RERC-DETR: A Real-Time Lightweight Transformer with Logical Distillation for Classroom Behavior Recognition

Xinghong Wang[1], Li Tao[1,†], Ying Wang[1], Yonghou He[1], Fei Jiang[2], Jingwei Qu[1], Zhaofang Yang[1]

***Abstract*—Student behavior recognition in classrooms is pivotal for smart education, yet achieving real-time performance without compromising accuracy remains a significant challenge. Although Transformers possess superior global modeling, their computational overhead often hinders latency-sensitive deployment. To address this, we propose RERC-DETR, a lightweight multi-scale Transformer detector optimized for classroom analytics. Specifically, we introduce RAD-Net, an efficient backbone designed to minimize redundancy, and a Recalibrated Cross-Scale Feature Pyramid Network (RCFPN) that preserves high-resolution semantics and resolves feature aliasing. Furthermore, we introduce a Quality-Aware Logical Distillation strategy utilizing teacher confidence to selectively transfer structural knowledge, enhancing the student detector without inference overhead. Experiments on STBD-08 and SC7 datasets demonstrate that RERC-DETR achieves an optimal accuracy-efficiency trade-off. Compared with the baseline RT-DETR, our model reduces the parameter count by 44% to only 11.2M, while attaining 93.9% mAP$_{50}$ at 156 FPS on STBD-08 and 89.7% mAP$_{50}$ at 80 FPS on SC7, outperforming multiple SOTA detectors under a compact parameter budget.**

***Index Terms*—Classroom Behavior Recognition, Real-Time Object Detection, Lightweight Transformer, Small Object Detection, Logical Distillation**

I. Introduction

Student behavior recognition in classrooms is a key component of smart education systems, providing an objective basis for evaluating instructional quality and student engagement [1], [2]. Compared with general action recognition, it must handle dense student layouts and large scale variations, while distinguishing fine-grained behaviors (e.g., "reading" vs. "writing") in real time to enable timely feedback and personalized interventions [3]. These requirements make both robust recognition and high-throughput inference essential in practical deployments.

Traditional assessment relies on manual protocols such as the Classroom Assessment Scoring System (CLASS) [4], which are labor-intensive, time-consuming, and subjective. In contrast, automated classroom analysis must satisfy both accuracy and real-time efficiency. Lightweight CNN-based detectors, such as YOLO [5], are computationally efficient but often struggle with dense student layouts, occlusion, and small back-row targets because their local receptive fields limit long-range context modeling. Transformer-based detectors, such as DETR [6], improve global reasoning through self-attention, yet their high computational cost and latency remain problematic for deployment in classroom scenarios. Even more efficient variants, such as RT-DETR [7], achieve better speed, but their small-object accuracy may still be limited in crowded scenes with large scale variation.

Motivated by these limitations, we propose RERC-DETR, a lightweight Transformer detector designed for classroom behavior recognition under real-time constraints. The proposed model combines hardware-aware backbone design, multi-scale feature recalibration, and confidence-guided distillation to improve localization precision without increasing inference overhead. Specifically, RAD-Net reduces redundant computation while preserving high-resolution cues, RCFPN strengthens cross-scale fusion for small objects, and QALD transfers reliable teacher knowledge to the student model. Together, these components allow RERC-DETR to target the practical regime where both speed and accuracy are required.

The main contributions of this paper are summarized as follows:

- **RERC-DETR**: We propose a lightweight Transformer detector that achieves an optimal accuracy–efficiency trade-off. By integrating the RAD-Net backbone with the RCFPN neck, it effectively resolves feature aliasing and small object detection challenges in dense classrooms.
- **QALD**: We introduce a Quality-Aware Logical Distillation strategy to transfer structural knowledge from a teacher model. It utilizes confidence-based weighting to suppress noise, boosting student performance without extra inference cost.
- **SC7 Dataset**: To address the scarcity of fine-grained benchmarks in complex scenes, we present SC7, a challenging university classroom dataset featuring simultaneous multi-view annotations, serving as a rigorous testbed to validate the robustness of our model across diverse classroom scenarios.

II. Related Work

This section reviews the literature across three key dimensions: classroom behavior analysis approaches, which

†Corresponding author

*This work is supported by Chongqing Academy of Science and Technology Basic Research Founding (No. 2022024030029 and 2022025030009)

[1]X. Wang, L. Tao, Y. Wang, Y. He, J. Qu, and Z. Yang are with the School of Computer and Information Science & School of Software, Southwest University, Chongqing, China; emails: wxh112024321341998@email.swu.edu.cn, tli@swu.edu.cn, waying95@swu.edu.cn, yhhe523@email.swu.edu.cn, qujingwei@swu.edu.cn, goodluck@swu.edu.cn.

