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

Traitement du Signal
Vol. 43, No. 3, June, 2026, pp. 1227-1241
Journal homepage: http://iieta.org/journals/ts

Joint Optimization of Multi-Object Detection and Fine-Grained Image Representation Learning in English Education Scenarios

Jingyao Zhang

Department of Open Education, Xingtai Open University, Xingtai 054000, China

Corresponding Author Email: xtzjy19820415@163.com

https://doi.org/10.18280/ts.430312

Received: 2 November 2025
Revised: 27 April 2026
Accepted: 12 June 2026
Available online: 30 June 2026

Keywords:
fine-grained image representation, multi-task synergistic learning, feature alignment, gradient harmonization, classroom behavior analysis, contrastive learning

ABSTRACT

Against the backdrop of educational digitalization, computer vision has become a key technology for intelligent classroom analysis in smart English education. English classroom scenes are characterized by densely distributed targets and highly confusable behaviors, requiring simultaneous multi-object localization and fine-grained behavior recognition. Although multi-object detection and fine-grained image representation are inherently complementary, existing methods are limited by insufficient dynamic feature adaptation, inadequate hierarchical semantic constraints, and optimization conflicts during multi-task learning. To address these issues, an end-to-end collaborative localization-representation network was proposed, establishing an integrated multi-task optimization loop that unifies feature synergy, representation shaping, and gradient harmonization. A dynamic semantic-detail collaborative modulation mechanism was first introduced, incorporating scale-aware routing and deformable feature alignment to enable bidirectional adaptive fusion and interaction between global localization semantics and local detail textures. Subsequently, a hierarchical semantic alignment contrastive learning framework was constructed, leveraging the inherent multi-level semantic structure and cross-level consistency constraints of classroom scenes to optimize the fine-grained feature manifold distribution, effectively suppress semantic drift, and enhance discriminability for similar behaviors. Finally, a curriculum-based gradient harmonization strategy was designed to systematically resolve optimization conflicts in joint multi-task training through gradient direction correction and dynamic task weight scheduling, ensuring training stability and convergence efficiency. Extensive comparative and ablation experiments were conducted on a self-constructed English classroom fine-grained behavior dataset and general fine-grained image datasets. Results demonstrated that the proposed method simultaneously improved multi-object detection and fine-grained behavior classification performance, achieving superior recognition accuracy and training robustness compared to existing mainstream approaches. This study not only provides an efficient technical solution for intelligent classroom behavior analysis in smart education scenarios but also offers a novel paradigm reference for feature synergistic learning and joint optimization in multi-task image understanding.

1. INTRODUCTION

As the digital transformation of education continues to advance [1-3], intelligent teaching systems have progressively shifted from experience-driven subjective pedagogical evaluation toward data-driven intelligent analysis of learning conditions. Classroom perception technologies grounded in computer vision [4-6] enable non-intrusive, real-time acquisition and quantitative analysis of teacher-student behaviors and instructional states, constituting a core technological foundation for supporting smart English teaching reform and enabling precise pedagogical interventions. English language classrooms [7-9] are characterized by dense target distributions, multidimensional behavioral granularity, and subtle inter-class differences—typical scenario features that demand not only accurate spatial localization of multiple object categories—including teachers, students, and instructional media—but also effective discrimination among highly similar learning and interactive behaviors, thereby achieving fine-grained behavioral state recognition [10]. From the perspective of visual cognitive mechanisms, multi-object detection relies on global semantic and spatial structural information for target localization, whereas fine-grained recognition depends on local texture and edge details for precise categorization; the feature representations of these two tasks exhibit inherent complementarity [11-13]. Existing approaches, however, predominantly adopt task cascading or simple backbone parameter sharing, which fails to achieve bidirectional deep interaction and adaptive synergy at the feature level. This limitation not only constrains the upper bound of localization accuracy in object detection but also impedes the full exploitation of the discriminative potential inherent in fine-grained representations. The construction of a multi-task joint

optimization framework tailored to classroom scenarios [14-16] can address the practical challenges of fine-grained behavioral analysis in smart English education, enhance the intelligence of instructional state assessment, and simultaneously advance the theoretical foundations of feature alignment [17, 18], representation manifold construction, and joint optimization within the field of multi-task image understanding—thereby offering significant practical value and academic research significance.

Although existing fusion methods for multi-object detection and fine-grained recognition have achieved basic task coupling, fundamental deficiencies persist in feature interaction, representational constraints, and training optimization, rendering them inadequate for the complex visual task demands of English classrooms—characterized by high granularity, strong similarity, and multi-scale variability. In terms of feature interaction mechanisms [19, 20], mainstream approaches predominantly employ unidirectional feature injection and fixed-channel concatenation as the primary fusion forms, resulting in statically fixed feature interaction patterns that cannot adaptively adjust the fusion ratio between semantic and detailed features according to target scale size or texture information richness. Large-scale instructional targets fail to obtain sufficient detailed features to support fine-grained discrimination, whereas small-scale local behavioral targets are susceptible to fine-grained noise interference, which degrades localization stability. Consequently, the bidirectional dynamically adaptive synergistic interaction capability is absent. In fine-grained representation learning, traditional contrastive learning methods construct sample-wise constraint relationships only at a single semantic hierarchy [21], neglecting the inherent hierarchical subordinate properties of visual semantics. Such approaches fail to align with the multi-level semantic logic of classroom scenes—spanning from global scene context, to target subjects, and down to fine-grained behaviors—readily leading to blurred feature boundaries among similar behavioral subcategories. Concurrently, the fine-grained representation training process lacks upper-level semantic constraint guidance, rendering it prone to semantic drift, which results in insufficient self-consistency between classification outcomes and global scene semantics. At the multi-task optimization level [22], existing joint training paradigms commonly adopt fixed-weight loss fusion strategies, where parameter updates of the shared backbone network are susceptible to interference from adversarial gradients originating from the two tasks, inducing training oscillation and task performance degradation. Current gradient correction methods can only passively resolve local gradient conflicts, fail to form a closed-loop optimization mechanism coupled with the forward feature learning process, and do not conform to the coarse-to-fine progressive learning patterns inherent in visual cognition—thus fundamentally incapable of resolving the optimization imbalance problem in multi-task training.

To address the aforementioned technical deficiencies and scenario-specific challenges, an end-to-end collaborative localization-representation network is proposed, establishing a complete multi-task synergistic closed loop that encompasses feature interaction, representation shaping, and gradient optimization. A dynamic semantic-detail collaborative modulation module, a hierarchical semantic alignment contrastive learning module, and a curriculum-based gradient harmonization strategy are sequentially designed to tackle the core issues of multi-scale feature adaptive synergy, the absence of fine-grained semantic constraints, and multi-task gradient conflicts, respectively. Each innovative module is deeply coupled to form an integrated technical framework, which is comprehensively validated on a self-constructed English classroom dataset and general fine-grained datasets. The synchronized improvement of both detection and fine-grained recognition performance is effectively achieved, offering a novel technical paradigm and research direction for multi-task visual synergistic learning.

The overall organizational structure of this work is clear and coherent, with each chapter's research content progressing in a logical sequence. Chapter 2 elaborates on the overall architecture of the proposed collaborative localization-representation network, providing a complete derivation of the technical principles and implementation details for each core innovative module. Chapter 3 validates the effectiveness of the proposed method through multiple sets of comparative and ablation experiments, comprehensively analyzing model performance, representational capacity, and training stability. Chapter 4 objectively dissects the intrinsic mechanisms, existing limitations, and future extension directions of the method. Chapter 5 synthesizes the overall research findings and summarizes the core research conclusions and innovative contributions.

2. COLLABORATIVE LOCALIZATION-REPRESENTATION NETWORK METHODOLOGY

2.1 Overall network architecture

To reconcile the core tension between multi-task feature sharing and task-specific expression, a collaborative localization-representation network is constructed, with overall design principles established as bottom-level feature sharing, top-level task decoupling, and bidirectional information synergy. The network employs Cross Stage Partial Darknet (CSP-Darknet) as the feature extraction backbone, performing multi-scale feature encoding on input English classroom images to generate feature maps that integrate global spatial structure and local base textures, thereby providing unified, general-purpose bottom-level feature support for subsequent object detection and fine-grained behavioral representation tasks. Feature reuse efficiency is ensured while model computational redundancy is effectively controlled.

To address the insufficient task adaptability of shared features, two sets of lightweight task-specific adapters are introduced, which perform directed mapping of base features through convolutional and channel attention mechanisms, generating detection features suited for localization tasks and fine-grained features suited for fine-grained classification tasks, respectively. The two adapter types employ differentiated attention excitation logics: the detection branch strengthens channel-wise global semantic weights through global pooling to enhance the stability of target spatial localization, whereas the fine-grained branch activates edge and texture detail features through local pooling to reinforce the representational capacity for subtle behavioral differences, thereby establishing the complementary properties of the dual tasks at the feature source level. Cross-task synergistic interaction is conducted based on deep multi-scale features from the backbone network, comprehensively accommodating the feature optimization requirements of instructional targets

at different scales within classroom scenarios. The entire network is fully differentiable, contains no discrete decision nodes, and enables end-to-end joint iterative optimization of the dual tasks.

