A Deep Learning-Based Method for Classroom Crowd Counting and Localization

Qin Ding[1] and Chunyan Yu[2](✉)

[1] Anhui University of Science and Technology, Anhui, China
[2] Chuzhou University, Anhui, China
yuchy@chzu.edu.cn

Abstract. In order to count the students' seating distribution and attendance in offline classroom, which is a better response to the students' learning and teaching situation. Based on the deep learning method, we propose a crowd localization and counting model for students' seating area in offline classroom. Firstly, we choose YOLOv8 to improve it, adding the SENet attention module after the backbone network reinforce the role of important channels and speed up model learning, designing a simple and efficient feature fusion method, using the anchor size and the number of detected heads which are more suitable for classroom scenarios and compute the overall loss of the model by using the loss of confidence and the loss of regression of prediction frames. Enhancement methods with Mosaic and cutout data to increase the generalization ability of the model. The improved network achieved 95.405% precision, 92.808% recall and 96.159% mAP on SCUT-HEAD Dataset and C University Dataset.

Keywords: Crowd Localization · Crowd Counting · Seating Distribution

1 Introduction

In all kinds of teaching activities, face-to-face offline teaching is still the most dominant form of teaching, and student attendance and seat distribution better reflect the learning situation of students and the teaching situation of teachers. However, in this scenario, both the answering and software check-in methods will take up the classroom time and affect the teacher's teaching. Therefore, utilizing surveillance video in the classroom is an effective and convenient way to take attendance. In addition, this method can quickly and significantly mark the position of students in the classroom, and obtain the seating distribution of students, and many studies have shown that classroom seating distribution has a great relationship with students' motivation in class, participation [1], interaction intensity [2] and so on.

Classroom crowd estimation under camera has the following challenges: (1) The size of head scale varies, the target sitting in the front row close to the camera is larger, while the head scale of students in the back row is very small, and it is often easy to miss or misdetect the target during detection. (2) High level of shelter, the distribution of students in the classroom environment is relatively dense, and high occlusion is easy to occur

W. Hong and G. Kanaparan (Eds.): ICCSE 2023, CCIS 2023, pp. 173–184, 2024.
https://doi.org/10.1007/978-981-97-0730-0_16

due to the inconsistency of each student's movements such as raising and lowering their heads. Highly overlapping targets have similar features and are difficult to distinguish during detection. (3) Blurred images, affected by hardware equipment, the cameras in most universities do not have high pixels. Especially for students sitting at the back position, the camera can only illuminate the outline of the human head, which is difficult to judge even with the naked eye in scenes with darker background environment such as desks.

Combining the existing methods and the above problems, the main work and contributions of this paper are as follows: (1) Designing a new feature fusion method, which can accelerate the model convergence speed and improve the model accuracy. (2) Proposing a crowd estimation model for classroom environments, which is used to detect the number and location of students in the classroom to assist teachers in attendance and teaching analysis.

2 Related Work

Deep learning based methods for crowd estimation can be categorized into two types: detection-based methods and map-based methods.

2.1 Detection-Based Methods

The detection method of crowd estimation is the application of target detection on heads. The method first detects the each head in the image and then counts it. In recent years, [3] employs a multi-column architecture with top-down feature modulation, which allows the network to jointly process multi-scale information, facilitating the network's ability to accurately locate people's heads. In particular, the bounding box of each head can be predicted with only point labeling information. Since the YOLO model is a target detection model for a wide range of applicable scenarios, many people choose to improve the YOLO model for head detection e.g., [4] used a modified yolov4 model for head detection and counting, which prevent the occurrence of missed detection by attenuating the score of adjacent detection frames. [5] proposed a self-training method that uses point annotations to directly supervise object centroids which makes data annotation faster. However, the method has high sensitivity to the annotation point density and is more suitable for scenes with uniform target distribution.