[2]F. Jiang is with the Chongqing Academy of Science and Technology, Chongqing, China; email: jiang.feiabc@163.com.

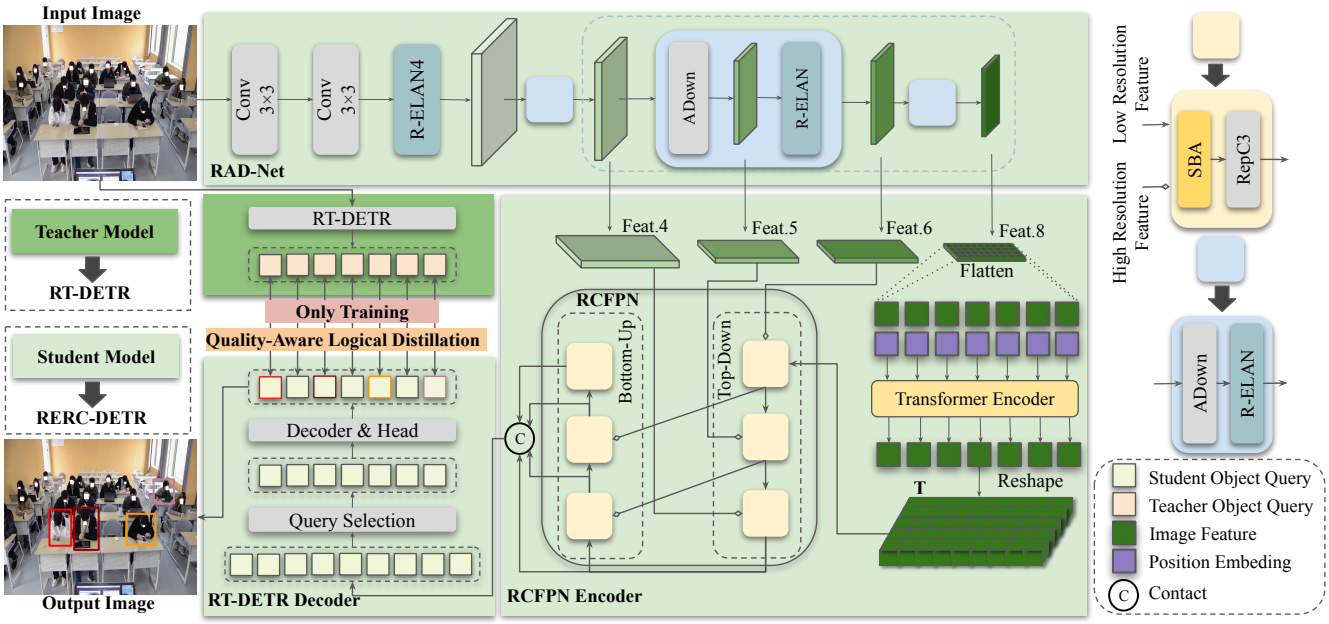


Fig. 1: Overall architecture of the proposed RERC-DETR detector. It incorporates the RAD-Net backbone and RCFPN-Encoder for efficient inference, along with a Quality-Aware Logical Distillation (QALD) strategy employed during training.

establish the application context of this study; deep learning-based recognition methods, which provide the technical foundation for our proposed detector; and knowledge distillation, which serves as a critical optimization strategy for efficient inference.

A. Classroom Behavior Analysis Approaches

Classroom behavior analysis is pivotal for evaluating instructional quality and student engagement. Early methodologies relied on manual observation protocols. For instance, Flanders' Interaction Analysis [8] focused on deriving teaching patterns from student-teacher interactions, while the CLASS [4] provided a standardized framework for assessing classroom processes. Subsequent approaches evolved from manual coding and handcrafted descriptors to automated systems using statistical models. Recently, the field has shifted towards deep learning paradigms, which offer superior capability in handling complex classroom scenes and are discussed in detail in the following subsection.

B. Deep Learning-based Classroom Behavior Recognition

CNN-based detectors, including two-stage (e.g.,Faster R-CNN [9]) and single-stage (e.g.,YOLO [5]) models, have been widely used to detect students and recognize classroom behaviors, benefiting from strong representation learning and end-to-end optimization. However, their inherently local receptive fields limit the ability to capture long-range dependencies, hindering performance in crowded scenes with strong interactions and small distant targets.Recently, Transformers have shown promise in classroom behavior recognition [10] by modeling global context through self-attention. Detection Transformers (e.g., DETR [6]) remove anchors and Non-Maximum Suppression, simplifying the detection pipeline but often incurring high computational cost. Efficient variants such as RT-DETR [7] improve speed, yet they may compromise small-object accuracy and remain heavy for dense classroom scenes. Our RERC-DETR builds on this line of work, integrating a lightweight architecture with rich multi-scale features tailored to the dense and fine-grained nature of classroom behaviors.

