Analysis, 32(3): 469-476. https://doi.org/10.1134/S1054661822030142

[10] Gurevich, I.B., Yashina, V.V. (2024). Multialgorithmic hierarchical image analysis system: Architecture and analysis model. Pattern Recognition and Image Analysis, 34(4): 959-965. https://doi.org/10.1134/S1054661824700949

[11] Bi, Y., Xue, B., Mesejo, P., Cagnoni, S., Zhang, M. (2022). A survey on evolutionary computation for computer vision and image analysis: Past, present, and future trends. IEEE Transactions on Evolutionary Computation, 27(1): 5-25. https://doi.org/10.1109/TEVC.2022.3220747

[12] Zhao, J., Wang, G., Zhou, B., Ying, J., Liu, J. (2023). SRA–CEM: An improved CEM target detection algorithm for hyperspectral images based on subregion analysis. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 16: 6026-6037. https://doi.org/10.1109/JSTARS.2023.3289943

[13] Wang, L., Dong, J., Wang, L., Liu, F., Wen, Q., Genc, B. (2025). A decoupled segmentation-classification strategy based on semantic-SAM for precise semantic segmentation in coal mine areas. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing., 18: 14820-14842. https://doi.org/10.1109/JSTARS.2025.3576285

[14] ChaitandasHadke, S., Mishra, R., Bankar, R.T., Chhabria, S.A., Chavate, S.P., Pinjarkar, L.S. (2025). An attention driven long short term memory based multi-attribute feature learning for shot boundary detection. Knowledge-Based Systems, 317: 113379. https://doi.org/10.1016/j.knosys.2025.113379

[15] Li, X., Liu, J., Lin, Z., Liu, X., Wang, Y., Zhang, S., Ye, B. (2024). Advancing RFID technology for virtual boundary detection. IEEE Transactions on Mobile Computing, 24(4): 3407-3422. https://doi.org/10.1109/TMC.2024.3514895

[16] Zhang, S.X., Yang, C., Zhu, X., Yin, X.C. (2023). Arbitrary shape text detection via boundary transformer. IEEE Transactions on Multimedia, 26: 1747-1760. https://doi.org/10.1109/TMM.2023.3286657

[17] Khalique, V., Kitagawa, H., Amagasa, T. (2023). BPF: a novel cluster boundary points detection method for static and streaming data. Knowledge and Information Systems, 65(7): 2991-3022. https://doi.org/10.1007/s10115-023-01854-1

[18] Patro, K.A.K., Acharya, B. (2023). An efficient two-level image encryption system using chaotic maps. International Journal of Information and Computer Security, 21(1-2): 35-69. https://doi.org/10.1504/IJICS.2023.131092

[19] Terhörst, P., Huber, M., Damer, N., Kirchbuchner, F., Raja, K., Kuijper, A. (2023). Pixel-level face image quality assessment for explainable face recognition. IEEE Transactions on Biometrics, Behavior, and Identity Science, 5(2): 288-297. https://doi.org/10.1109/TBIOM.2023.3263186