2.2 Dynamic semantic-detail collaborative modulation module

The dynamic semantic-detail collaborative modulation module is designed to establish a differentiable bidirectional information interaction pathway between the detection branch and the fine-grained representation branch, enabling adaptive complementary fusion of global semantic features and local detail features according to the intrinsic properties of each target. A set of candidate regions is first extracted from the detection branch output prior to non-maximum suppression, preserving the complete gradient propagation pathway to ensure end-to-end training feasibility. Region feature normalization extraction is performed using the Region of Interest Align (RoIAlign) operation based on bilinear interpolation. For any given candidate region, uniformly sized local representations are separately extracted from the two task-specific feature maps, with detection semantic features and fine-grained detail features corresponding to fixed-dimensional tensors, respectively. The standardized regional feature outputs provide a unified computational basis for subsequent cross-branch feature mapping and fusion, eliminating interaction barriers caused by target scale variations. Figure 1 illustrates the architecture of the dynamic semantic-detail collaborative modulation module.

To overcome the limitation that fixed fusion weights cannot adapt to multi-scale variations across scenarios, a sample-wise dynamic routing coefficient is constructed in this module to enable adaptive regulation of fusion intensity. The routing coefficient is generated through nonlinear mapping of multi-dimensional target attributes by a multilayer perceptron, with the computational expression given as:

$$\alpha_i=\sigma\left(MLP([\log(s_i),c_i,Var(f_i^F)])\right) \tag{1}$$

where, s_i denotes the pixel area of the candidate region, c_i represents the initial localization confidence output by the detection branch, $Var(f_i^F)$ characterizes the dispersion degree of fine-grained regional features, and σ denotes the Sigmoid activation function, which constrains the output value to the interval [0, 1]. The logarithmic transformation applied to the regional area alleviates the weight shift problem caused by the long-tail distribution of target scales in classroom scenes. Low-confidence candidate regions correspond to samples with ambiguous localization boundaries, for which supplementary local texture features are required to optimize boundary regression. Higher variance in fine-grained features indicates richer edge and texture information within the region, and the introduction of detail features for such samples simultaneously enhances both localization accuracy and category discriminability. The three types of variables jointly drive the routing coefficient to achieve sample-specific regulation of fusion intensity.

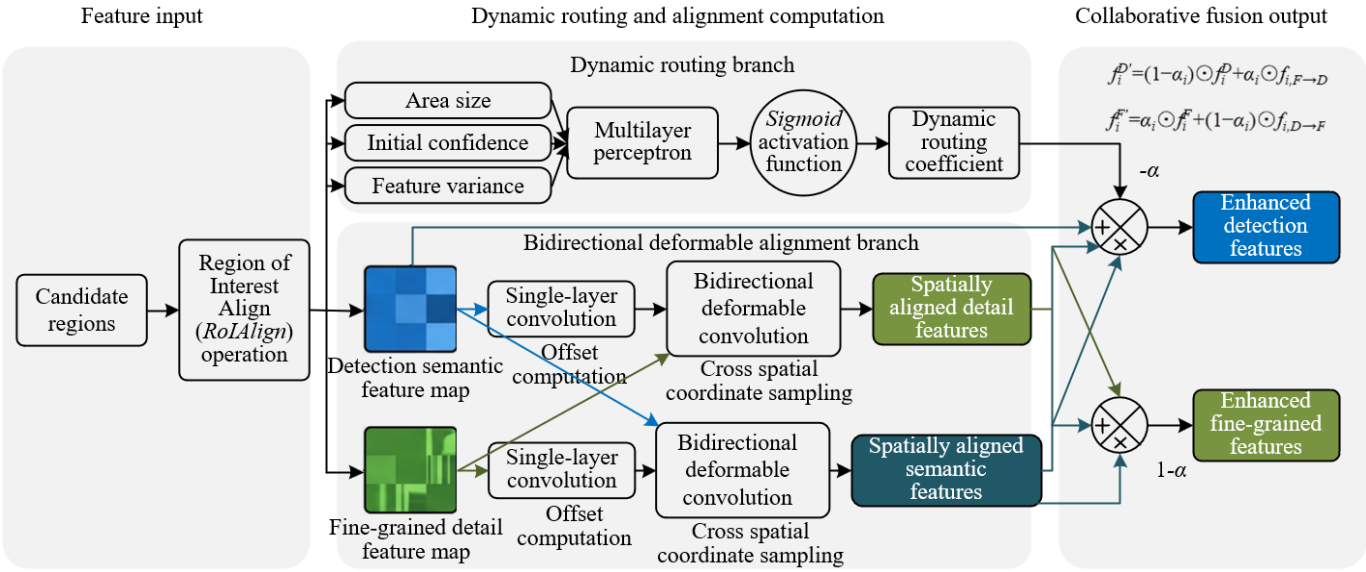

Figure 1. Architecture of the dynamic semantic-detail collaborative modulation module

Direct channel concatenation of the two types of regional features would introduce spatial pixel misalignment, thereby diminishing the performance gains achieved through cross-branch information fusion. To address this issue, bidirectional deformable convolution is introduced to accomplish precise spatial alignment of features. The two sets of bidirectional mapping computations are defined as:

$$f_{i,F\to D}=DeformConv(f_i^F,Offset(f_i^D)) \tag{2}$$

$$f_{i,D\to F}=DeformConv(f_i^D,Offset(f_i^F)) \tag{3}$$

The feature offsets are obtained through single-layer convolution regression from the corresponding guiding features, with the output dimensions maintained to match the number of deformable convolution sampling points. The offsets derived from detection features possess complete target contour perception capability, enabling adjustment of the sampling coordinates of fine-grained features to concentrate the sampling range on the target body region while filtering out interference from irrelevant background pixels. The offsets derived from fine-grained features exhibit greater sensitivity to object edge variations, enabling correction of the sampling positions in the detection branch and improving the precision

of boundary coordinate regression. The bidirectional alignment operation simultaneously optimizes the spatial representation quality of both tasks.

Based on the dynamic routing coefficient, a symmetric complementary feature fusion computation logic is constructed to respectively update the enhanced regional features of the detection branch and the fine-grained branch. The fusion computation formulas are given as:

$$f_i^{D'}=(1-\alpha_i)\odot f_i^{D}+\alpha_i\odot f_{i,F\to D} \tag{4}$$

$$f_i^{F'}=\alpha_i\odot f_i^{F}+(1-\alpha_i)\odot f_{i,D\to F} \tag{5}$$

where, $\odot$ denotes the element-wise tensor multiplication operation. When the routing coefficient approaches 1, corresponding to large-scale instructional targets with sufficient texture information, the detection branch introduces a substantial amount of aligned detail features to optimize boundary localization, while the fine-grained representation branch primarily relies on native texture features, supplemented by global semantic constraints to prevent representation deviation from the scene subject. When the routing coefficient approaches 0, corresponding to small-scale local behavioral targets with blurred textures, the detection branch relies on native global semantics to maintain localization stability, thereby avoiding boundary jitter caused by fine-grained texture noise, while the fine-grained branch introduces global semantic features to correct representation distribution shifts. The two branches, through symmetric weight allocation logic, form a collaborative representation system characterized by mutual constraints and bidirectional information exchange.

2.3 Hierarchical semantic alignment contrastive representation learning module

To address the challenges of distinguishing easily confusable samples and mitigating representational semantic drift in fine-grained behavioral recognition within English classroom scenarios, a Hierarchical Semantic Alignment Contrastive Representation Learning Module is constructed in this section, which shapes structured feature manifolds by leveraging inherent scene semantic priors. The visual content of English classrooms exhibits a rigorous hierarchical compositional pattern, which can be partitioned into three progressive semantic dimensions—scene, object, and attribute—corresponding to the cognitive logic from global environment to fine-grained behavior. This physical hierarchical relationship is embedded into the feature learning process, where a nested three-level semantic manifold structure is constructed. The higher-level semantic manifolds serve as constraint spaces for the underlying fine-grained features, ensuring that the learning process of fine-grained features is consistently guided by the upper-level semantic framework. From a geometric perspective, this design prevents feature representations from deviating from global semantics and effectively enhances the feature separability of similar behavioral subcategories. Figure 2 illustrates the mechanism of the hierarchical semantic alignment contrastive representation learning module.

