

Student-YOLOv8: A Privacy-Preserving Deep Learning Framework for Real-Time Student Detection and Automated Attendance Monitoring in Bangladeshi Classroom Environments

Md. Alamgir Hossain ¹

¹ Department of Computer Science and Engineering, Dhaka International University, Dhaka-1212, Bangladesh
Email: alamgir.cse@diu.ac

ARTICLE INFO

Keywords:

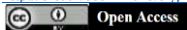
Deep Learning, YOLOv8, Student Detection, Privacy-Preserving Attendance, Classroom Analytics, Computer Vision, Bangladesh.

Vol.01, No.05, 2026

<https://doi.org/10.5281/zenodo.23432315>

Copyright © 2026 by author(s) and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CCBY4.0).

<http://creativecommons.org/licenses/by/4.0/>



ABSTRACT

Automatically tracking attendance in crowded university classrooms remains a major challenge. Currently, most schools rely on manual roll calls, which waste valuable class time and make it incredibly easy for students to skip class and have buddy fake their attendance. This paper presents Student-YOLOv8, a privacy-preserving deep learning framework built on the YOLOv8s architecture for real-time student detection and automatic headcount-based attendance estimate. Unlike biometric or facial-recognition systems, Student-YOLOv8 treats each student as a generic object, eliminating biometric data collection. The DIU-Student dataset-the first annotated classroom image benchmark from a Bangladeshi university-comprises 788 original images augmented to 3,320 samples using Mosaic, MixUp and geometric transformations. Trained for 300 epochs on an NVIDIA T4 GPU, the model achieves precision of 0.768, recall of 0.739, F1-score of 0.753, and mAP@0.5 of 0.737. The system gives per-frame attendance counts without any identity resolution, offering a scalable, ethically compliant and institutionally deployable solution for educational institutions across South Asia.

Introduction

While these traditional methods are simple to set up and don't require any special technology, they come with well-known drawbacks-especially in large classes. For starters, the roll call process itself eats up a significant chunk of teaching time. On top of that, mistakes made during manual recording often end up in official academic records and the whole system is wide open to proxy attendance, where one student simply answers or signs in for a friend who isn't there [1], [2]. In introductory undergraduate courses which frequently exceed one hundred students at both public and private universities in the region these inefficiencies are magnified, ultimately degrading both teaching quality and record accuracy.

The rapid advancement of computer vision and deep learning offers a practical pathway to automate these administrative tasks. Modern object detection models have demonstrated exceptional performance across a board spectrum of domains, ranging from autonomous vehical perception [3] to industrial inspection and medical imaging [4]. Among these, the You Only Look Once (YOLO) family of single-stage detectors has emerged as a leading solution. By combining near real-time inference speeds with competitive accuracy. YOLO models are particularly well-suited for resources constrained and latency-sensitive environments [5-7]. Prior studies in automatic attendance have investigated two prevalent approaches: the hardware-based approach, and the vision-based approach. Through the use of RFID tokens, fingerprint readers or smart cards [8], [9], or the facial recognition algorithms used are camera based and they are deep face [10], deep ID [11] and joint recognition [12]. proxying concerns, and are more complex than software-based systems. Hardware-based systems decrease proxying, require student engagement, and are more complex than software-based systems, which also create hygiene concerns. concerns for biometric readers, and require capital investment. The facial recognition systems provide a are more passive experience, but also pose considerable privacy threats. Storing and working with The General Data Protection Regulation (GDPR) is one of the regulations that govern facial

biometric data. One of the regulations governing facial biometric data is the General Data Protection Regulation (GDPR). Increasingly, GDPR and similar institutional policies of data governance are being passed along (GDPR2016) [13]. Universities in Bangladesh had adopted, thereby making it difficult from legal and ethical point of view to deploy.

This paper adopts another paradigm of estimating attendance as an object detection, not an identity verification problem but a crowd counting problem. Beacon Hill is set to become a major new landmark. detect and count the number of students that are in a classroom frame, it can estimate attendance. Without stating name of the students. This eliminates the need to collect biometric data altogether, Meeting the privacy needs at the architectural scope and not as an afterthought. This article introduces Student-YOLOv8, a mobile object detection system based on the YOLOv8s network. We introduce Student-YOLOv8, a detection system based on YOLOv8s. [7] was designed and trained using the DIU-Student dataset which was collected from real-world data. The classroom settings in Dhaka International University. The system is able to do real-time inference, Works with live video streams and static images, and generates an attendance count automatically Based on the number of detected boxes per frame. This work makes specific contributions of the following:

i) Model YOLOv8s detection pipeline for domain adaptation, PANet multi-scale neck and anchor-free detection head, specially designed and optimized for dense. In the context of a Bangladeshi classroom scenario, this paper examines how students can be detected when they are occluded and under varying scales.

