COMP4801 Final Year Project

Interim Report

AI-Powered Attention Monitoring System for Enhancing Online Learning Engagement

Tam Cheung Yan Shojin; 3035927673; BENG(CompSc)

Supervisor: Prof. Zhao Hengshuang

26-1-2025

Abstracts

Online classes have gained popularity since the pandemic, offering greater flexibility and accessibility [1]. However, research indicates that students are more susceptible to distraction during these sessions [2]. This project proposes an application that utilizes computer vision AI models to monitor student attentiveness through camera input. We are currently on schedule, having trained two models using various datasets, achieving satisfactory accuracy. The next steps involve sourcing additional datasets and developing a prototype for the application.

Acknowledgement

I would like to express my sincere gratitude to the Department of Computer Science for providing me with the opportunity to undertake this project. My heartfelt thanks go to Prof. Zhao HengShuang for his invaluable advice and guidance throughout the process.

Additionally, I wish to acknowledge Miss Mable Choi for helping me write this report.

Contents

List of Figures

List of Table

Abbreviations & Acronyms

AI - Artificial Intelligence

EAR - Eye aspect ratio

FPS - Frames Per Second

GPU - Graphics Processing Unit

MSCOCO - Microsoft Common Objects in Context

ONNX - Open Neural Network Exchange

UI - User Interface

YAR - Yawn aspect ratio

Notations & Symbols

S_{Happy} - The percentage of happy emotion predicted by the ResEmoteNet model

x - Attentiveness

Σ $parameters$ – Sum of all parameters

n – Number of all parameters

1. Solving inattentiveness with AI

Online classes have become essential since the pandemic, enabling students to participate remotely. Educational institutions have increasingly adopted online learning platforms, which allow for greater flexibility and accessibility [1]. However, research indicates that students are more susceptible to distractions during these sessions, adversely affecting their engagement and learning outcomes [2]. The lack of physical presence can lead to a decline in attentiveness [1]. Various studies suggest that students may multitask, engage with their phones [3], or even fall asleep during lectures [4]. This project proposes an innovative application that leverages advanced computer vision AI models to monitor and analyse student attentiveness through camera input, aiming to enhance online learning experiences by providing real-time feedback on student engagement.

1.1 Report outline

The report is structured as follows: Chapter 2 discusses the influential works that inspired this project, while Chapter 3 outlines its objectives. Chapter 4 covers the models utilised, detailing their training processes and the implementation of the program. Chapter 5 reviews the current progress, highlights the challenges encountered, and outlines future plans. Finally, Chapter 6 concludes the report with final insights and reflections.

2. Related works

Previous studies have explored various methodologies for monitoring student attentiveness, including:

Head Pose and Facial Landmark Tracking: Analyzing head orientation and facial features to gauge focus. For example, techniques that track head movement can indicate whether a student is looking at the screen [5].

Phone Detection: Identifying the presence of mobile devices to ascertain distractions. Research has shown that phone use during classes correlates with lower academic performance [6].

Eye Tracking: Monitoring eye movements to determine attention levels. Eye tracking can provide insights into where a student’s focus lies during a lecture [7].

Student's action detection: Classifying students' behaviours as high attention and low attention behaviour helps determine their attentiveness [8].

These approaches highlight the potential for integrating multiple data sources to create a comprehensive attentiveness monitoring system. Inspired by these studies, the proposed application will leverage open-source AI models to perform the four aforementioned tasks for assessing students' attentiveness.

3. Objectives

The proposed application will be built on the OpenCV framework for real-time computer vision. It aims to integrate multiple AI models:

1. L2CS-Net [9]: A gaze detection model to assess where the student is looking.
2. Facial Landmark Detection HRNet [10]: For detecting head orientation and facial landmarks.
3. ResEmoteNet [11]: To analyse the student's emotional state.
4. YOLOv8 [12]: For real-time detection of phones and monitoring of student actions.

These models will be trained on diverse datasets to enhance their predictive capabilities. The outputs will be synthesized to calculate a comprehensive attentiveness score for each student. A user-friendly interface will display these results alongside insights derived from the model predictions.

