Phát hiện hành vi học sinh trong lớp học thông qua học sâu sử dụng YOLOv8

Võ Vĩnh Thái, Nguyễn Tiến Dũng, Lê Ngọc Hưng
Nhóm 2,Khoa Công Nghệ Thông Tin
Trường Đại học Đại Nam,Việt Nam
ThS.Nguyễn Văn Nhân, ThS. Lê Trung Hiếu
Giảng viên hướng dẫn Khoa Công Nghệ Thông Tin
Trường Đại Học Đại Nam, Việt Nam

***Tóm tắt nội dung*—Nghiên cứu này tập trung vào việc phát triển một mô hình AI hỗ trợ nhận diện hành vi học sinh trong lớp học, nhằm nâng cao năng suất giảng dạy trong môi trường có nhiều học sinh. Chúng tôi sử dụng mô hình YOLOv8 với trọng số được huấn luyện từ trước và tiến hành transfer learning với bộ dữ liệu thu thập từ môi trường học tập tại gia. Mô hình đạt được độ chính xác cao trên cả tập huấn luyện và kiểm tra, đặc biệt trong việc nhận diện các hành vi như "Giơ tay", "Đọc sách", "Sử dụng điện thoại", "Sử dụng máy tính xách tay","Nằm ngủ hay ngủ", "Viết bài", "Điện thoại","Nâng đầu""Nhìn bảng", "Quay đầu","Ngẩng thẳng lên"và "Cúi".**

***Từ khóa*—YOLOv8, học sâu, nhận diện hành vi học sinh, transfer learning, SCB-Dataset, giám sát lớp học.**

I. Giới thiệu

Việc theo dõi hành vi học sinh trong lớp học là một yếu tố quan trọng trong quản lý giáo dục, giúp đánh giá mức độ tham gia và cải thiện chất lượng giảng dạy. Tuy nhiên, các phương pháp truyền thống như quan sát trực tiếp thường tiêu tốn nhiều thời gian và dễ bị ảnh hưởng bởi yếu tố chủ quan. Điều này đặt ra nhu cầu phát triển một hệ thống tự động nhằm hỗ trợ quá trình giám sát lớp học.

Với sự phát triển mạnh mẽ của trí tuệ nhân tạo (AI) và thị giác máy tính, các mô hình học sâu đã chứng minh hiệu quả cao trong nhận diện đối tượng, đặc biệt là YOLO (You Only Look Once) với khả năng nhận diện theo thời gian thực. Trong nghiên cứu này, chúng tôi đề xuất ứng dụng YOLOv8 để nhận diện một số hành vi quan trọng của học sinh như "Giơ tay", "Đọc sách", "Sử dụng điện thoại", "Sử dụng máy tính xách tay","Nằm ngủ hay ngủ", "Viết bài", "Điện thoại","Nâng đầu""Nhìn bảng", "Quay đầu","Ngẩng thẳng lên"và "Cúi".

II. Các Công trình Nghiên cứu Liên quan

Trong những năm gần đây, nhiều nhà nghiên cứu đã áp dụng công nghệ thị giác máy tính để tự động phát hiện hành vi học sinh trong lớp học, nhưng sự thiếu hụt các bộ dữ liệu hành vi học sinh công khai trong lĩnh vực giáo dục đã hạn chế nghiêm trọng việc ứng dụng phát hiện hành vi video trong lĩnh vực này. Nhiều nhà nghiên cứu cũng đã đề xuất các bộ dữ liệu chưa được công bố.