ii) The DIU-Student dataset; the first annotated classroom image benchmark from a Bangladeshi university, which is an extensive collection of 788 original images increased to 3,320 by a wide-ranging sampling strategy. Using the offline augmentation strategy (Mosaic, MixUp, horizontal flip, rotation, scaling), with a fixed 80/10/10 train-validation-test partition.

iii) A built-in real-time automatic attendance counter that counts the following: bounding-box detections per frame without biometric identification step, satisfying the criteria of the constraint. Bounding-box detections per frame without biometric identification step, satisfying the criteria of the constraint. privacy-by-design requirements architecturally.

iv) A systematic empirical assessment showing precision 0.768 and recall 0.739, The mAP@0.5 on a held-out test set is 0.737, and the F1-Score is 0.753, comprehensive with a held-out test set. Analysis of different occlusion and density conditions.

The rest of this paper is organised as follows. The related work is clarified in Section 2. Section 3 describes the overall research methodology. After that, Section 4 presents results and analysis. The findings and limitations are discussed in Section 5. Finally, the conclusion is described in Section 6.

Related Work

Hardware-Based Attendance Systems

Initially automated attendance systems were primarily hardware based. RFID In the systems, proposed in the work [8] students can be identified with passive cards swiping at the entrance. points. The RFID system cannot detect mid-session leaves. and students must have active token with them. The biometric systems include fingerprint readers and iris. These were then followed by scanners, vein-pattern recognition for better uniqueness. guarantees [14]. These have high errors under controlled conditions, but Classroom setup challenges: queuing at scanners takes setup time, surfaces are exposed to the wind, and so on. explain hygiene issues in post pandemic situations, and data collected is under strict protection. regulatory protections [13]. All indoor localization systems share the common feature of relying on the localisation systems can be used to track even to location, but only with dense hardware deployments. and active involvement of students.

Facial Recognition-Based Attendance

The surge of deep convolutional networks in the field of face recognition led to the interest of the face recognition system. With the development of deep convolutional networks for face recognition, it is inspiring people to study the face recognition system. camera-based biometric attendance. DeepFace [10] and Classroom to adopt a similar approach. Classroom followed the VGGFace [11] which achieved near-human verification performance. attendance proposals [12], [15]. Such systems are compared with

the detected the database faces against enrolled databases. They work well under controlled conditions, but break down in water. ambiguous classroom lighting and non-frontal head poses, partial occlusion, the accuracy decreases with camera-distance blur and with all but camera-distance blur. More importantly, facial recognition systems save facial information in the form of: biometric reference templates that are sensitive personal information according to GDPR [13]. Lakshmi et al. [15] introduced a facial recognition attendance system by using There was mention of improvements in automation, but also dependence on storing biometric data, via IoT cameras. This work does not follow this paradigm, but rather explicitly diverges from it. In this section, we will discuss object detection approaches and their variants—namely the YOLO family. Object detection is an alternative formulation: instead of detecting individuals, detect that There is someone there and count them. Two-stage detectors such as R-CNN [16] and by the way, following the region proposal networks, faster R-CNN [17] achieve good accuracy with faster speed. They are classified by the author, they are very heavy in terms of computation, and cannot be used in real classrooms. Single-stage detectors solve this problem by taking a single pass. SSD [18] and the YOLO family [5-7] have demonstrated that real-time multi-object detection can be accomplished with tolerable compromises in accuracy. YOLOv7 [6] introduced trainable bag-of-freebies to improve training efficiency without increasing inference cost. YOLOv8 [7] replaced the anchor-based detection head with an anchor-free decoupled head, adopted a new CSPDarknet backbone variant, and integrated PANet-based feature fusion for improved multi-scale detection. Domain-specific YOLO applications in Bangladeshi contexts have yielded strong results: Hossain et al. [3] deployed YOLOv7 for vehicle detection on the Padma Bridge achieving mAP@0.5 of 96.86%, establishing YOLO viability in South Asian real-world deployments. Kaur and Singh [19] applied YOLOv6 for traffic sign detection, reporting rapid convergence and superior localisation. No prior work has addressed student detection in Bangladeshi classroom environments.

Student Monitoring Systems

Several works have addressed broader classroom monitoring. Nag et al. [20] built a real-time deep learning classroom monitoring system emphasising dataset quality as a primary performance determinant and the inadequacy of general-purpose datasets for dense classroom scenarios. Vidhya et al. [21] proposed a student engagement tracking system combining detection with attention classification. These works extend beyond presence detection to behavioural analysis but continue relying on individual-level identification. The present work explicitly excludes identity resolution as a design constraint.