C. Knowledge Distillation

Knowledge Distillation (KD) compresses models by transferring knowledge from a complex, high-performance network (teacher) to a compact, lightweight network (student) [11]. Common strategies align features, responses, or relations [12] to boost lightweight detectors. However, most methods treat teacher predictions equally, risking the propagation of noise, especially in crowded scenes. Our Quality-Aware Logical Distillation addresses this by incorporating teacher confidence, encouraging the student to learn from high-quality predictions while suppressing false positives.

III. Proposed Method

To address the real-time recognition challenges of Transformer-based detectors in complex classroom scenarios, we propose *RERC-DETR*, a lightweight detector tailored for student behavior identification. Fig. 1 illustrates the overall architecture of RERC-DETR, along with the QALD strategy integrated during the training phase. First, the input image is processed by our computationally efficient *RAD-Net* backbone and a Transformer Encoder to extract multi-scale features with global context. Subsequently, RCFPN fuses these features using Selective Boundary Aggregation (SBA) [13] to effectively propagate semantic and localization cues. Then, uncertainty-aware query selection [7] initializes queries which are fed into the RT-DETR Decoder [7] for final predictions. Finally, a QALD strategy is integrated during training to leverage teacher knowledge without inference

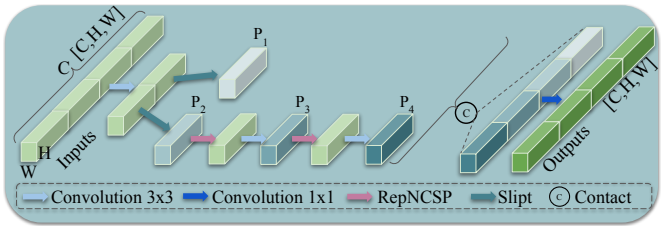


Fig. 2: Structure of the R-ELAN multi-branch enhancement block.

overhead. In the following subsections, we elaborate on the specific designs of these core components.

A. RAD-Net

RAD-Net is a lightweight backbone designed to preserve fine small-object cues while minimizing parameters and computational cost. It alternates Reparameterized Efficient Layer Aggregation Networks (R-ELAN) with ADown heterogeneous downsampling units [14] to produce strides $\{8, 16, 32\}$ (80×80, 40×40, 20×20) for a 640×640 input, maintaining high-resolution localization early and compact semantics later. The detailed architectures of these two core components are described as follows.

R-ELAN Block. As shown in Fig. 2, R-ELAN constructs multiple gradient paths by splitting, cascading, and concatenating feature branches. This design preserves shallow spatial details while enriching deeper semantic representations with low computational overhead.

ADown Module. ADown performs stride-2 reduction via complementary branches: a learnable conv path capturing local detail and a pooled path preserving salient responses with minimal cost. Given a split (X_a, X_b) of feature channels, the output Y is obtained by concatenating the processed branches:

$$Y = \mathrm{Concat}\big(\mathrm{Conv}_{3\times3}(X_a),\ \mathrm{Conv}_{1\times1}(\mathrm{MaxPool}_{3\times3}(X_b))\big) \tag{1}$$

B. RCFPN-Encoder

The *RCFPN-Encoder* serves as the neck, integrating a Transformer Encoder for global context with RCFPN. Traditional FPNs often suffer from information imbalance, where deep semantics dilute the high-resolution details essential for small objects. To mitigate this, we introduce two key mechanisms:

Global Context Enhancement. The deepest backbone feature (Feat.8) is processed by a Transformer Encoder to capture long-range dependencies critical for global reasoning. Specifically, Feat.8 is flattened, encoded by multi-head self-attention and a feedforward network, then reshaped to yield T:

$$Q = K = V = \mathrm{Flatten}(\mathrm{Feat.8}) \tag{2}$$

$$T = \mathrm{Reshape}\left(\mathrm{Encoder}(Q, K, V)\right) \tag{3}$$

where Q, K, and V denote query, key, and value; Encoder applies self-attention and feedforward interaction; Reshape restores the spatial form of Feat.8 to form a context-rich lateral feature.

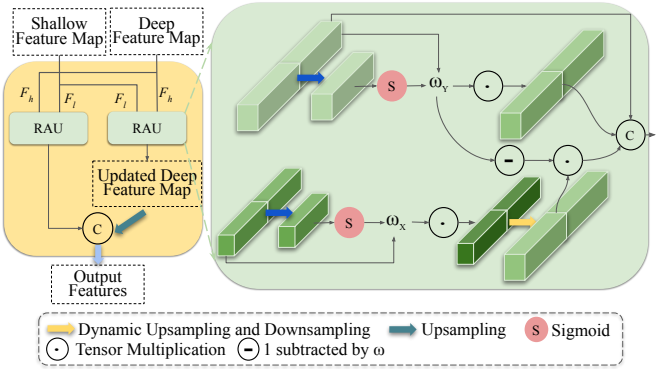


Fig. 3: Structure of the Selective Boundary Aggregation module.