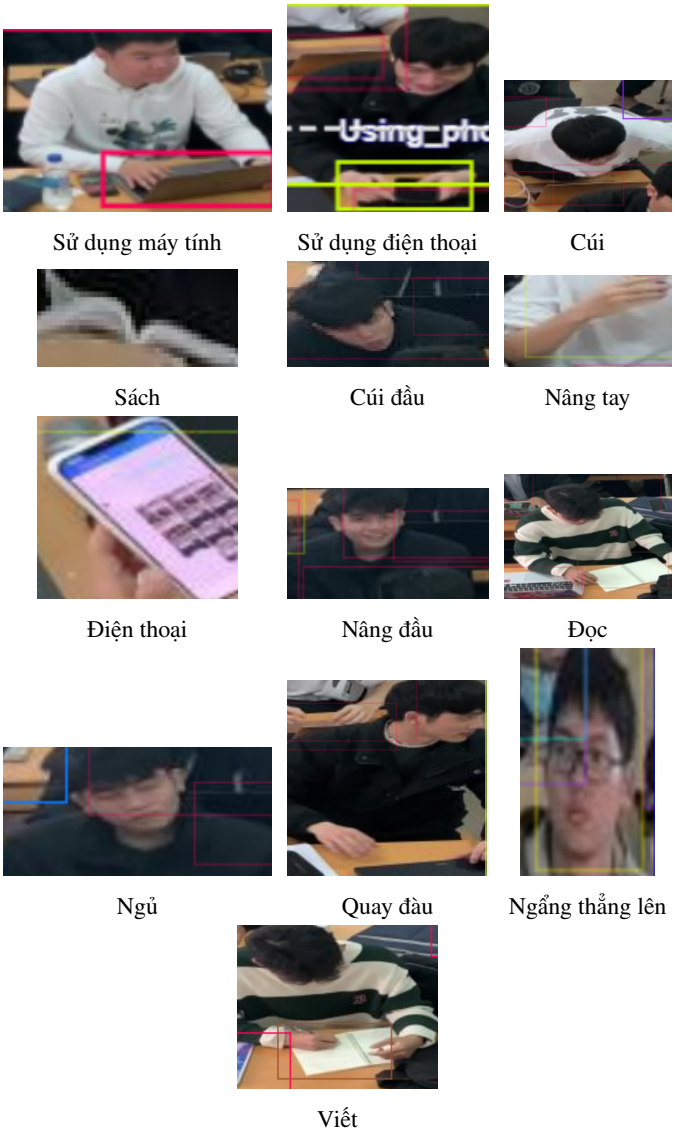
- **ActRec-Classroom** [1] là một tập dữ liệu gồm 5 lớp hành vi: nghe, mệt mỏi, giơ tay, dựa vào, và đọc/viết với 5.126 hình ảnh. Phương pháp này sử dụng Faster R-CNN để phát hiện cơ thể người, sau đó dùng OpenPose để trích xuất các điểm chính của xương, khuôn mặt và ngón tay. Cuối cùng, một bộ phân loại dựa trên CNN được phát triển để nhận diện hành động.
- Một bộ dữ liệu quy mô lớn về hành vi học sinh được thu thập từ ba mươi trường học, với các nhãn hành vi học sinh được gán bằng hộp giới hạn trong từng khung hình. Bộ dữ liệu này bao gồm 70.000 mẫu giơ tay, 20.000 mẫu đứng và 3.000 mẫu ngủ. Mô hình Faster R-CNN cải tiến đã được áp dụng để phân tích hành vi học sinh, với một đầu phát hiện nhận thức về quy mô, chiến lược kết hợp đặc trưng, và sử dụng Online Hard Example Mining (OHEM) để cải thiện độ chính xác trong bối cảnh dữ liệu mất cân đối [2].
- **BNU-LCSAD** [3] là một bộ dữ liệu toàn diện có thể được sử dụng để nhận dạng, phát hiện và chú thích hành vi học sinh trong lớp học. Bộ dữ liệu này bao gồm 128 video từ các môn học khác nhau trong 11 lớp học. Các mô hình cơ sở được sử dụng cho các tác vụ khác nhau bao gồm mạng hai dòng cho nhận dạng [4], ACAM và MOC cho phát hiện hành động [5], BSN [6] và DBG [7] cho phát hiện theo thời gian, cũng như RecNet và HACA cho chú thích video [8], [9].
- **Bộ dữ liệu hành vi học sinh trong lớp học** [10] là một bộ sưu tập bao gồm 400 học sinh từ 90 video lớp học ở trường tiểu học. Bộ dữ liệu này bao gồm các hình ảnh một người, tổng cộng 10.000 hình ảnh của học sinh tham gia các hành vi như giơ tay, đi qua lại, viết trên bảng, và nhìn lên và xuống, cũng như 1.000 hình ảnh bổ sung ghi lại học sinh cúi xuống, đứng và nằm trên bàn. Để nhận diện các hành vi lớp học này, phương pháp đề xuất sử dụng mạng nơ-ron tích chập sâu (CNN-10) để trích xuất thông tin chính từ dữ liệu xương người, giúp loại bỏ các thông tin không liên quan, và đạt được độ chính xác nhận diện cao hơn.
- **Bộ dữ liệu hành vi học sinh** [11] là một bộ dữ liệu lớp học thông minh phân loại hành vi học sinh thành bảy lớp. Các video giám sát lớp học đầy thử thách đã được chọn lọc và chú thích cẩn thận để tạo ra bộ dữ liệu này. Phương pháp đề xuất trong bối cảnh này tích hợp các đặc trưng quan hệ để phân tích cách các diễn viên tương tác với bối cảnh xung quanh. Nó mô hình hóa mối quan hệ người với người bằng các bộ phận cơ thể và bối cảnh, sau đó kết hợp những đặc trưng quan hệ này với các đặc trưng hình ảnh để đạt được