2.2 Map-Based Methods

The map-based methods refers to the transformation of the head estimation problem into a density map estimation problem. This method does not need to detect each person explicitly. [6] calculates the loss of pairs of density maps at different scales to achieve multi-scale head estimation, and fuses the density maps at different scales to improve the accuracy and stability of head estimation. Apart from this, a multi-task learning approach is used to obtain both the number and location information of the headcounts, which leads to a more comprehensive understanding of the crowd distribution. [7] introduces a deep prior to improve the accuracy of crowd localization and counting. The depth prior can provide additional information about the size and location of heads, thus helping the model to better understand the crowd scene. At the same time the method relies on depth information to assist in the localization and counting of heads. This means that without reliable depth information, the method may not be able to accurately localize and count heads. This limits its applicability on certain scenarios or devices. In [8], a topological constraint is proposed to address topological errors in crowd localization, and to enforce the constraint, a novel persistence loss based on persistent homotopy theory is proposed.

The existing deep learning based crowd estimation algorithms are improved to address the problems of existing methods and crowd estimation in classroom environments. The main improvements are (1) using smaller and denser anchors to alleviate the problem of model miss-detection and mis-detection on small targets, occlusion, and fuzzy problems. (2) Use bilateral three-path fusion of information from deep semantic features, shallow texture features, and mid-level features, and use anchor of two scales and three aspect ratios to detect heads of different scales and poses. (3) Add the SENet module after feature extraction to allow the model to focus on the features of the important channels to speed up learning and improve model accuracy.

3 Proposed Method

In order to solve the problem of high occlusion, multi-scale and low resolution problem of targets in classroom environment, this paper uses deep learning method to detect the number and location of students. The network structure in this paper refers to the feature extraction method of YOLOv8 and uses the SENet attention mechanism to strengthen the role of important features and weaken the role of background features. A bilateral three-path aggregation network is designed for feature fusion, fusing features from different sensory fields, and two detection heads are used to detect multi-scale targets in the classroom scene. The overall network structure is shown in Fig. 1.

The model firstly resizes the incoming image to 640*640 using non-distortion way, and then downsamples the image once, after which the image features are extracted using four C2F and downsampling operations, in which we obtain feature maps of three scales, i.e. $S_1 = 1/8$, $S_2 = 1/16$, and $S_3 = 1/32$ of the original image size. The C2F structure is a feature extraction module with residual structure in YOLOv8, which ensures lightweight while obtaining richer gradient flow information. Then, using the SENet module to respectively preprocess the S_1, S_2, S_3 features layers to enhance the important feature channels, and obtain the E_1, E_2, E_3 which are the same size as the input. The E_1, E_2, and E_3 feature layers are then fused using a bilateral three-path aggregation network. Finally, we get 20×20 size feature layer and 80×80 size feature layer, and use these two feature layers for head detection, the large feature layer is used mainly to detect small scale targets that are far away from the camera and the small feature layer is used mainly to detect large targets that are close to the camera.