4. Methodology

The following section covers the AI models utilised in the project, the calculation of attentiveness, and the implementation of the application. Chapters 4.1, 4.2, 4.3, and 4.4 present the outputs and attentiveness calculations for L2CS-Net, the Facial Landmark Detection HRNet model, ResEmoteNet, and YOLOv8, respectively. Chapter 4.5 provides an overview of the overall attentiveness calculation. Chapter 4.6 details the datasets used for training. Finally, Chapters 4.7, 4.8, and 4.9 discuss the implementation of training, deployment, and the user interface code.

4.1 L2CS-Net

L2CS-Net predicts the gaze direction of a student [9]. Attentiveness will be assumed if the student is looking forward; while inattentiveness will be assumed if the student is looking sideways, upward, downward, or backward.

4.2 Facial Landmark Detection HRNet

Facial Landmark Detection HRNet detects head pose, and facial landmarks [10]. For the facial landmarks, the model predicts 68 facial landmarks as proposed by the 300W database [13], which is shown in figure 1. Eye aspect ratio (EAR) and yawn aspect ratio (YAR) will be calculated, which are indicators of the drowsiness of the person, and will be used to calculate attentiveness [5].

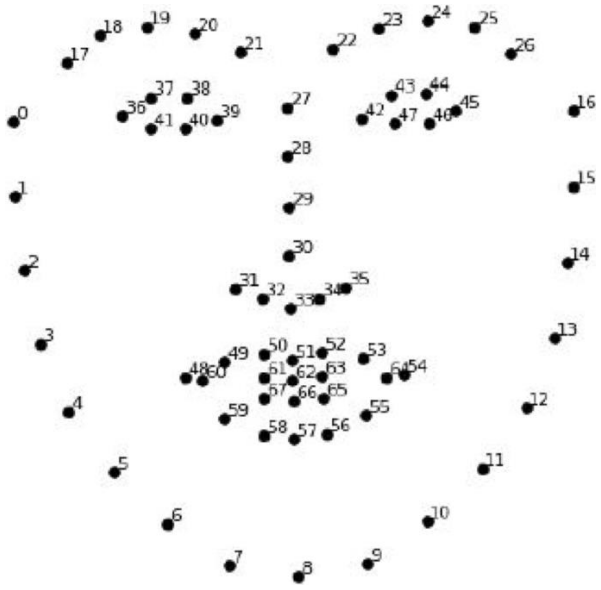

Figure 1: 68 Facial Landmarks [14]

The calculation of EAR is as follows [5]:

$$EAR = \frac{EAR1 + EAR2}{2} \quad (1)$$

$$EAR1 = \frac{||37 - 41|| + ||38 - 40||}{2||36 - 39||}$$

(2)

$$EAR2 = \frac{\left|\left|43-47\right|\right| + \left|\left|44-46\right|\right|}{2\left|\left|42-45\right|\right|}$$

(3)

The calculation of YAR is as follows [5]:

$$YAR = \frac{\left|\left|61-67\right|\right| + \left|\left|62-66\right|\right| + \left|\left|63-65\right|\right|}{2\left|\left|64-60\right|\right|}$$

(4)

EAR1 and EAR2 are the eye aspect ratio of left and right eyes respectively. For each landmark, a value will be calculated in (x-y) format, using the coordinates of the landmark [15]. The numbers in the formulas such as 37 and 41 refer to the values correlated to the 37th and 41st landmarks respectively.

As for the head pose, similarly, for gaze direction, attentiveness will be determined by the direction of the head pose.

4.3 ResEmoteNet

ResEmoteNet predicts the emotions of the student [11]. The model categorises emotions as one of the seven classes: angry, disgusted, fearful, happy, neutral, sad, and surprised. Attentiveness will be calculated based on the affective model showed in Figure 1 [16].

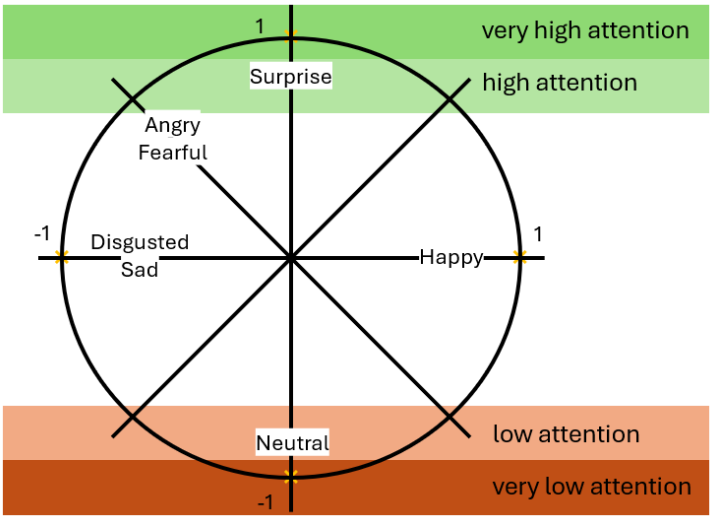

Figure 2: Affective Model modified from [16] to match with classes of ResEmoteNet