Positioning of the Present Work

Student-YOLOv8 occupies a distinct position in the literature. It requires no student-side hardware (cf. RFID), stores no biometric data (cf. facial recognition), produces per-frame bounding-box detections rather than only scalar headcounts (cf. generic counters), and is trained on a dedicated local dataset reflecting Bangladeshi university conditions. To the best of our knowledge, this is the first study to construct and evaluate a privacy-by-design student detection framework specifically trained on a Bangladeshi classroom dataset.

Methodology

The proposed methodology follows a five-stage pipeline see in Fig. 1. First of all, we collected real images data of the students from Dhaka International University classroom. Then the images are labelled by each of student id. After label the images, the image processing technique applied on the labelled images. Thereafter, the data augmentation technique applied on the dataset to enrich the dataset size. Then the dataset splitting into three parts: 80% as training, 10% as validation, and 10% as testing. Finally, the Student-YOLOv8 model applied on the dataset to detect the students acutely. The each of research methodology described in below by partly: (i) dataset collection, (ii) data preprocessing and annotation, (iii) augmentation strategy, (iv) model architecture and training, and (v) automated attendance counting.

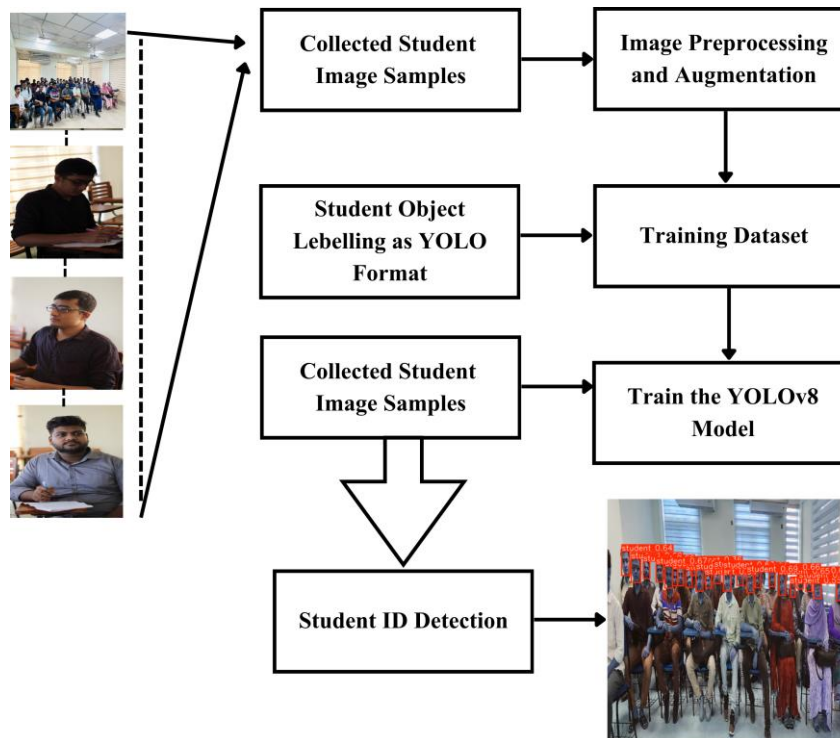


Fig. 1: Overall research methodology flowchart.

Dataset Collection

The DIU-Student dataset was constructed specifically for this study. Institutional permission was obtained from Dhaka International University (DIU) to capture classroom images across multiple teaching rooms, lecture halls, and laboratory spaces. Images were captured using high-resolution smartphone cameras (12-48 MP) and DSLR cameras under realistic classroom conditions multiple camera angles, seating configurations, student densities (5-43 students per frame), and lighting conditions (natural, fluorescent, mixed). A total of 788 original images were collected. All students visible in images are treated as members of a single class-Student-labelled with bounding boxes. No facial features, names, or identity information were recorded or retained. Institutional ethical clearance was obtained prior to data collection. Sample images are shown in Fig. 2.

Table 1. DIU-Student Dataset Summary

Category	Details
Source	Dhaka International University Classrooms, Dhaka, Bangladesh
Single Class Label	Student (43 sub-identifiers, Class ID: 0–42)
Image Resolution	640 × 640 pixels (YOLOv8 input)
Original Images	788
Dataset Split	633 Training (80%) / 78 Validation (10%) / 77 Testing (10%)
Augmented Training Images	3,165
Total Post-Augmentation	3,320 (3,165 Training + 78 Validation + 77 Testing)
Annotation Tool	LabelImg; normalised YOLO format
Ethical Clearance	Granted by DIU; no biometric data retained



Fig. 2: Sample images from the DIU-Student dataset showing diverse classroom conditions.