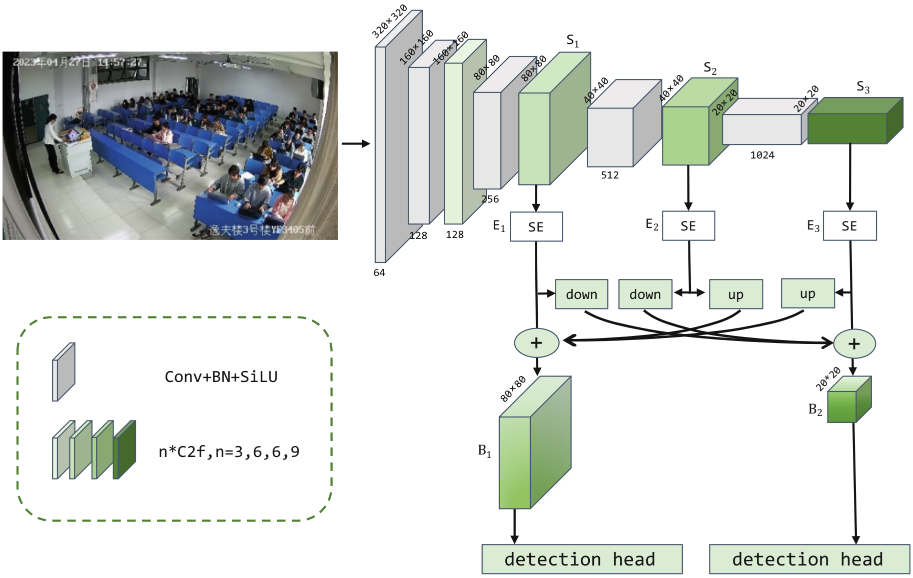

Fig. 1. Overall network architecture design

3.1 SENet Module

Since the model obtains a large number of features in the feature extraction stage, some features have important contributions to the model goal, while some features have smaller contributions, so before feature fusion, the SENet [9] module is added after each of the S_1, S_2, S_3 feature maps respectively, to speed up the model learning speed and to increase the detection accuracy. SENet is a lightweight network that considers the relationship between feature channels and automatically acquires the importance of each feature channel by learning to enhance the features that are important to the current task or suppress the unimportant features. SENet realizes the above functions by Squeeze module and Exciation module. The structure of SENet module is shown in Fig. 2.

The SENet module use global pooling for the three scales of the feature special to get $1 \times 1 \times 256$, $1 \times 1 \times 512$, and $1 \times 1 \times 1024$ feature layers, respectively, and then goes through a fully-connected layer to reduce the c channels of the model to c/r channels to reduce the amount of computation, using the ReLU activation function, and the second fully-connected layer serves to recover the number of channels, and the Sigmoid function is used to limit the range of values between 0 and 1, which is equivalent to the weights of the feature layer. Finally, the weights are multiplied with the original feature layer to get the new feature layer.

3.2 Feature Fusion Module for Bilateral Three-Path Module

Bilateral three-path aggregation module is uesd to fuse the global and local information of different sensory fields on the shallow, middle and deep networks. Firstly, the E_3 feature layer is up-sampled by 4 times and the E_2 feature layer is up-sampled by 2 times,

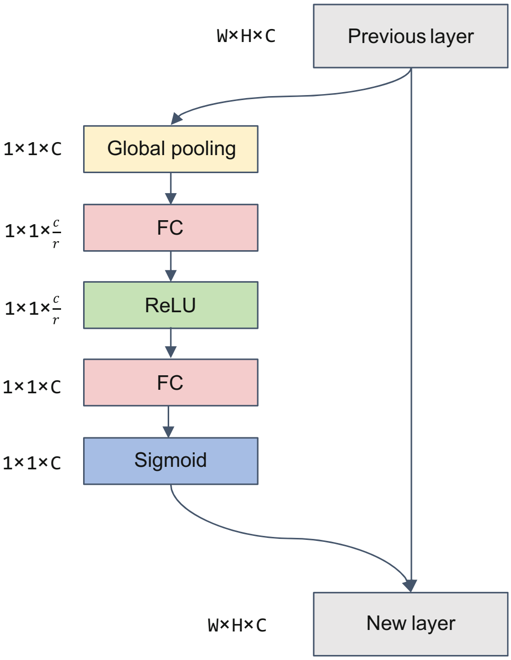

Fig. 2. SENet module

and the two obtained feature layers are fused with the E_1 feature layer to obtain B_1 feature layer of 80 × 80 size, which fuses the information of each layer. Then the E_1 feature layer is downsampled by 4 times, the feature layer of E_2 is downsampled by 2 times, and the obtained result is fused with the E_3 feature layer to obtain B_2 feature layer of 20 × 20 size, which fuses the information of each layer. The feature fusion module for bilateral three-path is shown in the Fig. 3.

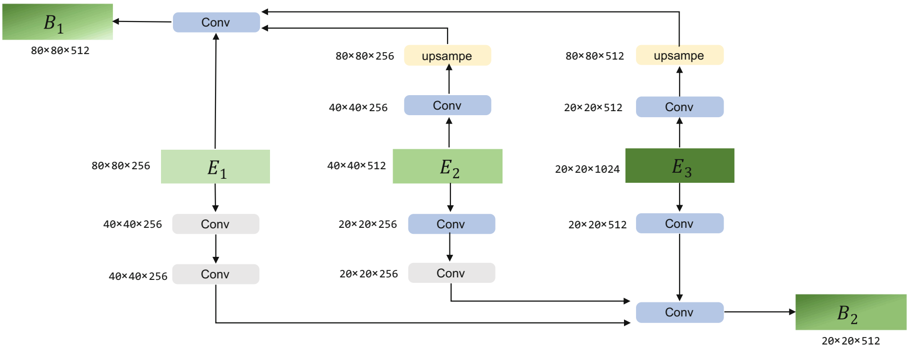

Fig. 3. Feature fusion module for bilateral three-path module. The convolution of the gray squares indicates downsampling, and the convolution of the blue squares indicates channel count adjustment.