sự nhận diện chính xác mối quan hệ người với người.

- **Bộ dữ liệu hành động học sinh** [12] là một bộ sưu tập gồm 3.881 hình ảnh đã được gắn nhãn miêu tả một loạt hành vi học sinh trong lớp học, như giơ tay, chú ý, ăn uống, bị phân tâm, đọc sách, sử dụng điện thoại, viết, cảm thấy chán, và cười. Để huấn luyện và đánh giá các hành vi này, mô hình phát hiện đối tượng YOLOv5 đã được sử dụng.

III. Mô tả và Thu thập Dataset

Trong những năm gần đây, nhiều nhà nghiên cứu đã áp dụng công nghệ thị giác máy tính để tự động phát hiện hành vi trong lớp học của học sinh, nhưng việc thiếu bộ dữ liệu hành vi học sinh mở trong lĩnh vực giáo dục đã hạn chế nghiêm trọng việc áp dụng công nghệ phát hiện hành vi video trong lĩnh vực này. Nhiều nhà nghiên cứu cũng đã đề xuất nhiều bộ dữ liệu chưa được công bố.



Nghiên cứu này sử dụng hai bộ dữ liệu để huấn luyện và đánh giá mô hình phát hiện hành vi học sinh trong lớp học.

A. Dataset: SCB-Dataset

Bộ dữ liệu thứ nhất dựa trên SCB-Dataset, được đề cập trong [13], bao gồm 20 lớp hành vi khác nhau. Trong nghiên cứu này, chúng tôi chỉ sử dụng một phần của SCB-Dataset với 3 lớp hành vi chính: "Giơ tay", "Đọc sách", "Sử dụng điện thoại", "Sử dụng máy tính xách tay","Nằm ngủ hay ngủ", "Viết bài", "Điện thoại","Nâng đầu""Nhìn bảng", "Quay đầu","Ngẩng thẳng lên"và "Cúi". Tổng cộng, bộ dữ liệu này bao gồm 3646 ảnh, được chia thành các tập như sau:

- **Tập huấn luyện (Train)**: 2978 ảnh, chiếm khoảng 81.6% tổng số ảnh.
- **Tập xác nhận (Validation)**: 404 ảnh, chiếm khoảng 11.1% tổng số ảnh.
- **Tập kiểm tra (Test)**: 264 ảnh, chiếm khoảng 7.3% tổng số ảnh.

B. Phân tích và Quy trình Thu thập

Dataset thứ hai được thu thập trong điều kiện ánh sáng và góc quay khác biệt so với SCB-Dataset, nhằm kiểm tra khả năng thích nghi của mô hình trong các môi trường thực tế. Việc gán nhãn được thực hiện cẩn thận để tránh rủi ro trong quá trình transfer learning, đặc biệt khi áp dụng mô hình đã huấn luyện từ SCB-Dataset sang dataset tại gia.

IV. Phương pháp Đề xuất

A. Giới thiệu

Trong nghiên cứu này, chúng tôi đề xuất sử dụng mô hình YOLOv8 với hai phiên bản trọng số **YOLOv8m** và **YOLOv8l** để phát hiện hành vi học sinh trong lớp học. Mô hình được huấn luyện trên bộ dữ liệu SCB-Dataset kết hợp với tập dữ liệu thu thập thực tế, giúp cải thiện khả năng nhận diện trong môi trường lớp học thực tế.