We calculate the attentiveness as follows [17]:

$$S_1 = S_{Happy} + (-1) \cdot S_{Sad} \quad (5)$$

$$S_2 = S_{Surprise} + (-1) \cdot S_{Neutral} \quad (6)$$

$$S_3 = S_{Disgusted} + S_{Fearful} \quad (7)$$

Where $S = \{S_{Happy}, S_{Sad}, S_{surprise}, S_{Neutral}, S_{Disgusted}, S_{Fearful}\}$ are the percentages of each of the emotions predicted by the model. The algorithm then calculates attentiveness x:

$$x = \sqrt{\left(\overline{S_1} + \frac{\overline{S_3}}{\sqrt{2}}\right)^2 + \left(\overline{S_2} + \frac{\overline{S_3}}{\sqrt{2}}\right)^2} \cdot \frac{\overline{S_2} + \frac{\overline{S_3}}{\sqrt{2}}}{\left|\overline{S_2} + \frac{\overline{S_3}}{\sqrt{2}}\right|} + 1 \quad (8)$$

Here, $\sqrt{\left(\overline{S_1} + \frac{\overline{S_3}}{\sqrt{2}}\right)^2 + \left(\overline{S_2} + \frac{\overline{S_3}}{\sqrt{2}}\right)^2}$ indicates the overall emotional intensity, while $\frac{\overline{S_2} + \frac{\overline{S_3}}{\sqrt{2}}}{\left|\overline{S_2} + \frac{\overline{S_3}}{\sqrt{2}}\right|}$ serves as the dimensional factor to assess whether a student is focused or distracted. The offset 1 is added so the range of the attentiveness becomes [-1,1].

4.4 Yolov8

Yolov8 is a detection model suitable for real-time detection [12]. It is selected due to its performance. The base model will be used to train two detection models: one for detecting phones, and the other for detecting student actions.

The phone detection model will be trained mainly with MSCOCO [18]. The student action detection model will be trained with EduNet dataset [19], pending permission from the dataset owner.

4.5 Overall attentiveness calculation

The overall attentiveness is the mean of all the parameters calculated with the above models, as suggested in [7].

$$Attentiveness = \frac{\Sigma\, parameters}{n} * 100 \quad (9)$$

4.6 Dataset preparation

Most AI models will be trained using the datasets that were used to develop the pretrained models provided by their authors. However, the student action detection model will be trained using the EduNet dataset [19]. Details of the datasets are presented in Table 4.1 and 4.2.

Table 4.1: Datasets used for training Facial Landmark Detection HRNet [20]

Name	Author	Published	#Marks	#Samples
300-W [13]	Imperial College London	2013	68	600
300-VW [21]	Imperial College London	2015	68	218597
AFW [22]	Imperial College London	2013	68	337
AFLW2000-3D [23]	Chinese Academy of Sciences	2015	68	2000
HELEN [24]	Imperial College London	2013	68	2330
IBUG [25]	Imperial College London	2013	68	135
LFPW [26]	Imperial College London	2013	68	1035
Total				**225034**

Table 4.2: Datasets used for training L2CS-Net, ResEmoteNet, and YOLOv8

AI model	Task	Dataset	#Classes	Dataset size
L2CS-Net	Gaze detection	MPIIGaze [27]	1	15 subjects, 213659 images
L2CS-Net	Gaze detection	Gaze360 [28]	1	238 subjects, 172000 images

ResEmoteNet	Emotion detection	FER2013 [29]	7	34034 images
YOLOv8	Student's action detection	EduNet [19]	20	7851 clips, 12 hours of clips
YOLOv8	Phone detection	MSCOCO [18]	80	330000 images, 11000 instances of cell phone

4.7 Models training

"All models will be trained from scratch using the prepared datasets, except for the phone detection model. Due to the size of the MSCOCO dataset, we will utilize a pretrained model for this task instead.