Annotation and YOLO Format

All student instances in the 788 original images were manually annotated using LabelImg, with each annotation stored in the normalised YOLO format: class_id x_center y_center width height, where all coordinate values are normalised by image dimensions. The single class identifier (Student, class_id = 0) is applied uniformly. Bounding boxes tightly enclose visible student bodies including occluded portions; partially visible students at frame edges were annotated if at least 30% of the body was visible. Annotation quality was validated through cross-reviewer spot-checks.

Data Augmentation Strategy

We used five different types of augmentation technique that enrich the training sample from 633 to 3,165. Validation and test sets were kept unchanged to ensure unbiased evaluation. The augmentation techniques are described in below:

- i) **Mosaic Augmentation:** Four randomly selected images combined into a single 640×640 composite, exposing the model to varied scales, positions, and co-occurrence patterns per sample.
- ii) **MixUp Augmentation:** Linear pixel-level interpolation of two images at a random blending ratio, improving generalisation by regularising decision boundaries.
- iii) **Horizontal Flipping** ($p = 0.5$): Mirrors images left-right, doubling effective viewpoint diversity.
- iv) **Random Rotation** ($\pm 10^\circ$): Simulates camera tilt and non-horizontal capture angles common in CCTV installations.
- v) **Scaling and Translation** ($\pm 20\%$): Random zoom and spatial shift to simulate varying camera distances and off-centre framing.
- vi) **Gaussian Blur, Median Blur, and CLAHE** (via Albumentations): Simulate camera defocus, motion blur, and low-contrast/uneven-lighting conditions respectively.

Model Architecture

Student-YOLOv8 is based on the YOLOv8s (small) variant [7], selected for its favourable accuracy-speed trade-off. The architecture (see Fig. 3) comprises three functional modules.

Backbone:

CSPDarknet: The backbone extracts hierarchical feature representations from the 640×640 input. YOLOv8s employs Cross-Stage Partial Darknet (CSPDarknet) with C2f modules modified CSP bottlenecks with two convolutional streams improving gradient flow and feature reuse. The backbone produces five feature maps (P1–P5) at progressively reduced spatial resolutions; P3 (80×80), P4 (40×40), and P5 (20×20) are passed to the neck. A Spatial Pyramid Pooling-Fast (SPPF) layer appended after the final backbone stage aggregates context at multiple receptive field sizes, critical for detecting students at widely varying distances from the camera.

Neck:

PANet Multi-Scale Feature Fusion: The neck aggregates backbone features across scales for multi-scale detection. YOLOv8s employs Path Aggregation Network (PANet) combining a top-down FPN path where high-level semantic features from P5 are upsampled and merged with P4 and P3 with a bottom-up PAN path reversing this flow. This bidirectional aggregation ensures both semantic richness (from deep layers) and spatial precision (from shallow layers) are available at each detection scale.

Head:

Anchor-Free Decoupled Detection: YOLOv8 adopts an anchor-free decoupled detection head, directly predicting bounding-box parameters (centre coordinates and dimensions) and class confidence scores as separate branches, avoiding the hyperparameter sensitivity of anchor-based designs. Detection is performed at three scales corresponding to P3 (small), P4 (medium), and P5 (large objects). The attendance count per frame is derived as the total number of bounding boxes predicted with confidence above $\tau = 0.25$.