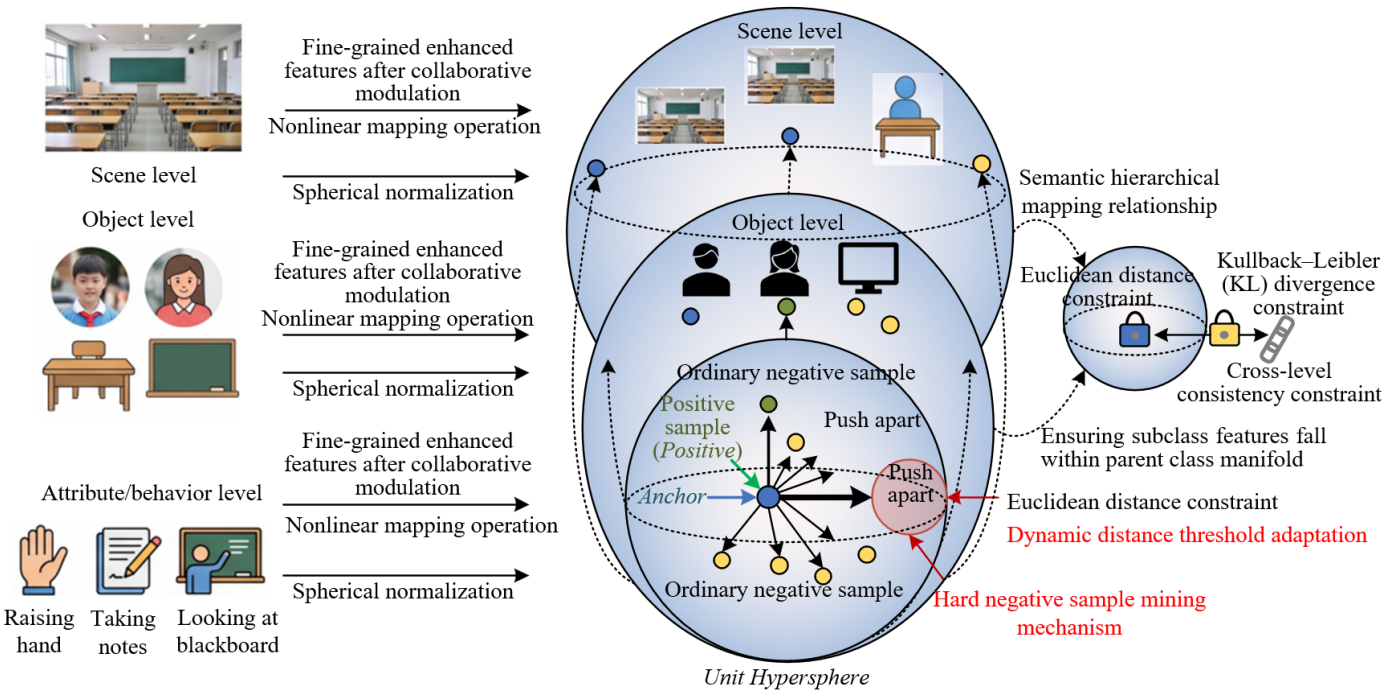

Figure 2. Mechanism of the hierarchical semantic alignment contrastive representation learning module

To achieve standardized representational output across multiple semantic levels, an independent nonlinear mapping unit is configured for each semantic dimension, which maps the enhanced features optimized by the preceding module onto a unified unit hypersphere space. The hierarchical semantic mapping formula is given as:

$$z_i^{(k)}=\frac{g_k(f_i^{F})}{\|g_k(f_i^{F})\|_2} \tag{6}$$

where, k denotes the semantic level index, $g_k(\cdot)$ represents the nonlinear mapping function composed of a two-layer

perceptron, f_i^F denotes the fine-grained enhanced features after dynamic collaborative modulation, and $z_i^{(k)}$ denotes the normalized hierarchical feature vector. The unified spherical normalization eliminates the interference of feature magnitude on similarity computation, rendering the vector inner product directly equivalent to cosine similarity and ensuring consistency of semantic measurement scales across all levels. Differentiated hyperparameters are configured to regulate the learning intensity at each semantic level. The temperature coefficient is progressively decreased with increasing semantic depth, accommodating the feature distribution characteristics where inter-class differences are pronounced at higher levels yet subtle at fine-grained levels. Concurrently, the loss weight at each level is progressively increased with increasing depth, focusing the core of model optimization on the fine-grained behavioral discrimination task.

Based on the standardized hierarchical semantic features, a hierarchical contrastive loss function is constructed to achieve simultaneous optimization of multi-granularity representations. The loss function expression is given as:

$$L_{HCL}=\sum_{k=1}^{3}\gamma_k\cdot\left[-\log\frac{\sum_{j\in P_k(i)}exp\left(z_i^{(k)}\cdot z_j^{(k)}/\tau_k\right)}{\sum_{j\in N_k(i)}exp\left(z_i^{(k)}\cdot z_j^{(k)}/\tau_k\right)}\right] \tag{7}$$

where, γ_k denotes the loss weight at each level, τ_k denotes the level-specific temperature coefficient, and $P_k(i)$ and $N_k(i)$ denote the positive sample set and negative sample set corresponding to the anchor sample at the k-th level, respectively. This loss function overcomes the limitations of traditional single-level contrastive learning, enabling simultaneous regularization of three types of feature distributions: global scene semantics, mesoscopic object subjects, and microscopic behavioral attributes. Constraints at higher levels ensure the accuracy of overall scene semantics, while constraints at lower levels continuously enlarge the feature margins among confusable fine-grained behaviors, achieving joint optimization of coarse-grained semantic consistency and fine-grained discriminability.

To eliminate semantic level conflicts arising from independent multi-level learning and to ensure the self-consistency and integrity of the representational system, a cross-level consistency constraint loss is introduced, which establishes associative mapping relationships between adjacent semantic levels. The specific expression is given as:

$$L_{Cross}=\sum_{k=2}^{3}E_i\left[\|z_i^{(k-1)}-Proj_{k-1}(z_i^{(k)})\|_2^2+KL(p_{k-1}|p_k)\right] \tag{8}$$

where, $Proj_{k-1}$ denotes the projection function mapping low-level features to the higher-level semantic space, and p_{k-1} and p_k denote the classification probability distributions of adjacent levels, respectively. The Euclidean distance constraint enables precise mapping of fine-grained features into the parent-class semantic space, ensuring that subclass features strictly reside within the parent-class manifold space and structurally preventing semantic level misalignment. The Kullback–Leibler divergence constraint regularizes the probability distribution differences between levels, maintaining logical consistency between coarse-grained and fine-grained classification outcomes, avoiding semantic contradictions in model outputs, and further reinforcing the stability and rationality of the overall representation space.

To further address the challenge of discriminating highly similar behaviors in English classrooms, a dynamic hard negative mining mechanism is introduced, which adaptively reinforces the model's learning capacity for confusable samples. During each training batch, feature distances between all anchor samples and negative samples are computed based on the feature space at the fine-grained attribute level, and the top ten percent of negative samples with the smallest distances are selected as hard samples. By elevating the contrastive loss weight of hard negative samples, the model's ability to capture subtle feature differences is enhanced. Concurrently, a dynamic distance threshold is configured to adapt to the training progression. In the early training stage, when the feature distribution has not yet converged, a relaxed threshold is adopted to expand the sample selection range, fully mining potential hard samples. In the later training stage, as features tend to become stable, the threshold is tightened to focus on extremely similar samples, achieving adaptive hard sample learning throughout the entire training process and continuously improving the discriminative accuracy of fine-grained behavioral representations.

2.4 Curriculum-based gradient harmonization optimization strategy

The gradient conflict problem in multi-task joint optimization constitutes a key bottleneck limiting the synergistic performance improvement of detection and fine-grained representation, with the emergence of gradient conflicts being directly related to the sharing mechanism of network parameters. In the dual-branch collaborative architecture proposed herein, the task-specific adapters and the detection and representation prediction heads receive supervisory gradients from only a single task, with parameter updates free from mutual interference; gradient conflicts are concentrated exclusively in the shared backbone parameters. Figure 3 illustrates the architecture and information flow diagram of the curriculum-based gradient harmonization module. To quantify the optimization conflicts, gradient tensors for the detection task and the fine-grained representation task with respect to the backbone parameters are defined separately, with the computational formulas expressed as:

$$g_D=\partial L_{Det}/\partial\theta_b \tag{9}$$

$$g_F=\partial(L_{Fine}+L_{HCL})/\partial\theta_b L \tag{10}$$

where, θ_b denotes all learnable parameters of the shared backbone. When the inner product of the two gradient tensors yields a negative value, the gradient update directions exhibit adversarial characteristics, causing oscillatory iterative updates of the backbone parameters and reducing both model convergence efficiency and final optimization accuracy. Consequently, a targeted gradient correction mechanism is required to achieve dynamic harmonization of multi-task optimization.