3.3 Loss Function

The loss function of this network is jointly determined by confidence loss and bounding box loss, the loss function formula is

$$Loss = \alpha * bbox_{loss} + \beta * obj_{loss} \tag{1}$$

where α and β are the proportion of confidence loss and bounding box loss, respectively, 1 and 0.1.

The confidence of each bounding box indicates the reliability of this bounding box, the larger the value means that the model thinks that this is the probability of the target is larger. The model uses the loss of confidence is the cross-entropy loss, the confidence loss of the bounding box consists of the positive sample loss of confidence and the negative sample loss of confidence together, the confidence of the bounding box is

$$obj_{loss} = \sum\nolimits_{i=0}^{K*K} \sum\nolimits_{j=0}^{M} I_{ij}^{obj}[log(C_i)] + \sum\nolimits_{i=0}^{K*K} \sum\nolimits_{j=0}^{M} I_{ij}^{noobj}[log(C_i)] \tag{2}$$

k is the grid size, M is the number of predefined anchors for each grid, I_{ij}^{obj} indicates whether the bounding box is a positive sample or not, when it is a positive sample, the change of value takes 1, otherwise it is 0.

The bounding box loss is a measure of how much the predicted box overlaps with the real box in target detection, the bounding box loss used in this paper is CIoU [10]. It is given by the formula

$$\mathrm{bbox_{loss}} = \sum\nolimits_{i=0}^{K*K} \sum\nolimits_{j=0}^{M} I_{ij}^{obj} L_{CIoU} \tag{3}$$

$$L_{CIoU} 1 - IoU + \frac{\rho^2(b, \widehat{b})}{c^2} + \alpha v \tag{4}$$

v is used to measure the consistency of the aspect ratio and it is expressed as

$$v = \frac{4}{\pi^2}(arctan\frac{\widehat{w}}{\widehat{h}} - arctan\frac{w}{h})^2 \tag{5}$$

α is the weight parameter, which has the expression

$$\alpha = \frac{v}{(1 - IoU) + v} \tag{6}$$

CIoU considers the distance and aspect ratio similarity between target and anchor on the basis of traditional IoU, which is more consistent with the target frame regression mechanism.

4 Experiment

4.1 Experimental Environment and Parameter Settings

Three anchors with different aspect ratios are set for each detector head to detect heads with different postures, and the preset anchor sizes for the two detector heads are [12,12, 14,17, 22,33] and [22,22, 25,32, 35,52]. Momentum is set to 0.937 and the number of iteration rounds is set to 200. The initial learning rate is 0.01, and the learning rate is dynamically adjusted during training based on the loss value of the model and the performance of the validation set. When the loss value of the model decreases or the performance of the validation set improves, the learning rate will decrease accordingly; while when the loss value of the model increases or the performance of the validation set decreases, the learning rate will increase accordingly. This allows the model to converge more stably and efficiently during the training process.

4.2 Experimental Dataset and Preprocessing

The data used in this experiment are (1) SCUT-HEAD Dataset [11], which contains two parts, PartA and PartB, with a total of 4342 images, PartA has a total of 2,000 camera data of a university, and PartB has 2342 images of students. (2) Surveillance data of 48 sections of eight classes in the last semester of 2023 in C University. For each class, 12 frames of images are extracted at regular intervals during class time, resulting in a cumulative total of 572 images. Consequently, the experiment utilizes a combined dataset of 4,914 images.

The SCUT-HEAD dataset encompasses a wide range of authentic student classroom scenarios, including diverse classroom surveillance camera data from different classrooms and time periods, as well as online classroom images from various contexts. The dataset exhibits a varied distribution of target quantities, ranging from 0 to 162, and encompasses images of varying sizes, ranging from 228 × 166 to 6280 × 4710 pixels. Notably, the image size distribution demonstrates a balanced representation across the dataset. Additionally, the C University dataset specifically focuses on classroom surveillance camera data from C University, with each image uniformly sized at 1920 × 1080 pixels. These comprehensive datasets effectively capture the complexities of real-world student classroom environments, encompassing diverse angles, crowd densities, lighting conditions, resolutions, and classroom settings. This rich and diverse collection of data serves as a valuable resource for academic research and analysis in the field.