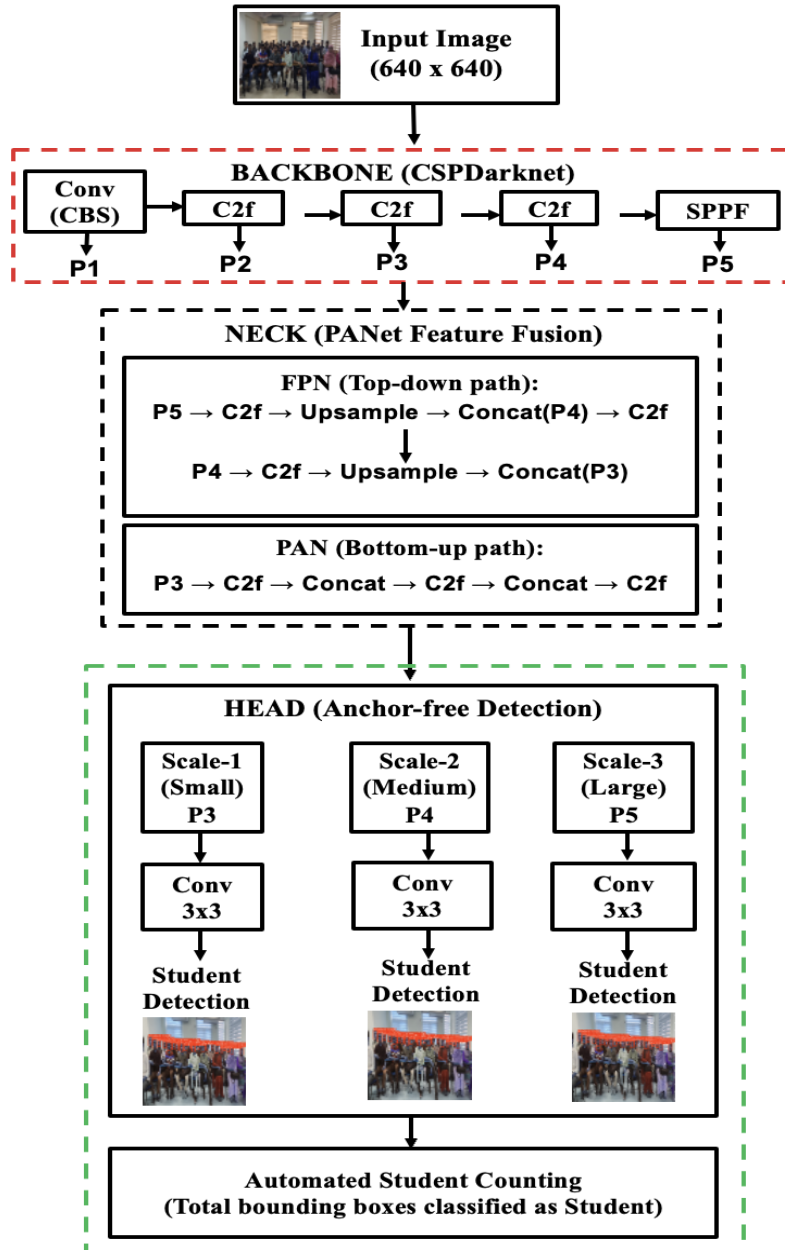


Fig. 3: Proposed Student-YOLOv8 model architecture

Training Configuration

Training was conducted on Google Colab Pro with an NVIDIA T4 GPU (15 GB VRAM) using Ultralytics YOLOv8 framework (v8.2.103), Python 3.10, and PyTorch 2.5.1 (CUDA 12.2). Automatic Mixed Precision (AMP) training was enabled to maximise GPU utilisation. Hyperparameters are summarised in Table 2

Table 2. Training Configuration Summary

Hyperparameter	Value
Model Variant	YOLOv8s (small)
Input Resolution	640 × 640 pixels
Epochs	300
Batch Size	16
Optimizer	AdamW
Learning Rate	0.000208 (auto-selected)
Momentum	0.9
Weight Decay	0.0005
Warmup Epochs	3
Confidence Threshold	0.25
NMS IoU Threshold	0.45
Pre-trained Weights	YOLOv8s COCO initialisation
Hardware	NVIDIA T4 GPU, 15 GB VRAM (Google Colab)
Framework	Ultralytics v8.2.103 / PyTorch 2.5.1 / CUDA 12.2

The composite loss function minimised during training comprises three terms: (i) CIOU bounding-box regression loss, penalising deviations in predicted centre coordinates, dimensions, and aspect ratio; (ii) binary cross-entropy classification loss for the single Student class; and (iii) Distribution Focal Loss (DFL) for refined bounding-box edge prediction.

Ethical Considerations: Full institutional permission was granted by Dhaka International University for data collection. No biometric data, facial features, or student identities were captured or stored at any stage. Images were used solely for training the object detection model and were collected under approved institutional ethical guidelines.

Automated Attendance Counting Mechanism

Following object detection on a classroom frame, the attendance count equals the number of predicted bounding boxes with confidence above $\tau = 0.25$ and class label Student. Since the system operates as a single-class detector, this count directly estimates students present. For video deployment, per-frame counts are aggregated and a temporal median computed over a configurable 30-frame window to reduce transient detection noise. The output is a scalar attendance estimate per time window delivered without any student identity resolution, satisfying privacy-by-design requirements.

Results and Analysis

Experimental Setup

The model was trained on Google Colab using an NVIDIA T4 GPU with 15 GB VRAM. The implementation used the Ultralytics v8.2.103 framework with Python 3.10 and PyTorch 2.5.1 (CUDA 12.2). Training ran for 300 epochs with batch size 16 and input resolution 640 × 640 pixels. AdamW was automatically configured with learning rate 0.000208 and momentum 0.9. Automatic Mixed Precision (AMP) training was enabled to maximise GPU memory utilisation. The augmented training set comprised 3,165 images; the validation and test sets remained fixed at 78 and 77 images respectively throughout all experiments.

Evaluation Metrics

Model performance is assessed using the standard MS COCO evaluation protocol [22]: Precision (P) and Recall (R) at $IoU \geq 0.5$; F1-Score as their harmonic mean; mAP@0.5 (mean average precision at IoU threshold 0.5, the primary detection benchmark); and mAP@0.5:0.95 (mean AP averaged over IoU thresholds $\{0.50, 0.55, \dots, 0.95\}$, the stricter COCO primary metric penalising imprecise localisation).