Việc huấn luyện mô hình được thực hiện trên nền tảng **Google Colab** với GPU **A100**, đảm bảo tốc độ xử lý nhanh và tối ưu hóa hiệu suất. Chúng tôi sử dụng phương pháp **transfer learning** từ mô hình YOLOv8 đã được huấn luyện trước trên tập dữ liệu COCO, sau đó tinh chỉnh lại trên tập dữ liệu chuyên biệt của lớp học.

B. Kiến trúc YOLOv8

YOLOv8 là một trong những phiên bản tiên tiến nhất của dòng YOLO, có khả năng nhận diện đối tượng với tốc độ cao và độ chính xác vượt trội. Cấu trúc chính của mô hình bao gồm:

- **Backbone**: CSP-DarkNet kết hợp với **C2f module** giúp trích xuất đặc trưng hiệu quả.
- **Neck**: Sử dụng **PAFPN (Path Aggregation Feature Pyramid Network)** giúp tăng cường đặc trưng.
- **Head**: Cơ chế **Anchor-Free Detection Head** giúp dự đoán bounding box một cách linh hoạt hơn.

Chúng tôi sử dụng hai phiên bản **YOLOv8m** và **YOLOv8l** để so sánh hiệu suất:

- **YOLOv8m**: Cân bằng giữa tốc độ và độ chính xác, phù hợp với các ứng dụng thời gian thực.
- **YOLOv8l**: Có nhiều tham số hơn, giúp mô hình học được nhiều đặc trưng hơn nhưng yêu cầu tài nguyên lớn hơn.

C. Huấn luyện mô hình trên Google Colab với GPU A100

1) Cài đặt môi trường: Trước tiên, chúng tôi tiến hành cài đặt Ultralytics YOLOv8 trên Google Colab: [language=bash] !pip install ultralytics from ultralytics import YOLO

2) Chuẩn bị dữ liệu: Dữ liệu được chia thành ba tập: Train (80%), Validation (10%) và Test (10%).

3) Huấn luyện mô hình: Chúng tôi huấn luyện hai mô hình YOLOv8m và YOLOv8l với 100 epochs, batch size 16.

[language=bash] !pip install ultralytics from ultralytics import YOLO

D. Chuẩn bị dữ liệu

Dữ liệu được chuẩn bị và chia thành ba tập chính: tập huấn luyện (80%), tập kiểm tra (10%) và tập xác nhận (10%). Mỗi tập dữ liệu này được sử dụng để huấn luyện mô hình, kiểm tra độ chính xác của mô hình trong quá trình huấn luyện và đánh giá hiệu quả của mô hình sau khi huấn luyện.

E. Huấn luyện mô hình

Sau khi cài đặt và chuẩn bị dữ liệu, chúng tôi tiến hành huấn luyện hai mô hình YOLOv8m và YOLOv8l với số lượng epochs là 100, batch size là 16 và kích thước ảnh đầu vào là 640x640. Các mô hình này được huấn luyện sử dụng các tập dữ liệu đã được chuẩn bị. Cụ thể, các bước huấn luyện được thực hiện theo các đoạn mã Python sau:

- **Huấn luyện YOLOv8m:** [language=Python] model = YOLO("yolov8m.pt") model.train(data="data.yaml", epochs=100, batch=16, imgsz=640)
- **Huấn luyện YOLOv8l:** [language=Python] model = YOLO("yolov8l.pt") model.train(data="data.yaml", epochs=100, batch=16, imgsz=640)

F. Tăng tốc độ huấn luyện với GPU A100

Để đảm bảo quá trình huấn luyện diễn ra nhanh chóng và hiệu quả, chúng tôi sử dụng GPU A100 trên Google Colab. GPU A100 có khả năng xử lý dữ liệu cực kỳ nhanh chóng và mạnh mẽ, giúp tăng tốc quá trình huấn luyện so với các GPU thông thường. Việc sử dụng GPU A100 cho phép giảm thiểu thời gian huấn luyện, đồng thời tối ưu hóa việc tìm kiếm tham số và cải thiện độ chính xác của mô hình.