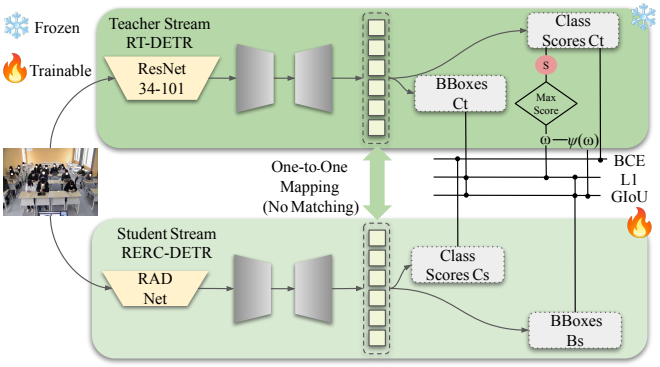


Fig. 4: Overview of the Quality-Aware Logical Distillation Framework.

Recalibrated Bidirectional Fusion. We employ a gated SBA module, as shown in Fig. 3, to dynamically calibrate information flow between adjacent scales (F_h, F_l). Unlike simple addition, SBA uses learnable gates (g_h, g_l) to selectively emphasize relevant features. Let $\mathcal{R}(X \rightarrow S_t)$ denote bilinear interpolation to the target size S_t. The gating mechanism is defined as:

$$g_h = \sigma(\mathrm{Conv}_{1\times1}(F_h)), \quad g_l = \sigma(\mathrm{Conv}_{1\times1}(F_l)) \tag{4}$$

Let $S_h = (H_h, W_h)$ and $S_l = (H_l, W_l)$ be the spatial sizes of F_h and F_l. We adopt a gated convex-like combination to prevent magnitude inflation:

$$\begin{aligned} \tilde{F}_l &= g_l \odot F_l + (1 - g_l) \odot \mathcal{R}(g_h \odot F_h \rightarrow S_l) \\ \tilde{F}_h &= g_h \odot F_h + (1 - g_h) \odot \mathcal{R}(g_l \odot F_l \rightarrow S_h) \end{aligned} \tag{5}$$

The calibrated output of SBA is then

$$\mathrm{SBA}(F_h, F_l) = \mathrm{Conv}_{3\times3}\Big(\mathrm{Concat}(\tilde{F}_h, \mathcal{R}(\tilde{F}_l \rightarrow S_h))\Big) \tag{6}$$

The SBA output is refined by a RepC3 block along the top-down path:

$$F_i^{td} = \mathrm{RepC3}\big(\mathrm{SBA}(F_h, F_l)\big) \tag{7}$$

Here, σ denotes the sigmoid function. RCFPN applies SBA in both top-down and bottom-up paths, allowing semantic and localization cues to be exchanged across scales before decoding.

C. Quality-Aware Logical Distillation

QALD aligns student outputs with teacher predictions using confidence-weighted logical supervision, reducing the impact of unreliable teacher boxes.

For the i-th query, we derive a quality indicator w_i from the teacher's confidence:

$$w_i = \max_{c \in \mathcal{C}} (\sigma(C_{i,c}^T)) \tag{8}$$

where $C_{i,c}^T$ is the teacher's logit for class c. The logical distillation loss $\mathcal{L}_{logical}$ combines classification (Binary Cross Entropy, BCE) and regression losses, weighted by w_i:

$$\mathcal{L}_{logical} = \frac{1}{N_{gt}} \sum_{i=1}^{N} \Big(\mathcal{L}_{cls}(P_i^S, P_i^T) + \mathcal{L}_{box}(B_i^S, B_i^T, w_i) \Big) \tag{9}$$

The regression loss $\mathcal{L}_{box}$ is dynamically modulated by w_i:

$$\mathcal{L}_{box} = \lambda_{L1} \cdot w_i \cdot ||B_i^S - B_i^T||_1 + \lambda_{GIoU} \cdot \psi(w_i) \cdot (1 - \text{GIoU}(B_i^S, B_i^T)) \tag{10}$$

To emphasize high-quality supervision, we apply a non-linear transformation $\psi(w_i)$ to the GIoU weight:

$$\psi(w_i) = \begin{cases} w_i^2, & \text{if } w_i < 0.5 \\ w_i^{0.5}, & \text{if } w_i \geq 0.5 \end{cases} \tag{11}$$

This suppresses low-confidence predictions while maintaining gradients for reliable ones.