Table 3. Performance Metrics of Student-YOLOv8 on the Test Set

Metric	Score
Precision (P)	0.768
Recall (R)	0.739
F1-Score	0.753
mAP@0.5	0.737
mAP@0.5:0.95	0.258

Quantitative Performance

Overall, Student-YOLOv8 demonstrates strong performance across all the metrics in Table III on the held-out test set. (77 images, all of which were not seen in training and validation). A precision of 0.768 means that 76.8% false positives are mostly caused by background regions that have texture similar to the student instances; Seated Human Figures (chairs, bags, camera artefacts). The recall of 0.739 indicates that 73.9% Most false non-detections are due to; Attributed to students that are significantly covered by other students in the front rows. The F1-score of 0.753 Has a good balance of precision and recall that is suitable for the application of attendance counting. The main detection quality indicator is mAP@0.5 of 0.737, which shows good performance at the typical threshold used for IoU estimation in attendance. The mAP@0.5:0.95 of 0.258 reflects the The challenge in precisely delineating the bounding-boxes in dense occlusion. This is in line with those in crowd-detection literature, which indicate that the mAP@0.5:0.95 range is 0.20–0.35. The number of pedestrian detectors that are routinely reported for single-class pedestrian detectors is 17.00. occlusion [23]. Some representative examples of results of detection are presented in Fig. 5.

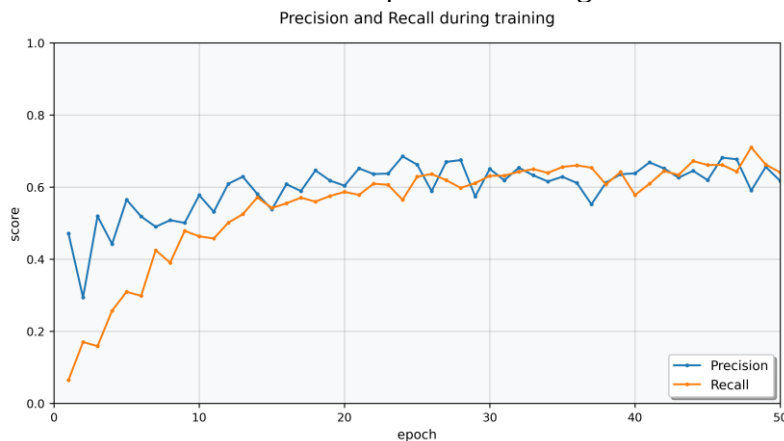


Fig. 4: Precision–recall curve of Student-YOLOv8 on the test set.

Training Behavior Analysis

The training loss curves for bounding-box regression (CIoU), classification (BCE), and distribution. All of the focal loss (DFL) exhibited downward trend for 300 epochs, which is consistent. The optimization without divergence and substantial overfitting. We validated the model and looked at the validation loss, which was equal to training loss. closely throughout. After



Fig. 5: Student-YOLOv8 detections on unseen classroom images with confidence scores.

epoch 50, the precision has been improving monotonically, as a result of progressive Advanced student body configuration rejection of background noise. Recall exhibited greater as students are mostly identifiable at a distance, the early variance is due to difficulty in detecting the partially occluded students, and stabilising after. epoch 120. The Precision-Recall curve (Fig. 4) illustrates a smooth trade-off curve that agrees with the reported mAP@0.5.

Confusion Matrix Analysis

The confusion matrix analysis (see Fig. 6) indicates that most of the predictions were correct. This corresponds to 382 correct Student detections, for reliable classification. False positive examples (115 examples) originate from background areas which have human-like visual features like Chairs and bags can be seen in the background, but they are partially obscured. The number of false negatives (135 missed) is mainly students. In densely packed classrooms behind the front row, with visibility limited by other students, there is a decrease in body area visible. Visible body area is decreased in dense classrooms at the back of the class with inter-student overlap. Below the level of detection. The main causes of errors are:

- i) No clear sightlines in seating arrangements with high density of students.
- ii) Classroom setting with similar background (chairs, bags) There are partial views of students in the frame corners and at the back of the frame. There is a partial view of students within the corners of the frame and at the back rows.