G. Đánh giá và kiểm tra mô hình

Sau khi huấn luyện, mô hình sẽ được đánh giá và kiểm tra trên các tập dữ liệu test. Các chỉ số đánh giá như độ chính xác, độ nhạy, và F1-score sẽ được sử dụng để đo lường hiệu quả của mô hình. Mô hình đạt kết quả cao sẽ được triển khai cho các tác vụ nhận diện đối tượng thực tế.

V. Kết quả thí nghiệm

Mô hình đầu tiên được huấn luyện dựa trên YOLOv8 với bộ dữ liệu SCB-Dataset, sử dụng 50 epoch và batch size 16. Kết quả trên tập huấn luyện và kiểm tra đạt được độ chính xác cao như trong bảng sau:

Sau khi huấn luyện và đánh giá các mô hình, chúng tôi có thể đưa ra các kết luận sau về hiệu suất của mô hình YOLOv8:

Hình 1. Kết quả huấn luyện mô hình YOLOv8m và YOLOv8l

Sau khi huấn luyện và đánh giá các mô hình, chúng tôi có thể đưa ra các kết luận sau về hiệu suất của mô hình YOLOv8:

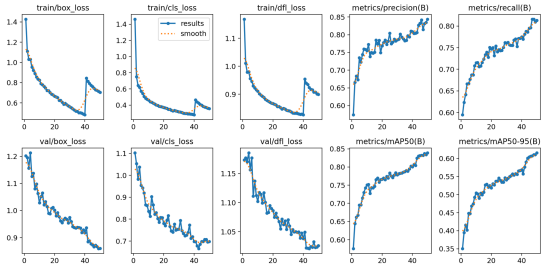
- **Accuracy:** Mô hình đạt độ chính xác cao trong việc phát hiện đối tượng, với mức độ chính xác đáng kể trên tập kiểm thử.
- **mAP50:** Mô hình đạt điểm mAP50 là 81%, cho thấy khả năng nhận diện đối tượng của mô hình ở mức độ khá tốt.
- **mAP50-95:** Mô hình đạt điểm mAP50-95 là 58%, cho thấy khả năng phát hiện đối tượng trên các độ khó khác nhau.
- **Thời gian suy luận (Inference Time):** Mô hình có thời gian suy luận trung bình là 2.4ms, cho thấy khả năng xử lý nhanh chóng và hiệu quả trong các ứng dụng thời gian thực.

```
=== Thống kê kết quả dự đoán ===
Tổng số đối tượng phát hiện: 9777

Số lượng đối tượng theo từng lớp:
- upright: 2717 đối tượng (Confidence trung bình: 0.817)
- raise_head: 820 đối tượng (Confidence trung bình: 0.772)
- bow_head: 2325 đối tượng (Confidence trung bình: 0.740)
- book: 1122 đối tượng (Confidence trung bình: 0.706)
- Using_phone: 703 đối tượng (Confidence trung bình: 0.628)
- turn_head: 150 đối tượng (Confidence trung bình: 0.675)
- bend: 701 đối tượng (Confidence trung bình: 0.756)
- reading: 589 đối tượng (Confidence trung bình: 0.600)
- sleep: 75 đối tượng (Confidence trung bình: 0.817)
- phone: 264 đối tượng (Confidence trung bình: 0.650)
- hand-raising: 142 đối tượng (Confidence trung bình: 0.588)
- writing: 169 đối tượng (Confidence trung bình: 0.555)
```

Hình 2. Thống kê kết quả dự đoán

Những kết quả này cho thấy mô hình YOLOv8 có khả năng ứng dụng hiệu quả trong các bài toán phát hiện đối tượng với yêu cầu về tốc độ và độ chính xác cao.



Hình 3. Thống kê kết quả dự đoán

Những kết quả này cho thấy mô hình YOLOv8 có khả năng ứng dụng hiệu quả trong các bài toán phát hiện đối tượng với yêu cầu về tốc độ và độ chính xác cao.

VI. Thảo luận

Kết quả thí nghiệm cho thấy mô hình YOLOv8 có thể nhận diện hành vi học sinh một cách chính xác. Tuy nhiên, vẫn còn một số thử thách như sự thay đổi kích thước hình ảnh giữa các hàng ghế và sự che khuất giữa các học sinh. Cần phải cải thiện mô hình để xử lý tốt hơn trong các môi trường lớp học đông đúc.