The final training objective combines the standard detection loss $\mathcal{L}_{det}$ (via Hungarian matching) with the distillation loss:

$$\mathcal{L}_{total} = \mathcal{L}_{det}(P^S, Y_{gt}) + \lambda_{distill} \cdot \mathcal{L}_{logical}(P^S, P^T) \tag{12}$$

where $\lambda_{distill} = 0.075$ balances the tasks. $\mathcal{L}_{logical}$ is applied solely during the distillation phase to transfer knowledge.

IV. Experimental Results and Analysis

A. Datasets

We evaluate our method on two datasets: the self-constructed SC7 and a refined subset of the public STBD-08 [15]. To ensure label reliability, we systematically cleaned the data by removing severe motion blur and correcting erroneous annotations, ensuring that detection difficulty remains representative of real-world scenarios. Specifically, the SC7 dataset was collected from real-world university classrooms featuring diverse frontal and side-view perspectives, comprising 2,392 images with 26,057 annotated instances across seven fine-grained behaviors. The STBD-08 dataset is derived from 131 classroom videos and contains 8,717 images with 257,728 annotations, where we focus on five key behaviors consistent with our analysis goals. Table I summarizes the instance distribution.

B. Implementation Details

Experiments were conducted on a workstation equipped with an NVIDIA RTX 4090 D (24GB) GPU, using Python 3.8.20 and PyTorch 2.0.1. Models were trained for 200 epochs with a batch size of 16 and an input resolution of 640×640. Early stopping was employed to mitigate overfitting. We utilized the AdamW optimizer with an initial learning rate of 1×10^{-4}, momentum of 0.9, and weight decay of 1×10^{-4}, accompanied by a 20-epoch warmup period.

C. Evaluation Metrics

We adopt standard metrics to comprehensively evaluate model performance from two perspectives:

- **Efficiency**: Computational cost and storage are measured via parameter count (Params), floating-point operations (GFLOPs), and model size (MB). Inference speed is reported as Frames Per Second (FPS) on an RTX 4090 D.
- **Accuracy**: Detection performance is evaluated via Precision (P), Recall (R), Mean Average Precision at IoU=0.50 (mAP_{50}), and the standard COCO metric $\text{mAP}_{50:95}$ (averaged over IoU thresholds 0.50–0.95).

All accuracy metrics are expressed in percentages (%).

D. Detection Model Comparison

We benchmarked RERC-DETR against state-of-the-art detectors to demonstrate its effectiveness. As presented in Table III, RERC-DETR achieves a competitive $\text{mAP}_{50:95}$ of 81.6% on the STBD-08 dataset, evidencing its robustness in small object detection. On the SC7 dataset, it attains balanced precision and recall with an mAP_{50} of 87.5%, underscoring its efficacy in diverse classroom scenarios. These findings highlight the advantages of our multi-scale feature fusion and uncertainty-aware query selection mechanisms in enhancing detection performance across varying object scales. Moreover, the integration of Quality-Aware Logical Distillation (RERC-DETR + QALD) further elevates performance, achieving 82.1% $\text{mAP}_{50:95}$ on STBD-08, surpassing all compact detectors.

E. Backbone Performance Comparison

To identify the optimal feature extractor, we evaluated various backbone architectures. For a fair comparison, all backbones were equipped with the default Efficient Hybrid Encoder and Decoder of RT-DETR. Table II presents the trade-offs between model complexity and detection accuracy. Our proposed RAD-Net effectively reduces parameters and computational cost compared to ResNet18, while maintaining competitive or superior accuracy. Notably, RAD-Net achieves an $\text{mAP}_{50:95}$ of 81.1% on the STBD-08 dataset, surpassing other compact backbones such as EfficientVit and Starnet. These results validate RAD-Net's efficacy in balancing efficiency with representational power, rendering it highly suitable for real-time applications in resource-constrained environments.

F. Ablation Study

Table IV elucidates the contributions of RAD-Net and RCFPN. Deploying RAD-Net alone significantly reduces complexity (10.45M params, 33.4 GFLOPs) while achieving 81.1% $\text{mAP}_{50:95}$ on STBD-08 at 179 FPS. Integrating RCFPN further boosts accuracy to 81.6% while maintaining real-time speed (156 FPS). On SC7, the full configuration attains 87.5% mAP_{50}. Although the gain on the smaller SC7 dataset is marginal, the consistent improvement on the larger STBD-08 confirms RCFPN's robustness in resolving feature aliasing. Thus, the full architecture offers the best generalization across varying data scales. These results corroborate

TABLE I: Instance counts per behavior for SC7 and STBD-08 datasets.