In this experiment, Mosaic [12] and cutout [13] method data enhancement methods are used to increase the generalization ability of the model.

Mosaic is a data enhancement technique commonly used in target detection tasks. It generates new training samples by stitching several different images together. Specifically, the Mosaic data enhancement method stitches together four randomly selected images in a certain ratio to form a new synthetic image. At the same time, the corresponding target frames need to be adjusted and transformed accordingly. In this way, more diverse and complex training samples can be generated by the Mosaic data enhancement method, which provides more perspective and background changes and enhances the generalization ability of the model. Mosaic data enhancement makes the training samples have more diversity and complexity, provides more background information, and enhances the generalization ability of the model. This helps the model to better adapt to various scenarios and changes.

The Cutout data enhancement method is a simulated random occlusion data enhancement method that improves the generalization ability of the model by cutting out random rectangular regions in the image during the training process. Specifically, Cutout randomly selects some pixels in the image and sets them to zero. There exists a 50% probability that the erased rectangular region is not exactly in the original image. The Cutout method does not have non-informative pixels during training and does not require the generation of additional images to increase the size of the training set as compared to traditional data enhancement methods. Therefore no extra cost is incurred during training. The results of training using Mosaic and Cutout methods are shown in Fig. 4.

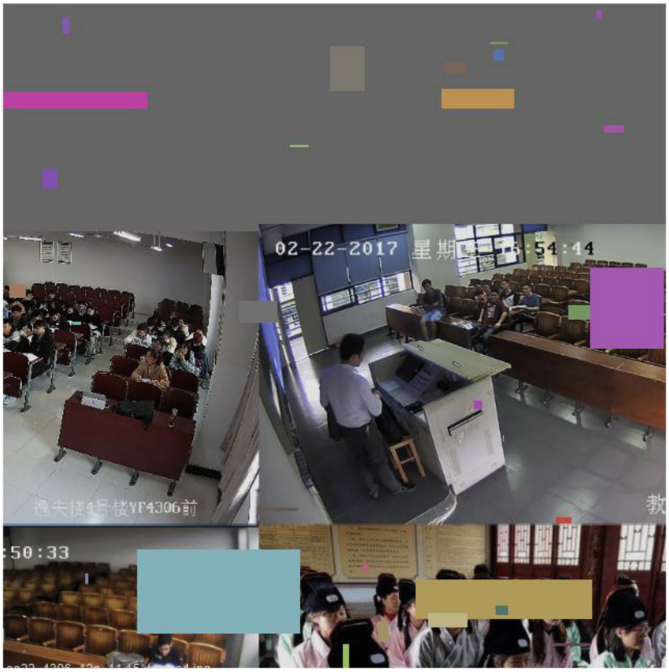

Fig. 4. Mosaic and Cutout. Using the Mosaic data enhancement method, four plots are randomly cropped and put together into a single plot, and the different colored and sized squares in the plot are the effect of Cutout data enhancement.

4.3 Experimental Results and Comparison

The experiments in this paper use precision, recall, regression frame loss, confidence loss and mAP to measure the model detection performance.

Precision: the proportion of correctly detected samples to the total detected samples, the formula is

$$\mathrm{P} = \frac{TP}{TP + FP} \tag{7}$$

Recall: the proportion of samples predicted to be positive to the actual positive samples, the formula is

$$\mathrm{R} = \frac{TP}{TP + FN} \tag{8}$$

mAP: AP is the average value of accuracy under equal interval recall rate, describing the overall situation of accuracy under different recall rates, which is used to react to the global performance of the model. The larger the value of AP, the higher the model accuracy, mAP is the average value of accuracy for each category, there is only one category in this paper, so AP is the same as mAP, and the formula is

$$\mathrm{AP} = \int_0^1 P(R)dR \tag{9}$$

$$\mathrm{mAP} = \frac{\Sigma AP}{n} \tag{10}$$

The regression frame loss, confidence loss formulas are shown in (2) and (3).