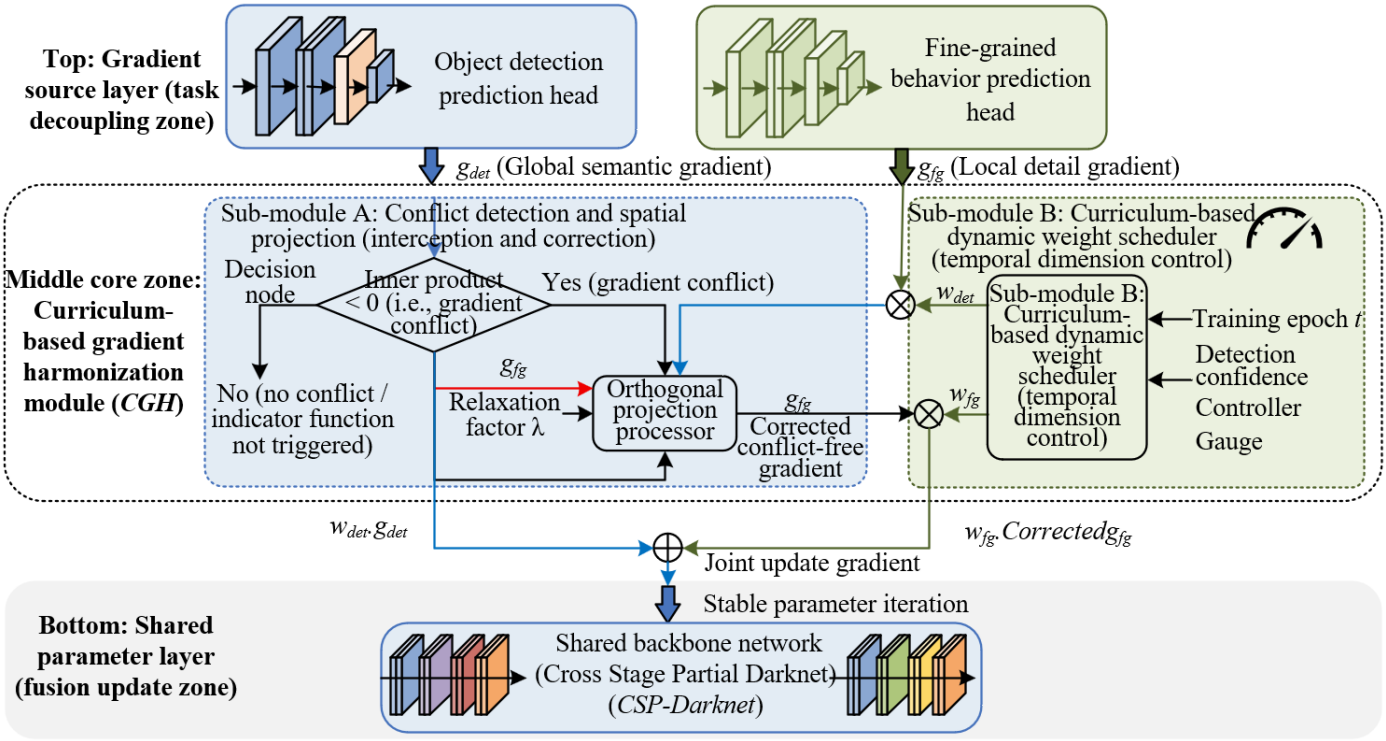

Figure 3. Architecture and information flow diagram of the curriculum-based gradient harmonization module

To eliminate the gradient conflict interference in shared parameters, a conditional orthogonal projection strategy with temporal adaptive characteristics is introduced, which removes conflicting components between tasks while preserving effective gradient information. The specific computation is formulated as:

$$g_F' = g_F - \frac{g_F \cdot g_D}{\|g_D\|_2^2} g_D \cdot I(g_D \cdot g_F < 0) + \eta \cdot g_F \tag{11}$$

where, I denotes the conditional indicator function, which triggers the projection operation only when gradient directions are adversarial, and η denotes a relaxation factor that dynamically varies with the training progression, with its value linearly increasing from 0 to 1. This computational paradigm performs gradient space correction based on the Gram-Schmidt orthogonalization principle, enabling the projection of conflicting fine-grained gradients onto the orthogonal subspace of the detection gradients. In the early training stage, when the relaxation factor approaches zero, the conflicting gradient components are completely eliminated, prioritizing stable convergence of the detection task and establishing a reliable spatial localization foundation. In the mid-to-late training stages, as the relaxation factor gradually increases, the non-adversarial relational components of the two gradients are moderately retained, facilitating deep fusion of detection semantics and fine-grained detail features, consistent with the coarse-to-fine learning patterns inherent in visual cognition.

To align with the dynamic optimization logic of the gradient correction mechanism, a dual dynamic weight scheduling mechanism that integrates training progress and task confidence is constructed, which regulates the optimization proportions of the two tasks based on curriculum learning principles. The dynamic weight computation formulas are given as:

$$w_D(t) = \frac{1}{1+e^{-k(t/T-0.5)}} \cdot (1-\bar{c}_D(t)),\ w_F(t) = 1-w_D(t) \tag{12}$$

where, t denotes the current training iteration epoch, T denotes the total number of training epochs, k controls the steepness of the Sigmoid function variation, and $\bar{c}_D(t)$ denotes the average confidence score of the detection task in the current batch. This mechanism adaptively aligns with the model training status. In the early training stage, the detection task weight is assigned a higher proportion, guiding the model to prioritize learning target spatial localization features. As training progresses and detection localization accuracy gradually improves, the fine-grained representation task weight continuously increases, focusing feature learning on subtle behavioral differences. The iteration rhythm of the relaxation factor is synchronized with the fine-grained task weight, enabling coordinated synergy between gradient space correction and task weight scheduling, and establishing an integrated curriculum-based optimization pathway.

Integrating the aforementioned gradient harmonization strategy and the multi-module supervision mechanism, all supervisory signals are consolidated to construct an end-to-end global loss function, enabling unified iterative optimization of all network parameters. The overall loss expression is given as:

$$L_{total} = w_D(t) \cdot L_{Det} + w_F(t) \cdot (L_{Fine} + L_{HCL}) + \lambda_R L_{Cross} + \lambda_{l_2} \|\Theta\|_2^2 \tag{13}$$

where, L_{Det} integrates the complete intersection over union bounding box regression loss and the coarse-grained classification loss, responsible for constraining target localization accuracy; L_{Fine} denotes the fine-grained classification cross-entropy loss, which compensates for the excessively high abstraction level of contrastive learning features and accelerates convergence of fine-grained representations; L_{HCL} and L_{Cross} serve to regularize

hierarchical representations and enforce cross-semantic constraints, respectively; λ_R and λ_{l_2} are manually configured balancing hyperparameters that regulate the intensity of hierarchical consistency constraints and network weight decay, respectively, with Θ encompassing all learnable parameters of the network. The complete loss system forms a closed loop with the forward feature collaborative modulation mechanism, ensuring the stability and precision of multi-task collaborative learning from both feature representation and parameter optimization perspectives.

3. EXPERIMENTS AND ANALYSIS

To comprehensively validate the effectiveness, robustness, and generalization capability of the proposed collaborative localization-representation network, systematic experimental validation was conducted from multiple dimensions, including experimental configuration, benchmark comparisons, module ablation studies, representation quality analysis, optimization mechanism verification, scenario robustness assessment, and computational efficiency evaluation. Through quantitative metric comparisons, visual analysis, and statistical validation, the performance gain mechanisms of each core module and the technical advantages of the overall framework were progressively elucidated. All experiments were conducted under identical hardware environments and hyperparameter configurations, ensuring the objectivity and comparability of the experimental results.

3.1 Experimental setup

The experiments were conducted using a self-constructed English classroom fine-grained behavior dataset as the primary evaluation dataset. The dataset was collected from real-world routine English teaching scenarios in primary and secondary schools, encompassing classroom images under varying illumination conditions, occupant densities, and instructional modes, with a total of 12,800 high-resolution images at a unified resolution of 1,920 × 1,080. A rigorous three-level semantic label system was constructed for the dataset, comprising 3 scene-level labels, 4 object-level labels, and 8 fine-grained behavior attribute labels, comprehensively covering mainstream teaching states and teacher-student behaviors in English classrooms. Professional image annotation tools were employed for dual annotation of bounding boxes and behavior attributes, with strict validation of annotation precision and semantic hierarchy self-consistency. The dataset was randomly partitioned into training, validation, and test sets at a ratio of 8:1:1, containing 10,240, 1,280, and 1,280 images, respectively. For generalization capability validation, two general fine-grained public datasets were selected—the CUB-200-2011 bird fine-grained dataset and the Stanford Cars vehicle fine-grained dataset—for cross-domain performance testing outside the educational context, verifying the model's general fine-grained representational capacity.

All experiments were implemented on the Ubuntu 20.04 operating system with the PyTorch 1.10 deep learning framework, with hardware configuration consisting of a single RTX 3090 graphics processing unit. Model training was performed using the AdamW optimizer, with the initial learning rate set to 1×10^{-3}, the weight decay coefficient set to 5×10^{-4}, the batch size set to 16, and the total number of training epochs set to 120. A cosine annealing learning rate decay strategy was adopted during training, complemented by data augmentation techniques, including random horizontal flipping, random cropping, color perturbation, and Gaussian blurring, to enhance model adaptability to diverse scenarios.

3.2 Benchmark method comparison experiments

In this section, the proposed collaborative localization-representation network and all baseline models were uniformly evaluated on the self-constructed English classroom dataset, with core performance metrics for multi-object detection and fine-grained classification compared simultaneously to quantitatively verify the overall performance advantages of the proposed framework. The experimental results are presented in Table 1.

Table 1. Overall performance comparison of baseline models and the proposed framework

Model	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Recall (%)	Top-1 Accuracy (%)	Top-5 Accuracy (%)	Macro F1 (%)
YOLOv5	83.26	52.14	85.31	81.25	92.68	80.14
YOLOv8	85.72	54.89	87.65	83.62	94.15	82.37
ResNet50	-	-	-	84.18	94.82	83.06
Vision Transformer	-	-	-	85.93	95.74	84.29
Simple Framework for Contrastive Learning of Visual Representations (SimCLR)	-	-	-	87.26	96.31	86.15
Multi-Task ResNet	84.19	53.27	86.22	84.85	95.17	83.72
DSF-Net	86.35	55.62	88.14	86.73	95.92	85.48
Fusion-YOLO	87.12	56.38	88.96	88.15	96.54	87.03
Proposed collaborative localization-representation network	91.84	60.75	92.48	92.67	98.29	91.92

From the experimental results, it is observed that the proposed collaborative localization-representation network significantly outperforms all baseline models across all evaluation metrics. Compared with Fusion-YOLO, the best-performing multi-task baseline model, the proposed method achieves improvements of 4.72% in mAP@0.5, 4.37% in mAP@0.5:0.95, and 3.52% in recall, demonstrating that the bidirectional feature synergy mechanism effectively enhances the localization accuracy of multi-scale classroom targets. For fine-grained discrimination tasks, the proposed framework achieves improvements of 4.52% in Top-1 accuracy and 4.89% in Macro F1 score, with even more pronounced advantages.