Dataset	Listening	Eating/Drinking	Using Phone	Using Tablet	Using Laptop	Reading	Writing	Raising Hand	Standing
SC7	11520	195	5786	1661	6053	523	319	-	-
STBD-08	117561	-	-	-	-	57985	73405	4687	4090

TABLE II: Backbone Comparison Experiments on Different Datasets.

Backbone	Params(↓)	GFLOPs(↓)	SC7				STBD-08			
			P	*R*	mAP_{50}	$mAP_{50:95}$	*P*	*R*	mAP_{50}	$mAP_{50:95}$
ResNet18 [16]	19878180	57.0	**88.9**	80.3	86.2	72.3	89.0	89.7	**93.7**	80.8
StarNet [17]	11992964	31.8	83.4	87.2	**87.9**	**74.4**	89.1	89.8	93.4	80.3
MobileNetV4 [18]	11315428	39.5	86.5	82.4	86.1	70.3	**90.6**	88.1	93.0	80.0
FasterNet [19]	10813224	28.5	87.2	80.9	84.3	68.6	89.4	**89.9**	93.3	80.5
EfficientViT [20]	10707748	**27.2**	81.3	71.0	74.7	58.9	89.2	89.8	93.1	79.5
RAD-Net	**10452868**	33.4	85.1	**87.3**	87.8	73.8	89.7	89.8	93.6	**81.1**

TABLE III: Detection Model Comparison Experiments on Different Datasets.

Dataset	Model	Type	P	R	mAP_{50}	$mAP_{50:95}$
SC7	YOLOv5	CNN	81.0	82.3	86.9	71.0
	YOLOv6 [21]	CNN	83.7	83.2	87.5	73.2
	YOLOv8	CNN	86.3	79.4	87.5	72.2
	YOLOv11	CNN	82.9	80.8	86.7	72.1
	SSDLite [22]	CNN	73.9	77.4	79.8	61.6
	RTMDet [23]	CNN	76.1	80.8	85.4	66.5
	RT-DETR [7]	Trans.	**88.9**	80.3	86.2	72.3
	RERC-DETR	Trans.	86.4	84.4	87.5	73.8
	RERC-DETR + QALD	Trans.	85.5	**88.1**	**89.7**	**76.9**
STBD-08	YOLOv5	CNN	87.8	88.7	93.1	79.7
	YOLOv6 [21]	CNN	85.7	88.3	92.2	77.9
	YOLOv8	CNN	88.0	89.5	93.7	80.7
	YOLOv11	CNN	87.3	88.8	93.2	79.7
	SSDLite [22]	CNN	85.3	84.8	89.8	69.4
	RTMDet [23]	CNN	82.9	87.7	91.6	74.9
	RT-DETR [7]	Trans.	89.0	89.7	93.7	80.8
	RERC-DETR	Trans.	**90.3**	89.0	**94.0**	81.6
	RERC-DETR + QALD	Trans.	89.5	**90.8**	93.9	**82.1**

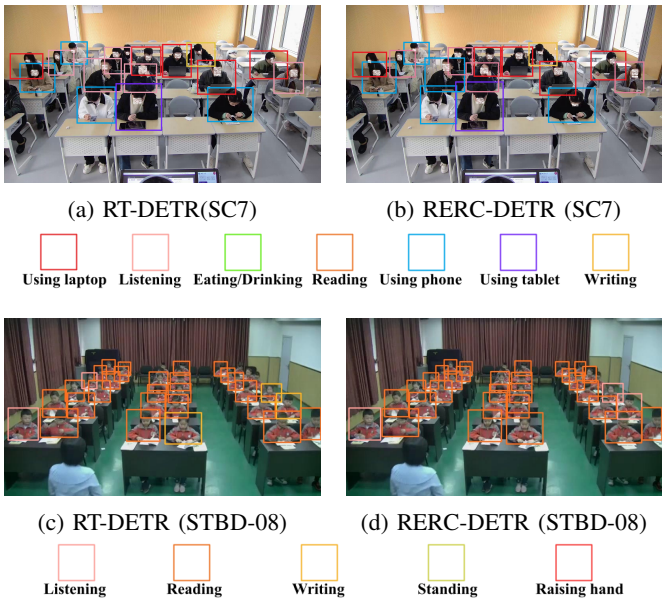
the effectiveness of the proposed components in addressing classroom behavior recognition challenges while ensuring high efficiency.