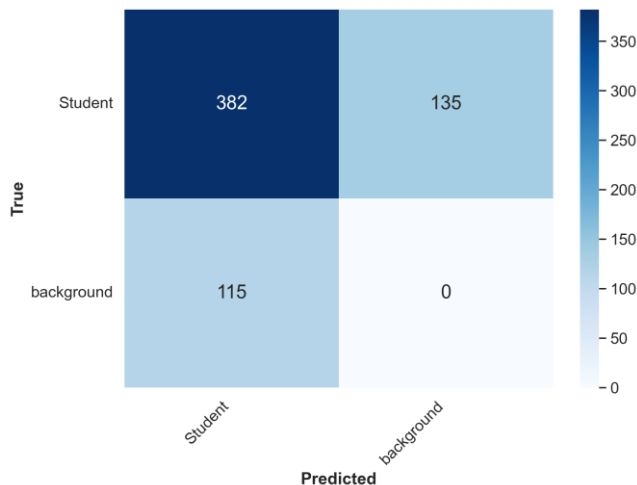


Fig. 6: Confusion matrix for Student-YOLOv8 on the held-out test set.

These difficulties notwithstanding, the overall detection reliability is still high, substantiating the practical feasibility. In authentic classroom contexts.

Comparative Evaluation

Student-YOLOv8 is compared to state-of-the-art YOLO-based models in Tab. IV. The vehicle-YOLOv7 [3] and YOLOv6 [19] were tested on transportation-domain datasets-domains in which objects are well separated, have low mutual occlusion, and occur in structured spatial arrangements. In contrast, this work addresses severe inter-object occlusion, irregular spatial distributions, and high intra-class appearance variability.

Table 4. Comparison with State-of-the-Art YOLO-Based Detectors

Ref.	Model	Domain	P	R	F1	mAP@0.5
[3]	Vehicle-YOLOv7	Traffic / Bridge	0.972	0.980	0.968	0.969
[19]	YOLOv6	Traffic Signs	0.952	0.962	0.955	0.935
Proposed	Student-YOLOv8	Classroom (Dense)	0.768	0.739	0.753	0.737

Student-YOLOv8' models are competitive and reliable in detecting the target domain. The difference between the metric and transportation-domain models is similar to what has been reported. of dense multi-person detection versus sparse, well-separated vehicle detection. Understanding of cross domain comparison must be considered in context, not equivalent-condition benchmarking.

Discussion

Practical Implications

Student-YOLOv8's major practical impact is a privacy-preserving attendance system that can be institutionally used at Bangladeshi universities without any changes in the data. governance policies or investment in biometric infrastructure. The system requires only a single One ceiling or wall-mounted camera and one processing unit per classroom, able to process YOLOv8s. inference. In a 50-minute lecture for 40 students it takes an estimate of 5-8 minutes for manual attendance. The amount of time spent in lecture-about 10-16% of lecture length. Recovering this Time is a pedagogically measurable resource, especially when there are many students in the class.

Limitations

Several limitations merit acknowledgment. Firstly, the DIU-Student dataset is limited to a single institution. Domain shift may reduce performance when deploying to classrooms with substantially different architectural features, lighting conditions, or cultural dress norms. Dataset expansion to multiple Bangladeshi universities is the highest-priority future work. Secondly, the system is a presence detector, not an identity resolver. It can estimate how many students are present but cannot determine which specific students are absent. This limits direct applicability to generating roster-level attendance records without integration with a complementary entry-point system (e.g., RFID-based entry scanning). Thirdly, the mAP@0.5:0.95 of 0.258 indicates strict bounding-box localisation remains a challenge in the most densely occluded scenes. While this does not impair the primary attendance counting use case, future architectures incorporating explicit occlusion-aware components, multi-institution dataset expansion, and edge device deployment are identified as the primary directions for future research.

Conclusion

This paper presented Student-YOLOv8, a privacy-preserving deep learning framework for real-time student detection and automated attendance monitoring in Bangladeshi classroom environments. The model, trained on the DIU-Student dataset — the first annotated classroom benchmark from a Bangladeshi university and achieved precision of 0.768, recall of 0.739, F1-score of 0.753, and mAP@0.5 of 0.737 on a held-out test set. The system processes live video streams and automatically produces per-frame student counts without capturing, storing, or processing any biometric or identity data, satisfying privacy-by-design requirements at the architectural level. The study takes into account three main disadvantages of existing automated attendance systems: Identification systems, and the issue of security, will also be discussed. The challenges of hardware-based token systems, privacy concerns of biometric facial identification systems and security issues will also be discussed. The lack of any education-specific benchmark data for the education system in Bangladesh, and the lack of any existing

evidence of children being identified as having special educational needs. contexts. By making attendance an object detection rather than an identity verification process, Student-YOLOv8 is scalable, ethical and institution deployable solution Which can be adopted without any regulatory hurdles. Future work will be directed towards expanding the DIU-Student dataset to more than one Bangladeshi. universities, assessing backbone variants based on transformer (RT-DETR, YOLOv10) for better performance. occlusion handling, deployment to edge devices (Jetson Nano, Raspberry Pi) with the model, etc. Provide real-time classroom applications, inference speed benchmarking, and expand it to multi-class detection, which separates students, teachers and empty seats for better classroom analytics.