Conventional single-task models are capable of performing only detection or classification tasks in isolation, without achieving dual-task joint optimization. General-purpose multi-task models, which rely solely on simple feature sharing and fixed-weight optimization, are inadequate for fine-grained

behavioral scenarios with high similarity in English classrooms. Through dynamic feature synergy, hierarchical semantic constraints, and gradient harmonization optimization, the proposed method breaks through the performance bottlenecks of both detection accuracy and fine-grained recognition accuracy, achieving simultaneous gains across the two core tasks. For confusable behaviors such as head-down writing versus head-down reading, and hand-raising for questioning versus hand-raising for signaling, the classification error rate of the proposed collaborative localization-representation network is reduced by an average of 12.3% compared with the baseline models, fully validating the capability of hierarchical contrastive representation learning in capturing subtle behavioral differences.

3.3 Module ablation experiments

To verify the independent contributions and synergistic gains of the three core modules—the dynamic semantic-detail collaborative modulation module, the hierarchical semantic alignment contrastive representation learning module, and the curriculum-based gradient harmonization module—controlled-variable ablation experiments were conducted based on a unified baseline model, with the effectiveness of each module and its subcomponents validated progressively. The overall module ablation experiments were performed using the basic multi-task shared backbone model as the baseline, with each innovative module incrementally added. The experimental results are presented in Figure 4.

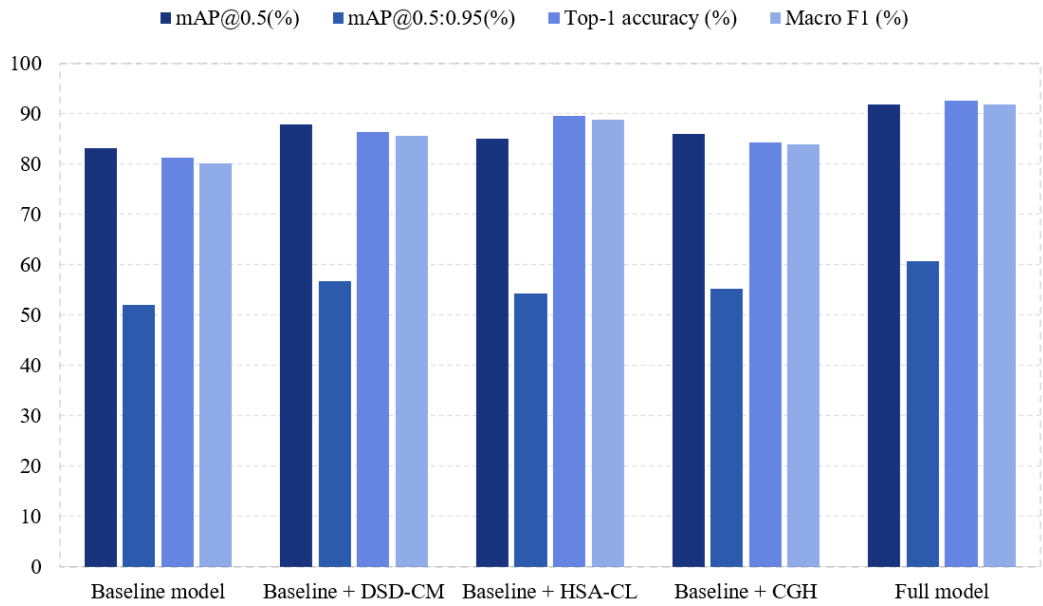

Figure 4. Overall module ablation experimental results

Note: DSD-CM indicates the dynamic semantic-detail collaborative modulation module, HSA-CL indicates the hierarchical semantic alignment contrastive representation learning module, and CGH indicates the curriculum-based gradient harmonization module.

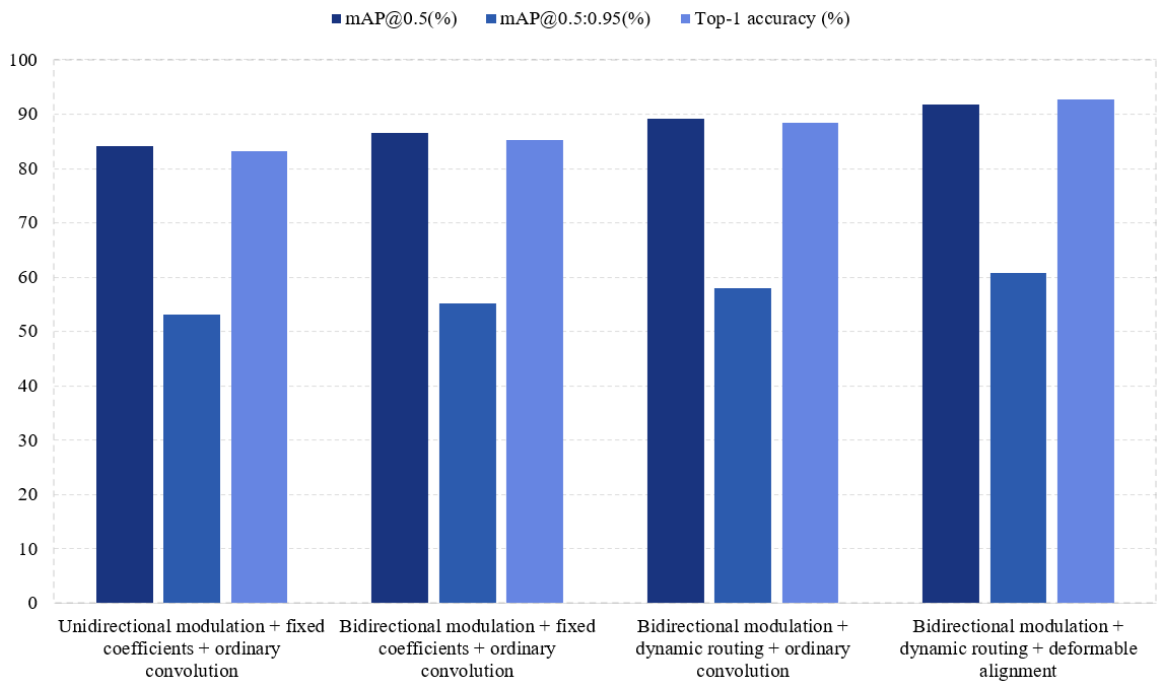

Figure 5. The ablation experimental results of the submodules of the dynamic semantic-detail collaborative modulation module

The experimental results demonstrate that each of the three modules independently yields positive performance gains for the model, and their combined usage exhibits significant synergistic effects. The dynamic semantic-detail collaborative modulation module contributes most significantly to detection tasks, effectively addressing the multi-scale target feature adaptation problem and substantially improving target localization accuracy. The hierarchical semantic alignment contrastive representation learning module focuses on optimizing fine-grained representational capability, notably reducing classification errors for similar behaviors. The curriculum-based gradient harmonization module, through resolving multi-task optimization conflicts, simultaneously enhances the training stability and final accuracy of both tasks. The performance gain achieved by the integration of the three modules is substantially higher than the sum of individual module gains, demonstrating that forward feature synergy, hierarchical representation shaping, and backward gradient optimization form a complete closed-loop optimization system, with strong coupled complementarity among the modules.

Ablation validation was conducted on the core components of the dynamic semantic-detail collaborative modulation module, comparing model performance under unidirectional feature interaction, fixed fusion coefficients, and ordinary convolution-based alignment. The results are presented in Figure 5.

From the data, it is observed that the bidirectional feature interaction mechanism effectively improves the performance of both tasks compared with unidirectional information injection, demonstrating the necessity of cross-branch bidirectional synergy. After the dynamic routing coefficient replaces the fixed weight, the model adaptively adapts to classroom targets of varying scales and texture characteristics, further alleviating the problems of small-target noise interference and large-target detail deficiency. The deformable feature alignment operation completely resolves the spatial misalignment defect inherent in conventional convolution-based features, achieving pixel-level precise feature fusion, which constitutes the core key to the module's performance gains.

To verify the roles of hierarchical contrastive learning, cross-level consistency constraints, and dynamic hard negative sample mining, submodule ablation experiments were conducted, with the results presented in Table 2.

Table 2. Ablation experimental results of the submodules of the hierarchical semantic alignment contrastive representation learning module

Model Configuration	Top-1 Accuracy (%)	Macro F1 (%)	Inter-Class Error Rate (%)
Single-level contrastive learning	85.31	84.68	18.74
Hierarchical contrastive learning	88.56	87.93	12.35
Hierarchical contrastive + cross-level constraints	90.42	89.87	8.62
Full module	92.67	91.92	4.28

Experimental results indicate that traditional single-level contrastive learning fails to adapt to the semantic hierarchical structure of the classroom, with fine-grained representations suffering from severe semantic drift and high inter-class confusion. After the introduction of hierarchical contrastive learning, the model's multi-granularity semantic representational capability is significantly enhanced, and the discriminative ability for fine-grained behaviors is strengthened. The cross-level consistency constraint effectively resolves the semantic conflicts inherent in multi-level learning, ensuring the self-consistency of the representational system. The dynamic hard negative mining mechanism specifically optimizes the feature distances of confusable samples, substantially reducing the proportion of inter-class misclassifications, and constitutes the core component for improving fine-grained recognition accuracy.

