

Original Research Article

Optimized YOLOv11n for real-time behavior detection in classrooms

Jiangteng Long^{1,2}, M. Kazem Chamran¹, Jiaojie Song²

¹Faculty of Information Technology, City University Malaysia, Petaling Jaya, Selangor, 46100, Malaysia

²Jiangxi Vocational & Technical College of Information Application, Nanchang, Jiangxi, 330043, China

Abstract: This research proposes a multi-strategy joint optimization framework for classroom behavior detection based on the improved YOLOv11n. It combines the Normalized Wasserstein Distance (NWD), Shape - Intersection over Union (Shape-IoU), and Shape-NWD fusion loss function, and introduces the C2PSA attention enhancement module and a full - lifecycle privacy protection mechanism. In response to the problems of gradient vanishing and inaccurate matching of the traditional CIoU loss for small targets, this paper introduces the NWD index, replacing the overlap - based similarity measure with the Gaussian distribution distance to effectively stabilize the training gradient of sparse - pixel small targets.

Keywords: image recognition; YOLOv11n; real-time response; deep learning; human pose estimation

1. Introduction

Artificial intelligence technology has achieved considerable development, and its combination with education has become increasingly close. Artificial intelligence-assisted teaching has been widely used and has become an inevitable trend in the future development of education ^[1].

This study endeavors to incorporate the modified YOLOv11 artificial intelligence model in the improvement of classroom teaching ^[2]. Through the analysis of human movements in video surveillance by the YOLOv11 model, real-time monitoring and identification of students' behaviors in the classroom are conducted. The classroom behaviors of students are summarized and analyzed to assist teachers in adjusting their teaching strategies promptly and to facilitate schools' evaluation of the learning effect in the classroom.

2. Methodology

This study proposes an optimization of the YOLOv11n architecture for real-time behavior detection in classrooms. At the data processing level, considering the perspective distortion characteristics of classroom surveillance, an SCB-Dataset^[3] containing samples from multiple angles such as top-down and side views is constructed. Through multi-dimensional technological innovations, the detection bottleneck of the traditional YOLOv11 model in complex teaching scenarios is overcome. By adopting the NWD (Normalized Wasserstein Distance)^[4] metric mechanism, reconstructing the detection head structure, and combining the Shape-IoU^[5] loss function with the Shape-NWD joint optimization strategy, the problems of small targets and target confusion are effectively addressed^[6].

2.1. Dataset description

The SCB-Dataset3 (Student Classroom Behavior Dataset)^[7] is specifically designed for classroom behavior detection tasks, composed of 5,686 images and 45,578 labels, it focuses on six behaviors: Raising hands, reading, writing, using mobile phones, looking down, and leaning on the desk. The images are extracted from real classroom surveillance footage with a resolution of 1080p. The dataset exhibits a diverse class distribution, with behaviors like raising hands accounting for 15%, reading for 20%, and sleeping for 10%.

2.2. Input preprocessing

YOLO model performs well in real-time and large target detection, it has some drawbacks in recognizing small targets. By combining new loss functions such as Normalized Wasserstein Distance (NWD), Shape-IoU loss function, and Shape-NWD fusion improvement, the YOLO model can be optimized to enhance the performance of small target detection.

To quantitatively verify the effectiveness of the proposed NWD, Shape-IoU, and Shape-NWD optimization strategies. The only variable is the bounding box regression loss function, which ensures that the performance differences can be uniquely attributed to the improved loss optimization mechanism. The detailed configuration of each experimental group is shown in **Table 1**: Experimental Group Table.

Table 1. Experimental group table.

Experimental Group	Model Configuration	Modification Scope
I	Original YOLOv11n	No modification, standard CIoU loss as baseline
II	YOLOv11n + NWD	Replace the original regression loss with NWD metric loss
III	YOLOv11n + Shape-IoU	Replace the original regression loss with Shape-IoU loss
IV	YOLOv11n + Shape-NWD	Adopt the proposed Shape-NWD fusion joint optimization loss

2.3. Evaluation methods

The evaluation system uses mAP (mean Average Precision) to measure detection accuracy and FPS (Frames Per Second) to measure inference speed, forming a two-dimensional evaluation framework of accuracy-speed. Use of the same dataset (SCB-Dataset3) and evaluation metrics to ensure fairness in comparison experiments. Detection performance metric mAP (mean Average Precision), speed and resource metric FPS (Frames Per Second). The metric system covers detection accuracy, fine positioning capability, error suppression ability, inference efficiency, and model computational complexity. Specifically, mAP@0.5 and mAP@0.5:0.95 are utilized to evaluate overall detection performance and precise bounding box regression capability.