4.8 Program coding

The training and deployment code will be implemented using Python within the OpenCV framework, which is known for its efficiency and flexibility in handling computer vision tasks. Python will be used for ease of programming.

4.9 UI

The UI will be developed using Qt, providing a user-friendly experience. It will display model predictions, overall attentiveness scores, and visual feedback on student engagement. If inattentiveness is detected, the UI will offer actionable tips and strategies for improving focus, such as reminders to take breaks or suggestions for reducing distractions in the environment.

4.10 Summary

This chapter outlines the rationale behind our project. We introduced the following AI models: L2CS-Net, the Facial Landmark Detection HRNet model, the ResEmoteNet, and YOLOv8. We also defined the methods for calculating attentiveness, dataset preparation, and the framework that will be used for the program. The next chapter will discuss our current progress.

5. Current Progress

5.1 Overview

This chapter outlines the current progress of the project. Chapter 5.2 provides details on model training. Chapter 5.3 presents the results of the training, while Chapter 5.4 showcases the outcomes of real-time inference. Chapter 5.5 includes the project schedule. Finally, Chapters 5.6 and 5.7 discuss the challenges we may face and the next steps for the project.

5.2 Models training

Currently, L2CS-Net, ResEmoteNet are trained, while Facial Landmark Detection HRNet is being trained. All three models were trained using the official training code provided in their respective GitHub repositories. Table 5.1 shows the configuration of the training.

Table 5.1: Configuration of the training

AI model	Datasets	Batch size	#Epochs	Learning rate
L2CS-Net	MPIIGaze	16	50	0.00001
ResEmoteNet	FER2013	16	80	0.001
Facial Landmark Detection HRNet	300-W, 300-VW, AFW, AFLW2000-3D, HELEN, IBUG, LFPW	32	80	0.001

5.3 Training results

Testing for L2CS-Net and ResEmoteNet was conducted using the testing sets of the datasets. The training results of the models are shown in Table 5.2.

Table 5.2: Training results of L2CS-Net and ResEmoteNet

AI model	Name of Metric	Value
L2CS-Net	Gaze Angular Error	2.307
ResEmoteNet	Accuracy	0.900

While the accuracy is satisfactory, it indicates that incorporating additional datasets could help achieve higher performance.