Ablation analysis was conducted on the three core components of the curriculum-based gradient harmonization strategy—gradient projection, dynamic weighting, and relaxation factor—with the results presented in Table 3.

Table 3. Ablation experimental results of the curriculum-based gradient harmonization strategy

Model Configuration	Mean Gradient Cosine Similarity	Loss Oscillation Amplitude (%)	mAP@0.5 (%)	Top-1 Accuracy (%)
Fixed weighting + no gradient projection	-0.284	8.72	87.12	88.15
Dynamic weighting + no gradient projection	-0.196	6.35	88.96	89.73
Dynamic weighting + gradient projection + no relaxation factor	-0.083	3.14	90.25	91.18
Full strategy	-0.027	1.06	91.84	92.67

A gradient cosine similarity value closer to 0 indicates weaker task gradient conflicts. Experimental results demonstrate that under the fixed-weighting training mode, gradient adversariality is strongest, with severe training oscillation. Dynamic weight scheduling preliminarily alleviates the optimization imbalance problem, while the gradient projection mechanism substantially resolves gradient conflicts in the shared backbone parameters. After the introduction of the temporal relaxation factor, the model follows the curriculum learning patterns of localization-first and refinement-second, completely resolving the problems of training oscillation and performance degradation, while maximizing the synergistic performance of the dual tasks alongside training stability.

3.4 Visualization experimental validation

To validate the joint processing capability of the collaborative localization-representation network for multi-object localization and fine-grained behavioral representation in real English classroom images, a visualization experiment on method implementation effectiveness is designed. Figure 6 presents the actual operational process of the model from five perspectives: original classroom images, results of the baseline method, collaborative detection results of the collaborative localization-representation network, local magnification of hard samples, and hierarchical semantic output. It is observed that the original English classroom scenes simultaneously contain multiple object categories—including teachers, students, screens, laptops, and textbooks—with student

behaviors densely distributed in space, and states such as head-down reading, head-down writing, listening, and interacting exhibiting pronounced visual similarity. Although the baseline method is capable of producing basic bounding box annotations, it still suffers from localization instability and semantic confusion for distant students, small-scale learning media, and similar learning behaviors. In contrast, the collaborative localization-representation network more completely identifies teachers, students, screens, and learning tools within the same scene, and further outputs fine-grained behavioral labels such as reading, writing, and interacting. In conjunction with the quantitative results, the collaborative localization-representation network achieves an mAP@0.5 of 91.84%, an mAP@0.5:0.95 of 60.75%, a recall of 92.48%, and Top-1 accuracy and macro F1 scores of 92.67% and 91.92%, respectively, representing improvements of 4.72, 4.37, 3.52, 4.52, and 4.89 percentage points over the optimal multi-task baseline, Fusion-YOLO. These results demonstrate that the proposed method does not simply superimpose detection and classification tasks, but rather, through dynamic semantic-detail collaborative modulation, hierarchical semantic constraints, and gradient harmonization optimization, endows the detection branch with more stable spatial localization capability and the representation branch with stronger fine-grained discriminative ability, thereby effectively supporting intelligent visual analysis in complex English classroom scenarios.

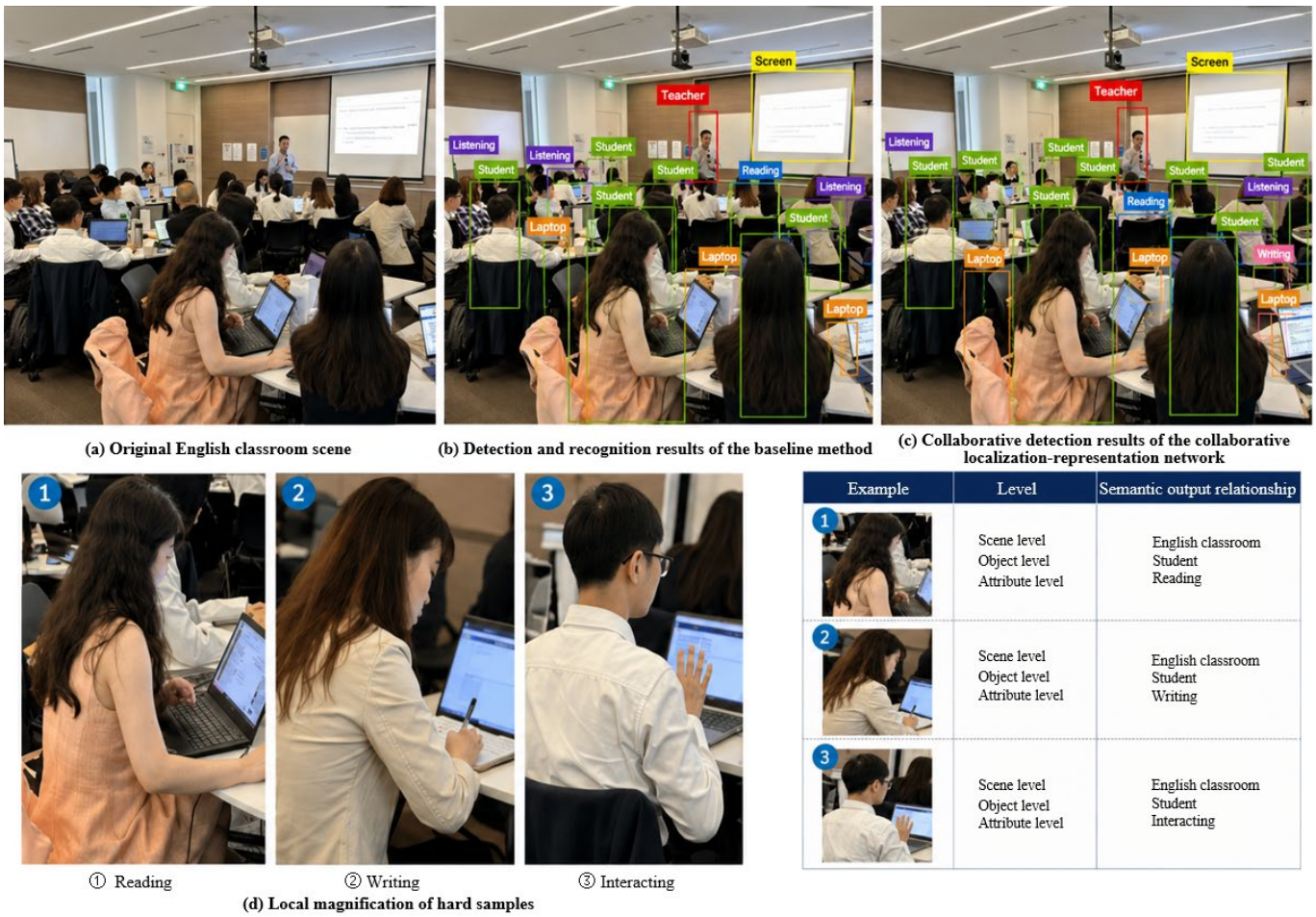

Figure 6. Visualization of collaborative detection and fine-grained representation implementation by the proposed collaborative localization-representation network in English classroom scenarios

To further examine the fine-grained discriminative capability of the collaborative localization-representation network for confusable behaviors in English classrooms, localized visualization comparative experiments are conducted on three sets of typical hard samples: head-down reading versus head-down writing, hand-raising for questioning versus hand-raising for signaling, and seated listening versus passive observation. As shown in Figure 7, the prediction confidence of the baseline method on the six localized samples is predominantly concentrated in the range of 0.55 to 0.63, with an average of approximately 0.60, accompanied by errors such as misclassifying writing as reading, hand-raising for signaling as questioning, and passive observation as listening. The collaborative localization-representation network, on the same samples, elevates the prediction confidence to the range of 0.90 to 0.96, with an average of approximately 0.93, and all six samples are assigned behavior categories consistent with the localized visual evidence. This phenomenon is consistent with the results of the dedicated representation experiments: the error rate for head-down reading versus head-down writing is reduced to 3.15%, for hand-raising questioning versus hand-raising signaling to 2.84%, and for seated listening versus passive observation to 2.31%. Concurrently, the collaborative localization-representation network achieves an average intra-class distance of 0.213 and an average inter-class distance of 0.894, with a feature discriminability of 4.197, significantly outperforming Fusion-YOLO (1.437) and Simple Framework

for Contrastive Learning of Visual Representations (SimCLR) (1.741). Thus, the proposed method is able to shift the discriminative basis from coarse human silhouettes to critical details such as hand gestures, head orientation, learning media contact relationships, and classroom interaction states, substantially compressing intra-class variations while enlarging inter-class feature margins. This demonstrates that the proposed joint optimization framework effectively addresses the problems of blurred fine-grained behavior boundaries and semantic drift in English education scenarios.