In order to find the most suitable size of detection head for the classroom scenario, we did a comparison test using three different combinations of heads. The detection effect of three different scale combinations of detection heads are shown in Table 1.

Table 1. Comparison of the precision and recall of different sizes of detection head.

Size of detection heads	Precision	Recall
40 × 40 and 80 × 80	95.102	91.243
20 × 20 and 40 × 40	93.635	75.147
20 × 20 and 80 × 80	**95.405**	**92.808**

As shown in the table, the best results in terms of precision and recall were achieved using the 20 × 20 and 80 × 80 sized detection heads. From Fig. 5, it can be seen that the 40 × 40 and 80 × 80 detection heads repeat the detection of large scale heads close to the camera, while the 20 × 20 and 40 × 40 detection heads miss many small targets.

Figure 6 shows the variation of regression frame loss and confidence loss for the training and test sets. The bbox loos and confidence loss decreases rapidly in the first 50 rounds, then the training set loss decreases slowly but still into a decreasing trend in the following 150 rounds, but the test set has converged to a straight line with no decreasing trend.

Through the above comparative experiments with different detection head sizes, we have obtained best precision and recall after 200 rounds respectively. The comparison

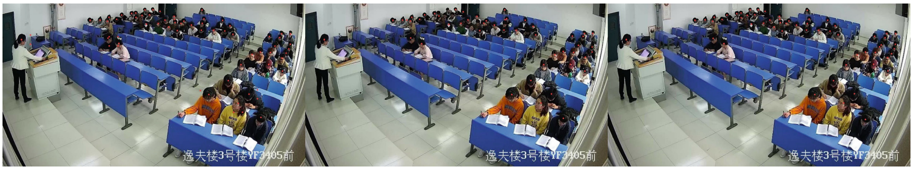

Fig. 5. The prediction effect of different detection head sizes. The left picture is the prediction effect after 200 rounds of model training corresponding to 40 × 40 and 80 × 80 sized detection heads, the middle picture is the prediction effect after 200 rounds of model training corresponding to 20 × 20 and 40 × 40 sized detection heads. The figure on the right is the prediction effect after 200 rounds of model training corresponding to 40 × 40 and 80 × 80 sized detection heads. These three models are identical except for the different detection head sizes.

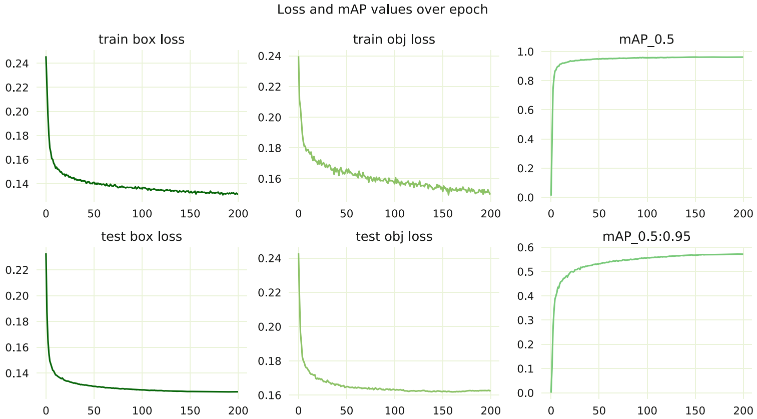

Fig. 6. LOSS and mAP over epoch. The left two plots show the bbox loss for the training and test sets, the center shows the confidence loss for the training and test sets, the top right plot shows the IoU of 0.5 is the mAP value, and the bottom right is the mean value of the mAP when the IoU is from 0.5 to 0.95, with an interval of 0.5.

Table 2. Comparison of precision and recall between our model and YOLOv8 model.

Method	Precision	Recall
YOLOv8	93.429	90.361
Ours	**95.405**	**92.808**

of precision and recall results between our model and YOLOv8 after 200 rounds are shown in Table 2.

Our model has nearly 2% higher precision and 2.5% higher recall than YOLOv8. Since the seat of each person is basically unchanged during a lesson, more C University lesson data can be collected in the future, and for the image frames of the same lesson, we can assist each other's training to further solve the omission detection caused by the occlusion problem.