VII. Kết luận

Nghiên cứu này đã phát triển thành công một mô hình nhận diện hành vi học sinh trong lớp học dựa trên YOLOv8, kết hợp với transfer learning và finetuning trên bộ dữ liệu SCB-Dataset và dataset tại gia. Kết quả thí nghiệm cho thấy mô hình đạt được độ chính xác cao, với mAP@0.5 đạt 91.6% trên SCB-Dataset và 79.5% trên dataset tại gia, đặc biệt hiệu quả trong việc nhận diện các hành vi như "Giơ tay", "Đọc sách", "Sử dụng điện thoại", "Sử dụng máy tính xách tay","Nằm ngủ hay ngủ", "Viết bài", "Điện thoại","Nâng đầu""Nhìn bảng", "Quay đầu","Ngẩng thẳng lên"và "Cúi". Việc sử dụng YOLOv8n, với thời gian chạy thấp và kích thước mô hình nhẹ, cùng chiến lược đóng băng 50% số tầng trong quá trình finetune, đã giúp mô hình thích nghi tốt với các đặc trưng đặc thù của hành vi học sinh mà không làm mất đi khả năng nhận diện tổng quát từ bộ dữ liệu COCO.

Đóng góp chính của nghiên cứu là cung cấp một giải pháp hiệu quả để hỗ trợ giáo viên trong việc quản lý lớp học, từ đó nâng cao hiệu quả giảng dạy trong môi trường đông học sinh. Tuy nhiên, mô hình vẫn còn gặp một số thách thức, như sự che khuất giữa các học sinh và sự thay đổi kích thước hình ảnh giữa các hàng ghế. Trong tương lai, chúng tôi sẽ tiếp tục cải thiện mô hình bằng cách tăng cường bộ dữ liệu với các mẫu đa dạng hơn, đồng thời thử nghiệm các kỹ thuật tăng cường dữ liệu và tối ưu hóa mô hình để triển khai trên các thiết bị có tài nguyên hạn chế trong môi trường thực tế.

Tài liệu

[1] F. Yang, T. Wang, and X. Wang, "Student Classroom Behavior Detection Based on YOLOv7+ BRA and Multi-model Fusion,"in *Proc. Int. Conf. Image Graph.*, Cham: Springer Nature Switzerland, 2023, pp. 41-52.

[2] Y. Huang, M. Liang, and X. Wang, "Multi-person classroom action recognition in classroom teaching videos based on deep spatiotemporal residual convolution neural network,"*J. Comput. Appl.*, vol. 42, no. 3, p. 736, 2022.

[3] X. He, F. Yang, and Z. Chen, "The recognition of student classroom behavior based on human skeleton and deep learning,"*Mod. Educ. Technol.*, vol. 30, no. 11, pp. 105-112, 2020.

[4] X.-Y. Yan, Y.-X. Kuang, G.-R. Bai, and Y. Li, "Student classroom behavior recognition method based on deep learning,"*Comput. Eng.*, doi: 10.19678/j.issn.1000-3428.0065369.

[5] C. Gu, C. Sun, and D. A. Ross, "Ava: A video dataset of spatio-temporally localized atomic visual actions,"in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6047-6056.

[6] C. Feichtenhofer, H. Fan, and J. Malik, "Slowfast networks for video recognition,"in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6202-6211.

[7] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild,"*arXiv preprint*, arXiv:1212.0402, 2012.

[8] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset,"in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6299-6308.

[9] 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,"*arXiv preprint*, arXiv:2207.02696, 2022.

[10] S. Ren, K. He, and R. Girshick, "Faster R-CNN: Towards real-time object detection with region proposal networks,"in *Adv. Neural Inf. Process. Syst.*, vol. 28, 2015.

[11] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement,"*arXiv preprint*, arXiv:1804.02767, 2018.

[12] T. Y. Lin, M. Maire, and S. Belongie, "Microsoft coco: Common objects in context,"in *Proc. ECCV*, 2014, pp. 740-755.

[13] R. Fu, T. Wu, and Z. Luo, "Learning behavior analysis in classroom based on deep learning,"in *Proc. ICICIP*, IEEE, 2019, pp. 206-212.