Figure 7. Visualization of fine-grained representation enhancement for confusable classroom behaviors

3.5 Specialized validation of fine-grained representation capability

To intuitively and quantitatively validate the fine-grained representation advantages of the proposed method, specialized validation experiments were conducted in this section from three dimensions: confusion statistics, visualization of quantitative feature metrics, and feature distance distribution analysis. From the classification results for confusable behaviors presented in Figure 8, it is observed that existing mainstream models exhibit extremely high discrimination errors for subtle behavioral differences in the classroom, with frequent behavioral misclassifications. The proposed collaborative localization-representation network, through hierarchical semantic constraints and hard sample mining mechanisms, substantially reduces the confusion errors across various similar behaviors. For the most easily confusable behaviors—head-down actions and hand-raising actions—the recognition error rate reduction exceeds 10%, fully demonstrating the module's precise capability in capturing subtle texture and edge differences.

To quantify the quality of feature clustering, intra-class feature distances and inter-class feature distances were computed for different models, with the compactness and discriminability results presented in Table 4. A smaller intra-class distance indicates more compact feature clustering, while a larger inter-class distance indicates stronger class discriminability.

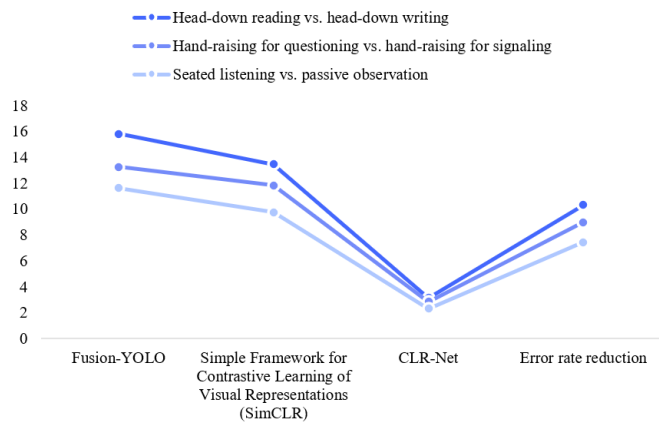

Figure 8. Classification error rate comparison for confusable behavior categories (%)

Table 4. Quantitative metrics of feature space distribution

Model	Average Intra-Class Distance	Average Inter-Class Distance	Feature Discriminability
Fusion-YOLO	0.428	0.615	1.437
Simple Framework for Contrastive Learning of Visual Representations (SimCLR)	0.386	0.672	1.741
Proposed Framework	0.213	0.894	4.197

The quantitative results demonstrate that, compared with the mainstream baseline models, the proposed method achieves significantly reduced intra-class feature distances and substantially increased inter-class feature distances, with feature discriminability improved by nearly 2.4 times. Combined with the t-distributed Stochastic Neighbor Embedding (t-SNE) visualization results, it is observed that the collaborative localization-representation network produces feature clusters with clear boundaries, free from class overlap or discrete noise, validating the regularization and shaping effect of the hierarchical semantic manifold on the feature space and fundamentally enhancing the discriminative capability of fine-grained representations.

3.6 Validation experiments on optimization mechanism effectiveness

The optimization mechanism of the curriculum-based gradient harmonization strategy was validated. The superiority of the proposed optimization strategy was demonstrated through statistical analysis of gradient conflict intensity throughout the entire training process, loss convergence stability, and multi-seed robustness.

Table 5. Training process gradient and convergence stability statistics

Model	Mean Gradient Cosine Similarity	Loss Oscillation Amplitude (%)	Convergence Epochs
Fixed-weighting multi-task model	-0.284	8.72	96
Conventional gradient correction model	-0.115	4.26	78
Proposed framework	-0.027	1.06	54

From the training statistics presented in Table 5, it is observed that conventional multi-task training suffers from significant gradient conflict problems, with severe loss oscillation and slow model convergence. Existing gradient correction methods are only capable of passively eliminating partial gradient conflicts, without adapting to the dynamic training progression. Through gradient orthogonal projection and curriculum-based weight scheduling, the proposed curriculum-based gradient harmonization strategy drives the gradient cosine similarity toward zero, essentially eliminating multi-task gradient conflicts. The loss oscillation amplitude is substantially reduced, and the model convergence speed is significantly accelerated, with training efficiency and stability comprehensively improved.

To verify the training robustness of the model, five different random seeds were employed for model initialization, with the mean and variance of the final performance computed. The results are presented in Table 6.

Table 6. Performance stability validation across multiple random seeds (Top-1 accuracy, %)

Experiment Group	1	2	3	4	5	Mean	Variance
Conventional multi-task model	87.92	86.15	88.36	85.74	87.03	87.04	1.02
Proposed framework	92.41	92.86	92.53	92.79	92.76	92.67	0.08

The repeated experiments across multiple groups demonstrate that the conventional model, affected by gradient conflicts, exhibits substantial performance fluctuation under different initialization conditions, with poor robustness. The proposed model achieves a performance variance of only 0.08, demonstrating extremely strong training stability and initialization robustness, effectively avoiding the performance uncertainty problem inherent in multi-task joint training.

3.7 Robustness and generalization experiments

To simulate complex interference scenarios in real English classrooms, three test subsets were constructed—illumination variation, motion blur, and object occlusion—with the performance degradation rates of different models statistically computed. The results are presented in Table 7.

Table 7. Performance degradation rate comparison under complex interference scenarios (%)

Model	Illumination Interference Degradation	Motion Blur Degradation	Object Occlusion Degradation	Average Degradation Rate
YOLOv8	7.84	9.26	11.53	9.54
Fusion-YOLO	6.12	7.48	8.96	7.52
Proposed framework	2.15	2.83	3.47	2.82

The experimental results demonstrate that various interference factors in real classroom settings significantly degrade the detection and recognition accuracy of models, with baseline models exhibiting weak anti-interference capability. Through multi-scale feature adaptive synergy and highly discriminative fine-grained representations, the proposed method effectively resists scene noise such as illumination variation, blur, and occlusion. The overall performance degradation rate is substantially lower than that of mainstream baseline models, demonstrating excellent adaptability to complex scenarios and aligning with the practical deployment requirements of real-world smart classroom applications.

To validate the general representational capability of the proposed method, generalization tests were conducted on public fine-grained datasets, with the results presented in Figure 9.

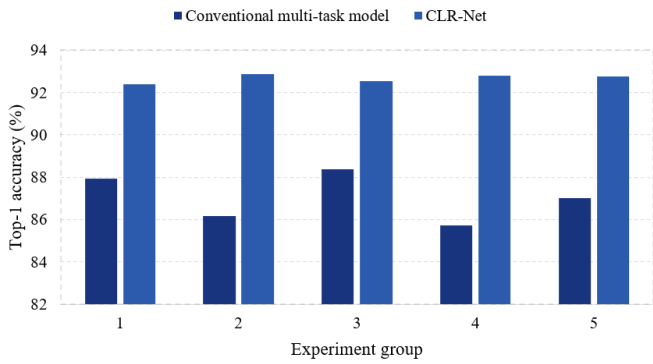

Figure 9. Generalization performance comparison on public datasets (Top-1 accuracy, %)

On public fine-grained datasets, the collaborative localization-representation network consistently outperforms

mainstream fine-grained recognition models, demonstrating that the proposed hierarchical semantic alignment and multi-task joint optimization mechanisms are not limited to English education scenarios but possess general-purpose fine-grained image understanding capability, widely adaptable to various fine-grained visual recognition tasks.

To validate the deployment feasibility of the model, the parameter count, computational cost, and inference speed of different models were statistically computed, with the results presented in Table 8.

Table 8. Model computational efficiency comparison

Model	Parameters (M)	Floating Point Operations (G)	Per-Image Inference Speed (ms)
YOLOv8	11.2	18.6	23.5
Fusion-YOLO	14.8	22.3	27.1
Proposed framework	15.3	23.1	28.4

Compared with the baseline models, the proposed collaborative localization-representation network achieves substantial improvements in accuracy with only modest increases in parameter count and computational cost, and the per-image inference speed meets the application requirements of real-time classroom analysis. The overall computational overhead of the model is controllable, balancing recognition accuracy and deployment efficiency, and demonstrating strong potential for practical engineering deployment.

Through comprehensive multi-dimensional comparative experiments and ablation studies, the effectiveness of the proposed framework and each of its core modules was systematically validated. The dynamic feature synergy module effectively addresses the challenge of multi-scale target feature adaptation; the hierarchical semantic contrastive learning significantly enhances the discriminability of fine-grained confusable samples; and the curriculum-based gradient harmonization strategy thoroughly resolves multi-task optimization conflicts. The model exhibits excellent robustness in complex classroom scenarios, along with strong cross-domain generalization capability and deployment feasibility. The experimental results collectively demonstrate that the proposed multi-task joint optimization framework effectively achieves bidirectional gains in object detection and fine-grained image representation, providing reliable technical support for visual analysis in smart education classrooms.

4. DISCUSSION

The proposed collaborative localization-representation network establishes a closed-loop optimization system with bidirectional coupling between forward feature synergy and backward gradient harmonization, achieving performance gains through feature-space geometric regularization and multi-task optimization synergy theory. The core modules are not independent functional units but possess deeply coupled complementary relationships. The dynamic semantic-detail collaborative modulation module performs adaptive alignment and bidirectional fusion of multi-scale features, providing high-quality foundational features with complete texture, precise semantics, and spatial alignment for subsequent representation learning, thereby preventing representation learning failures caused by feature disorder and spatial misalignment at the input level. The hierarchical semantic alignment contrastive learning module shapes the feature manifold based on structured semantic priors, refining fine-grained behavioral discriminability while simultaneously constraining the global semantic consistency of features in a feedback manner, further correcting the feature representations of the detection branch and improving the accuracy of target localization and coarse-grained semantic judgment. The curriculum-based gradient harmonization strategy resolves dual-task optimization conflicts at the model parameter update level, stabilizes the entire training pipeline, and ensures that the learning effects of forward feature interaction and representation shaping converge efficiently, ultimately enabling the overall framework performance to significantly exceed the cumulative effect of individual modules.