Through our model, we have successfully extracted the number and spatial coordinates of students at different time intervals within each class. In the subsequent analysis, we aim to calculate the average distance of students from the podium, assess the dispersion of student positions, and investigate potential variations in student engagement across different grades, universitys, and course types. Additionally, we intend to explore the relationship between student motivation and academic performance, leveraging our approach to address the inherent perspective challenges present in the images. Furthermore, we plan to employ tracking algorithms to monitor and analyze the movements of teachers within the classroom, examining their walking routes and the seating arrangements of students in proximity to these routes. This comprehensive analysis will provide valuable insights into classroom dynamics and contribute to a deeper understanding of student-teacher interactions.

5 Conclusion

In order to solve the problem of statistical crowd estimation in the classroom, a model for multi-scale feature fusion based on YOLO algorithm is proposed. Firstly, the backbone network of yolov8 was fine-tuned, and the SEnet module was used to improve the quality of the feature extraction before the special fusion. Then, we replaced the PANnet fusion in the YOLO model by using a bilateral three-path aggregation network. Confidence loss and bounding box regression loss were used to jointly calculate the model loss, and according to the needs of the scene, the anchor and detection head were designed to meet the needs of the classroom scene to detect targets of different scales. Finally, Mosaic and Cutout data enhancement methods were used to increase the generalization ability of the model and the detection ability of occluded targets.

Acknowledgements. The work was supported by science and technology innovation 2030—major project of "New Generation Artificial Intelligence" (2022ZD0115905), the Key Project of Chuzhou University (2022XJZD13) and Anhui University of Science and Technology Graduate Innovation Projects (2023cx2130).

References

1. Yang, X., Zhou, X., Hu, J.: Students' preferences for seating arrangements and their engagement in cooperative learning activities in University English blended learning classrooms in higher education. High. Educ. Res. Dev. **41**(4), 1356–1371 (2022)
2. Juhaňák, L., Cigán, J.: Effects of seating arrangement on students' interaction in group reflective practice. J. Exp. Educ. **91**(2), 249–277 (2023)
3. Sam, D.B., Peri, S.V., Sundararaman, M.N., Kamath, A., Radhakrishnan, V.B.: Locate, size and count: accurately resolving people in dense crowds via detection. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 2739–2751. IEEE (2020)
4. Zhang, Z., Xia, S., Cai, Y., Yang, C., Zeng, S.: A Soft-YoloV4 for high-performance head detection and counting. Mathematics **9**(23), 3096 (2021)

5. Wang, Y., Hou, J., Hou, X.: A self-training approach for point-supervised object detection and counting in crowds. In: Transactions on Image Processing, pp. 2876–2887. IEEE (2020)
6. Zand, M., Damirchi, H., Farley, A., Molahasani, M., Greenspan, M., Etemad, A.: Multiscale crowd counting and localization by multitask point supervision. In: International Conference on Acoustics, Speech and Signal Processing, pp. 1820–1824. IEEE (2022)
7. Lian, D., Chen, X., Li, J., Luo, W., Gao, S.: Locating and counting heads in crowds with a depth prior. In: Transactions on Pattern Analysis and Machine Intelligence, pp. 9056–9072. IEEE (2021)
8. Abousamra, S., Hoai, M., Samaras, D., Chen, C.: Localization in the crowd with topological constraints. ArXiv, abs/2012.12482 (2020)
9. Hu, J., Shen, L., Sun, G.: Squeeze-and-Excitation Networks. In: 2018 Conference on Computer Vision and Pattern Recognition, pp. 7132–7141. IEEE/CVF (2018)
10. Zheng, Z., et al.: Enhancing geometric factors in model learning and inference for object detection and instance segmentation. In: Transactions on Cybernetics, pp. 8574–8586. IEEE (2021)
11. Peng, D., Sun, Z., Chen, Z., Cai, Z., Xie, L., Jin, L.: Detecting heads using feature refine net and cascaded multiscale architecture. In: International Conference on Pattern Recognition (ICPR), pp. 2528–2533. IEEE (2018)
12. Bochkovskiy, A., Wang, C., Liao, H.M.: YOLOv4: optimal speed and accuracy of object detection. ArXiv, abs/2004.10934 (2020)
13. Devries, T., Taylor, G.W.: Improved Regularization of convolutional neural networks with cutout. ArXiv, abs/1708.04552 (2017)