G. Logical Distillation Experiment

To verify the effectiveness of the proposed Quality-Aware Logical Distillation strategy, we conducted experiments using RT-DETR with different backbones (ResNet34, ResNet50, ResNet101) as teachers, while the student model is our proposed RERC-DETR. As shown in Table V, the distillation process consistently improves the student model's performance across both datasets. On the challenging SC7 dataset, using ResNet50 as the teacher yields the best result, boosting the student's $mAP_{50:95}$ to 76.9%, a significant improvement over the baseline. Interestingly, the strongest teacher (ResNet101) does not necessarily lead to better results on SC7, possibly due to the domain gap or overfitting on smaller datasets. However, on the larger STBD-08 dataset, the strongest teacher (ResNet101) leads to the highest performance of 82.1% $mAP_{50:95}$. These results demonstrate that our logical distillation strategy effectively transfers structural knowledge from powerful teachers to the lightweight student, enhancing its detection capability without increasing inference cost.

H. Visualization of Detection Results

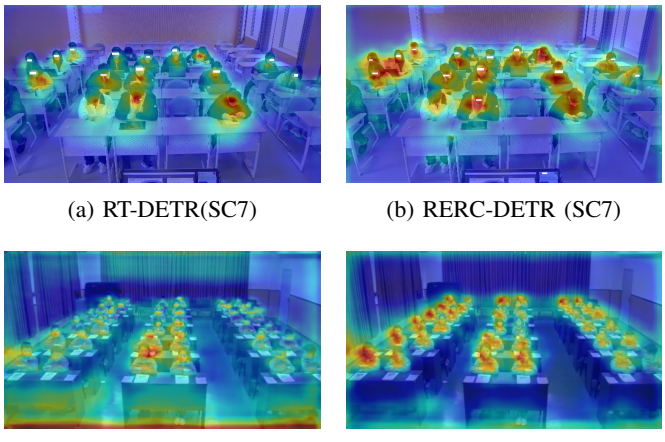
Fig. 5 presents qualitative detection results in classroom scenes, showing the localization performance of RT-DETR and RERC-DETR on SC7 and STBD-08. Fig. 6 further visualizes the heatmaps generated after detection. Compared with RT-DETR, RERC-DETR localizes dense and small student behaviors more accurately and focuses more consistently on behavior-relevant regions, demonstrating stronger robustness in complex classroom environments.



(a) RT-DETR(SC7) (b) RERC-DETR (SC7)

(c) RT-DETR (STBD-08) (d) RERC-DETR (STBD-08)

Fig. 5: Qualitative visualization of detection results.



(a) RT-DETR(SC7) (b) RERC-DETR (SC7)

(c) RT-DETR (STBD-08) (d) RERC-DETR (STBD-08)

Fig. 6: Heatmap visualization after detection.

V. Conclusion and Future Work

This paper introduces RERC-DETR, a lightweight Transformer detector that resolves the accuracy–efficiency dilemma in classroom behavior recognition. By integrating

TABLE IV: Ablation Study Experiments on Different Datasets.

Dataset	RAD-Net	RCFPN	Params(↓)	GFLOPs(↓)	Size(↓)	FPS	P	R	mAP_{50}	$mAP_{50:95}$
SC7	✗	✗	19878180	57.0	38.6	87	**88.9**	80.3	86.2	72.3
	✗	✓	19494532	84.4	38.1	50	86.2	83.8	86.6	**73.9**
	✓	✗	**10452868**	**33.4**	**20.7**	**92**	85.1	**87.3**	**87.8**	73.8
	✓	✓	11155492	39.3	22.4	80	86.4	84.4	87.5	73.8
STBD-08	✗	✗	19878180	57.0	38.6	166	89.0	89.7	93.7	80.8
	✗	✓	19494532	84.4	38.1	122	89.8	89.7	93.9	81.3
	✓	✗	**10452868**	**33.4**	**20.7**	**179**	89.7	**89.8**	93.6	81.1
	✓	✓	11155492	39.3	22.4	156	**90.3**	89.0	**94.0**	**81.6**

TABLE V: Quality-Aware Logical Distillation Experiments on Different Datasets.

Dataset	Teacher	Params(M)	GFLOPs	Student			
				P	R	mAP_{50}	$mAP_{50:95}$
SC7	ResNet34	31.1	88.8	**87.6**	85.0	88.2	75.1
	ResNet50	41.9	125.7	85.5	**88.1**	**89.7**	**76.9**
	ResNet101	60.9	186.3	85.8	83.2	88.2	75.5
STBD-08	ResNet34	31.1	88.8	89.0	89.7	93.1	80.8
	ResNet50	41.9	125.7	**89.5**	89.4	93.2	81.0
	ResNet101	60.9	186.3	**89.5**	**90.8**	**93.9**	**82.1**

the efficient RAD-Net backbone, RCFPN, and a Quality-Aware Logical Distillation strategy, our method effectively preserves small-object semantics and leverages teacher knowledge without extra inference cost. Extensive experiments on STBD-08 and SC7 datasets confirm that RERC-DETR outperforms state-of-the-art compact detectors, providing a robust solution for real-time smart education analytics.