2.4. Ethics and constraints

This study established a multi-level informed consent system before data collection. Informed consent from students, parents, and teachers is prioritized to ensure ethical and transparent monitoring. Voluntary principle: Participants (and their guardians) are clearly informed that they can withdraw from the study at any time and that their collected data will be permanently deleted from all storage media within 72 hours after withdrawal. Transparency principle: Explain the purpose of the camera, data flow and protection measures to students in plain language to avoid psychological pressure caused by "hidden monitoring"^[9].

3. Experimental results and analysis

The experimental results strongly verify the significant positive synergistic effect between NWD and Shape-IoU mechanisms. The overall mAP@0.5 of the Shape-NWD model is improved by 2.9%, which exceeds the simple superposition gain of individual NWD (+1.7%) and Shape-IoU (+1.4%) optimization, proving that the fusion strategy can mutually compensate the respective limitations of the two single-loss functions. Benefiting from the dual enhancement of NWD scale robustness and Shape-IoU adaptive scale weighting, the detection accuracy of the mobile phone playing category is drastically improved from 64.2% to 73.2%, with a maximum increment of 9.0% among all behavior categories. In addition, slender targets (hand raising) and occluded targets (head bending) also achieve optimal localization performance under the joint constraint of distribution distance and geometric shape penalty. In terms of inference efficiency, the FPS of the Shape-NWD model is only slightly reduced from 128.6 to 128.3. Since all improvements belong to training-stage loss optimization without introducing extra computational parameters in the inference process, the proposed strategy realizes a zero-cost accuracy promotion, which perfectly maintains the high real-time performance of the original YOLOv11n and meets the deployment requirements of classroom real-time monitoring systems. The final result is as follows **Table 2** Performance Comparison between.

Table 2. Performance comparison between.

Configuration	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Mobile Phone AP (%)	Hand Raising AP (%)	Head Bending AP (%)	FPS
YOLOv11n (CIoU)	77.4	53.9	64.2	82.3	78.6	128.6
YOLOv11n +NWD	79.1	56.2	70.8	83.1	79.8	128.4
YOLOv11n+Shape-IoU	78.8	56.8	67.5	85.6	80.4	128.5
YOLOv11n+Shape-NWD	80.3	58.4	73.2	86.1	81.7	128.3

4. Conclusion

This study takes complex classroom student behavior detection as the research object, aiming at the problems of poor environmental robustness, insufficient small-target detection accuracy, low regression precision of irregular human targets, and difficulty in balancing accuracy and real-time performance existing in the original YOLOv11n model in educational scenarios. The goal of this research revolves around optimizing the YOLOv11 model to address the challenges of student behavior detection in complex classroom environments, including occlusion, small target recognition, and environmental noise, while also considering privacy protection.

References

- [1] O. Zawacki-Richter, V.I. Marín, M. Bond and F. Gouverneur, "Systematic Review of Research on Artificial Intelligence Applications in Higher Education – Where Are the Educators?" *Int. J. Educ. Technol. High. Educ.*, vol. 16, no. 1, pp. 1-27, 2019, doi: 10.1186/s41239-019-0171-0.
- [2] R. Khanam and M. Hussain, "YOLOv11: An Overview of the Key Architectural Enhancements," *Corr*, vol. abs/2410.17725, pp. 1-9, 2024.
- [3] F. Yang, "SCB-dataset: A Dataset for Detecting Student Classroom Behavior," *Arxiv (Cornell University)*, vol. abs/2304.02488, 2023, doi: 10.48550/arxiv.2304.02488.
- [4] Y. Zhou, K. Cao, De Li and J. Piao, "Fine-YOLO: A Simplified X-ray Prohibited Object Detection Network Based on Feature Aggregation and Normalized Wasserstein Distance," *Sensors*, vol. 24, 2024, doi: 10.3390/s24113588.
- [5] H. Zhang and S. Zhang, "Shape-IoU: More Accurate Metric Considering Bounding Box Shape and Scale.," *Corr*, vol. abs/2312.17663, 2023, doi: 10.48550/arxiv.2312.17663.
- [6] C. Rao, W. Li, X. Xie and G. Cheng, "A Shape-Based Quadrangle Detector for Aerial Images.," *Pattern Recognition and Computer Vision, Prcv 2023, Pt IV*, vol. 14428, pp. 368-379, 2024, doi: 10.1007/978-981-99-8462-6_30.
- [7] F. Yang and T. Wang, "SCB-Dataset3: A Benchmark for Detecting Student Classroom Behavior.," *Corr*, vol. abs/2310.02522, 2023, doi: 10.48550/arxiv.2310.02522.
- [8] I. Bartoletti, "AI in Education," *The Ethics of Artificial Intelligence in Education*, pp. 74-90, 2022, doi: 10.4324/9780429329067-5.