References

1. O. Masurekar, O. Jadhav, P. Kulkarni, and S. Patil, "Real time object detection using YOLOv3," *International Research Journal of Engineering and Technology (IRJET)*, vol. 07, no. 03, pp. 3764–3768, 2020. https://www.academia.edu/44168535/IRJET_Real_Time_Object_Detection_Using_YOLOv3
2. G. Danezis and S. Gurses, "A critical review of 10 years of privacy technology," in *Surveillance Cultures: A Global Surveillance Society*, 2010. <https://homes.esat.kuleuven.be/~sgurses/publications.html>
3. M. A. Hossain, M. A. Rahman, R. Khanom, S. Sultana, and A. Hossain, "Real-time vehicle detection and classification on the Padma multipurpose bridge using deep learning," in *Proceedings of the National Conference on Interdisciplinary Mathematics (NCIM)*, 2023, pp. 1–6. <https://ieeexplore.ieee.org/document/10212549>
4. A. Farid et al., "A fast and accurate real-time vehicle detection method using deep learning," *Applied Sciences*, vol. 13, no. 5, p. 3059, 2023. <https://doi.org/10.3390/app13053059>
5. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788. <https://doi.org/10.1109/CVPR.2016.91>
6. C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022. <https://doi.org/10.1109/CVPR52729.2023.00721>
7. G. Jocher, A. Chaurasia, and J. Qiu, "YOLO by ultralytics," <https://github.com/ultralytics/ultralytics>, 2023. <https://github.com/ultralytics/ultralytics>
8. R. Want, "An introduction to RFID technology," *IEEE Pervasive Computing*, vol. 5, no. 1, pp. 25–33, 2006. <https://doi.org/10.1109/MPRV.2006.2>
9. L. M. Ni, Y. Liu, Y. C. Lau, and A. P. Patil, "LANDMARC: Indoor location sensing using active RFID," *Wireless Networks*, vol. 10, pp. 701–710, 2004. <https://doi.org/10.1023/B:WINE.0000044029.06344.dd>
10. Y. Taigman, M. Yang, M. A. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1701–1708. <https://doi.org/10.1109/CVPR.2014.220>
11. O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2015. <https://doi.org/10.5244/C.29.41>
12. K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016. <https://doi.org/10.1109/LSP.2016.2603342>
13. European Parliament, "General data protection regulation (GDPR)," *Official Journal of the European Union*, vol. L 119, pp. 1–88, 2016. <https://eur-lex.europa.eu/eli/reg/2016/679/oj>
14. A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 4–20, 2004. <https://doi.org/10.1109/TCSVT.2003.818349>
15. M. V. Lakshmi, M. S. Siva Priya, K. Manivannan, and M. Vivekanandan, "Real-time implementation of an automated student attendance monitoring system with computer vision

technology,” in Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC), 2024, pp. 1651–1654. <https://ieeexplore.ieee.org/document/10689967>

16. R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 580–587. <https://doi.org/10.1109/CVPR.2014.81>

17. S. Ren, K. He, R. Girshick, and J. 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, 2017. <https://doi.org/10.1109/TPAMI.2016.2577031>

18. W. Liu et al., “SSD: Single shot multibox detector,” in Proceedings of the European Conference on Computer Vision (ECCV), ser. Lecture Notes in Computer Science, vol. 9905. Springer, 2016, pp. 21–37. https://doi.org/10.1007/978-3-319-46448-0_2

19. R. Kaur and J. Singh, “Local regression based real-time traffic sign detection using YOLOv6,” in Proceedings of the International Conference, Dec. 2022, pp. 522–526. <https://ieeexplore.ieee.org/document/10074236>

20. A. Nag, J. Ferdous, S. B. Baha Uddin, and R. Debnath, “Real-time and automated student attention monitoring system in classroom using deep learning techniques,” in Proceedings of the International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (ICCIT), Cox’s Bazar, Bangladesh, Dec. 2024. <https://ieeexplore.ieee.org/document/11021871>

21. K. Vidhya et al., “Smart focus: A deep learning-based student engagement tracking system,” in Proceedings of the International Conference on Intelligent Robotics, Control, and Automation (ICIRCA), Jun. 2025. <https://doi.org/10.1109/ICIRCA65293.2025.11089732>

22. T.-Y. Lin et al., “Microsoft COCO: Common objects in context,” in Proceedings of the European Conference on Computer Vision (ECCV), ser. Lecture Notes in Computer Science, vol. 8693. Springer, 2014, pp. 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

23. Q. Wang, J. Gao, W. Lin, and Y. Yuan, “Learning from synthetic data for crowd counting in the wild,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8198–8207. <https://doi.org/10.1109/CVPR.2019.00839>