In future work, we aim to incorporate temporal modeling to better distinguish fine-grained actions and expand our benchmark to include diverse instructional scenarios.

REFERENCES

[1] Zhifeng Wang, Wenxing Yan, Chunyan Zeng, Yuan Tian, and Shi Dong, "A unified interpretable intelligent learning diagnosis framework for learning performance prediction in intelligent tutoring systems," *International Journal of Intelligent Systems*, vol. 2023, no. 1, pp. 4468025, 2023.

[2] Xu Xin, Yu Shu-Jiang, Pang Nan, Dou ChenXu, and Li Dan, "Review on a big data-based innovative knowledge teaching evaluation system in universities," *Journal of Innovation & Knowledge*, vol. 7, no. 3, pp. 100197, 2022.

[3] Qin Ni, Yanjin Zhu, Lingying Zhang, Xiaochen Lu, and Lei Zhang, "Leverage learning behaviour data for students' learning performance prediction and influence factor analysis," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 5, pp. 2422–2433, 2023.

[4] Karen M La Paro, Robert C Pianta, and Megan Stuhlman, "The Classroom Assessment Scoring System: Findings from the prekindergarten year," *The Elementary School Journal*, vol. 104, no. 5, pp. 409–426, 2004.

[5] Zhifeng Wang, Jialong Yao, Chunyan Zeng, Wanxuan Wu, Hongmin Xu, and Yang Yang, "YOLOv5 enhanced learning behavior recognition and analysis in smart classroom with multiple students," in *Proceedings of the International Conference on Intelligent Education and Intelligent Research (IEIR)*. IEEE, 2022, pp. 23–29.

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko, "End-to-end object detection with Transformers," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 213–229.

[7] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen, "DETRs beat YOLOs on real-time object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 16965–16974.

[8] Ned A Flanders, "Abstracted from interaction analysis in the classroom a manual for observers," *Classroom Interaction Newsletter*, vol. 3, no. 2, pp. 1–5, 1968.

[9] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[10] Wei Lu, Xiangpeng Liu, Yulin Peng, Maria Kyrarini, Kang An, and Yuhua Cheng, "PACR-DETR: A real-time end-to-end object detector for behavior recognition in various classroom scenarios," *IEEE Transactions on Instrumentation and Measurement*, 2025.

[11] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, 2021.

[12] Zhihui Li, Pengfei Xu, Xiaojun Chang, Luyao Yang, Yuanyuan Zhang, Lina Yao, and Xiaojiang Chen, "When object detection meets knowledge distillation: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 10555–10579, 2023.

[13] Feilong Tang, Zhongxing Xu, Qiming Huang, Jinfeng Wang, Xianxu Hou, Jionglong Su, and Jingxin Liu, "DuAT: Dual-aggregation transformer network for medical image segmentation," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 2023, pp. 343–356.

[14] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao, "YOLOv9: Learning what you want to learn using programmable gradient information," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2024, pp. 1–21.

[15] Jinhua Zhao and Hongye Zhu, "CBPH-Net: A small object detector for behavior recognition in classroom scenarios," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–12, 2023.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[17] Xu Ma, Xiyang Dai, Yue Bai, Yizhou Wang, and Yun Fu, "Rewrite the stars," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 5694–5703.

[18] Danfeng Qin, Chas Leichner, Manolis Delakis, Marco Fornoni, Shixin Luo, Fan Yang, Weijun Wang, Colby Banbury, Chengxi Ye, Berkin Akin, et al., "MobileNetV4: Universal models for the mobile ecosystem," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2024, pp. 78–96.

[19] Jierun Chen, Shiu-hong Kao, Hao He, Weipeng Zhuo, Song Wen, Chul-Ho Lee, and S-H Gary Chan, "Run, don't walk: Chasing higher FLOPs for faster neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 12021–12031.

[20] Xinyu Liu, Houwen Peng, Ningxin Zheng, Yuqing Yang, Han Hu, and Yixuan Yuan, "EfficientViT: Memory efficient vision transformer with cascaded group attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 14420–14430.

[21] Chuyi Li, Lulu Li, Hongliang Jiang, Kaiheng Weng, Yifei Geng, Liang Li, Zaidan Ke, Qingyuan Li, Meng Cheng, Weiqiang Nie, et al., "YOLOv6: A single-stage object detection framework for industrial applications," *arXiv preprint arXiv:2209.02976*, 2022.

[22] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4510–4520.

[23] Chengqi Lyu, Wenwei Zhang, Haian Huang, Yue Zhou, Yudong Wang, Yanyi Liu, Shilong Zhang, and Kai Chen, "RTMDet: An empirical study of designing real-time object detectors," *arXiv preprint arXiv:2212.07784*, 2022.