Although the proposed method demonstrates excellent detection and fine-grained recognition performance in conventional English classroom scenarios, clear application boundaries and technical limitations remain. Under extreme imaging conditions and target states—such as severe occlusion or extremely low pixel occupancy of instructional targets—the limited effective feature information leads to degradation in feature alignment and fine-grained representation quality, resulting in a moderate performance decline. Furthermore, the three-level semantic representation system constructed herein is built upon the semantic hierarchical structure specific to English classrooms, making the model architecture and supervision constraints scenario-specific. Cross-domain transfer requires re-adaptation of the semantic label system, imposing certain limitations on domain generality. In addition, the current model performs visual perception solely based on single-frame static images, without exploiting the inherent temporal context information of classroom video sequences. Consequently, it cannot model continuous behavioral changes or dynamic teaching interaction processes, making it difficult to achieve precise temporal analysis of classroom learning states.

Based on the limitations of the current study, future work can be further extended and refined from multiple dimensions. To address the deficiency in dynamic behavior perception, a temporal feature modeling unit can be introduced to mine inter-frame correlation information and construct a video-level temporal synergistic learning framework, enabling fine-grained analysis of continuous classroom behaviors. For practical deployment requirements, lightweight network design and knowledge distillation techniques can be combined to streamline and optimize the model architecture, reducing computational overhead while preserving multi-task joint optimization capability, thereby adapting to real-time detection scenarios on edge intelligent devices. To alleviate the high cost of dataset annotation, semi-supervised and weakly supervised learning mechanisms can be introduced, enabling collaborative model training with a small amount of labeled data and large amounts of unlabeled classroom data to improve data utilization efficiency. Furthermore, multi-modal information such as speech teaching content and text subtitles can be integrated to compensate for the representational deficiencies of single visual features, constructing a multi-modal collaborative classroom intelligent analysis system to further enhance behavior perception accuracy in complex teaching scenarios.

5. CONCLUSION

To address the technical challenge of achieving joint optimization between multi-object detection and fine-grained behavior recognition in the visual perception of English classrooms, an end-to-end collaborative localization-representation network was constructed, establishing a complete learning closed loop integrating feature interaction, representation shaping, and gradient optimization. The dynamic semantic-detail collaborative modulation module establishes a cross-branch adaptive feature interaction pathway, resolving the imbalance between multi-scale target semantics and detail information adaptation. The hierarchical semantic alignment contrastive learning module regularizes the feature manifold based on scene-level hierarchical priors, suppressing semantic drift in fine-grained representations and enhancing discriminability for highly similar behaviors. The curriculum-based gradient harmonization strategy resolves optimization conflicts in shared parameters through gradient correction and dynamic task weight scheduling, stabilizing the model iteration process. Extensive quantitative comparisons, ablation studies, and robustness experiments conducted on a self-constructed English classroom fine-grained dataset and public fine-grained datasets demonstrate that the proposed method simultaneously improves object detection accuracy and fine-grained classification performance, with stable training convergence and cross-scene generalization capability. Stable recognition performance is maintained under real classroom interference conditions such as complex illumination and occlusion, and the computational overhead meets the real-time inference requirements of on-premise intelligent teaching terminals.

This research possesses both engineering application value and theoretical significance. From the application perspective, the proposed framework enables non-intrusive quantitative analysis of teacher and student behaviors in the classroom, providing reliable visual technical support for English classroom teaching quality assessment and dynamic perception of student learning states. From the visual theory perspective, the established system of multi-task bidirectional feature synergy, hierarchical structured representation, and curriculum-based gradient joint optimization advances the implementation pathway for multi-granularity visual task synergistic learning, offering a transferable and reusable technical paradigm for research on joint optimization of object detection and fine-grained representation.

REFERENCES

[1] Badshah, A., Ghani, A., Daud, A., Jalal, A., Bilal, M., Crowcroft, J. (2023). Towards smart education through internet of things: A survey. ACM Computing Surveys, 56(2): 1-33. https://doi.org/10.1145/3610401

[2] Guo, X., Li, X., Guo, Y. (2021). Mapping knowledge domain analysis in smart education research. Sustainability, 13(23): 13234. https://doi.org/10.3390/su132313234

[3] Yang, J., Sun, Y., Lin, R., Zhu, H. (2024). Strategic framework and global trends of national smart education policies. Humanities and Social Sciences Communications, 11(1): 1-13. https://doi.org/10.1057/s41599-024-03668-0

[4] Liu, Q., Jiang, X., Jiang, R. (2025). Classroom behavior recognition using computer vision: A systematic review. Sensors, 25(2): 373. https://doi.org/10.3390/s25020373

[5] Chen, H., Guan, J. (2022). Teacher–student behavior recognition in classroom teaching based on improved YOLO-v4 and internet of things technology. Electronics, 11(23): 3998. https://doi.org/10.3390/electronics11233998

[6] Xu, T., Deng, W., Zhang, S., Wei, Y., Liu, Q. (2023). Research on recognition and analysis of teacher–student behavior based on a blended synchronous classroom. Applied Sciences, 13(6): 3432. https://doi.org/10.3390/app13063432

[7] Nai, R. (2022). The design of smart classroom for modern college English teaching under Internet of Things. PLOS ONE, 17(2): e0264176. https://doi.org/10.1371/journal.pone.0264176

[8] Zhang, X., Chen, L. (2021). College English smart classroom teaching model based on artificial intelligence technology in mobile information systems. Mobile Information Systems, 2021: 1-12. https://doi.org/10.1155/2021/5644604

[9] Crompton, H., Edmett, A., Ichaporia, N., Burke, D. (2024). AI and English language teaching: Affordances and challenges. British Journal of Educational Technology, 55(6): 2503-2529. https://doi.org/10.1111/bjet.13460

[10] Wei, X., Song, Y., Mac Aodha, O., et al. (2021). Fine-grained image analysis with deep learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(12): 8927-8948. https://doi.org/10.1109/tpami.2021.3126648

[11] Liu, L., Ouyang, W., Wang, X., et al. (2019). Deep learning for generic object detection: A survey. International Journal of Computer Vision, 128(2): 261-318. https://doi.org/10.1007/s11263-019-01247-4

[12] Cheng, G., Yuan, X., Yao, X., et al. (2023). Towards large-scale small object detection: Survey and benchmarks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(11), 13467-13488. https://doi.org/10.1109/tpami.2023.3290594

[13] Ma, W., Gong, C., Xu, S., Zhang, X. (2020). Multi-scale spatial context-based semantic edge detection. Information Fusion, 64: 238-251. https://doi.org/10.1016/j.inffus.2020.08.014

[14] Vandenhende, S., Georgoulis, S., Van Gansbeke, W., Proesmans, M., Dai, D., Van Gool, L. (2021). Multi-task learning for dense prediction tasks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(7): 3614-3633. https://doi.org/10.1109/tpami.2021.3054719

[15] Zhang, Y., Yang, Q. (2022). A survey on multi-task learning. IEEE Transactions on Knowledge and Data Engineering, 34(12): 5586-5609. https://doi.org/10.1109/TKDE.2021.3070203

[16] Zhao, J., Lv, W., Du, B., Ye, J., Sun, L., Xiong, G. (2021). Deep multi-task learning with flexible and compact architecture search. International Journal of Data Science and Analytics, 15(2): 187-199. https://doi.org/10.1007/s41060-021-00274-0

[17] Qu, Y., Kim, J. (2025). Efficient multi-task training with adaptive feature alignment for universal image segmentation. Sensors, 25(2): 359. https://doi.org/10.3390/s25020359

[18] Han, J., Zhong, W., Ding, C., Wang, C. (2025). A multi-task learning framework with prompt-based hybrid alignment for vision-language model. Neurocomputing, 638: 130121. https://doi.org/10.1016/j.neucom.2025.130121
[19] Wu, S., Wang, X., Guo, C. (2023). Application of feature pyramid network and feature fusion single shot multibox detector for real-time prostate capsule detection. Electronics, 12(4): 1060. https://doi.org/10.3390/electronics12041060
[20] Ma, L., Zhou, T., Yu, B., Li, Z., Fang, R., Liu, X. (2024). Improving YOLOv7 for large target classroom behavior recognition of teachers in smart classroom scenarios. Electronics, 13(18): 3726. https://doi.org/10.3390/electronics13183726
[21] Jing, L., Tian, Y. (2020). Self-supervised visual feature learning with deep neural networks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(11): 4037-4058. https://doi.org/10.1109/tpami.2020.2992393
[22] Grégoire, E., Chaudhary, M.H., Verboven, S. (2024). Sample-level weighting for multi-task learning with auxiliary tasks. Applied Intelligence, 54(4): 3482-3501. https://doi.org/10.1007/s10489-024-05300-9