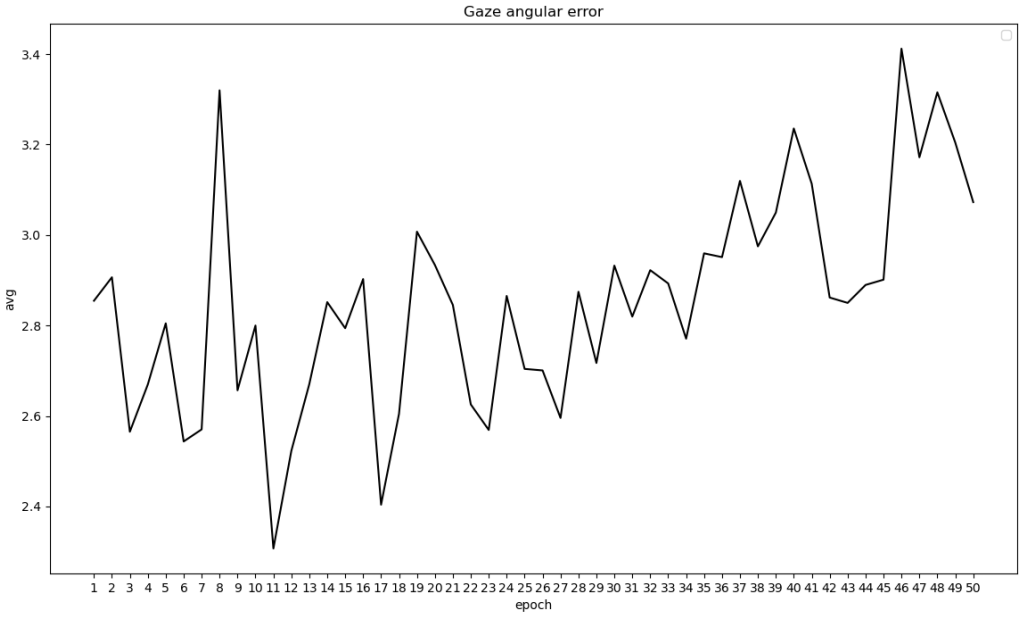

Figure 3: Learning curve of L2CS-Net

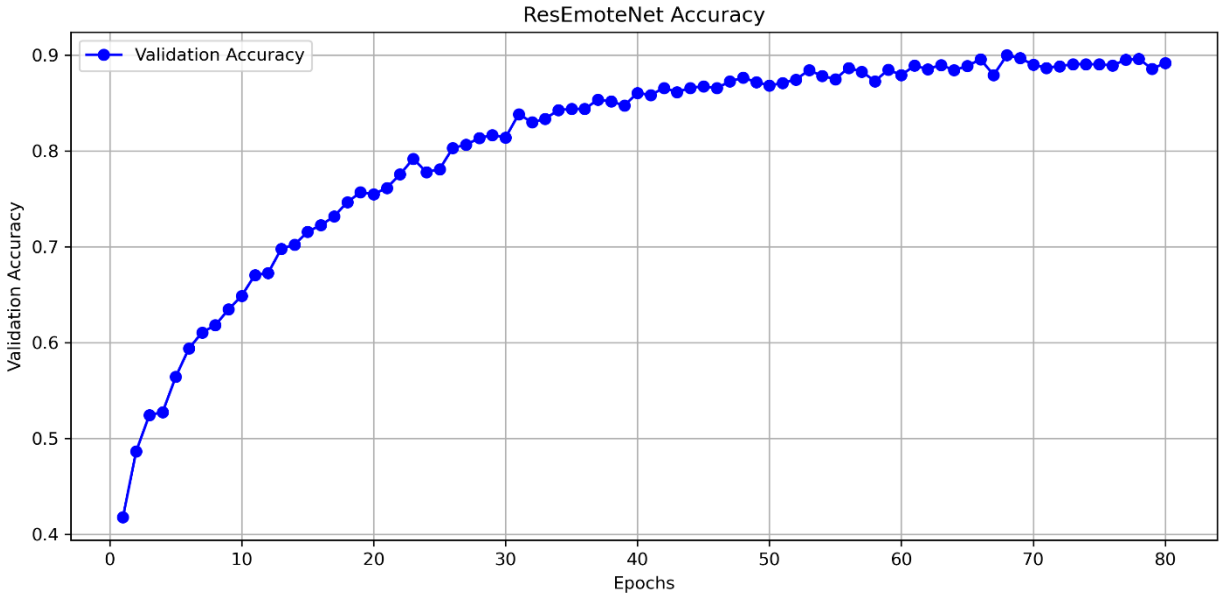

Figure 4: Learning curve of ResEmoteNet

For L2CS-Net, epoch 11 yielded the best results, as illustrated in Figure 3. Throughout training, the error value exhibited fluctuations, notably rising after epoch 11, indicating the onset of overfitting. To address this issue, adjustments such as tuning the learning rate or increasing the dataset size could be beneficial.

In contrast, ResEmoteNet demonstrated a stable learning curve, as depicted in Figure 4. This stability suggests that the current configurations are well-optimized, effectively avoiding overfitting issues.

5.4 Real-time inference

Real-time inference was conducted to evaluate the speed of the AI models. The frames per second (FPS) for each model were recorded in Table 5.3.

Table 5.3: Frames per second of AI models

AI model	Average FPS
L2CS-Net	23.40
ResEmoteNet	32.53

These results suggest that performance improvements are necessary, as the speeds are relatively slow.

5.5 Project Schedule

Table 5.2 and 5.3 show the project schedule for semester 1 and 2. We have completed the deliverables of Phase 1, which is a detailed project plan, and the deliverables of Phase 2, which is the interim report. We are currently training the AI models.

Table 5.4 Semester 1 schedule

Oct 1, 2024	**Deliverables of Phase 1** • **Detailed project plan**
Jan 6, 2025	Finalise first presentation and interim report

Table 5.5 Semester 2 schedule

Jan 13-17, 2025	First presentation
Jan 26, 2025	**Deliverables of Phase 2** • **Preliminary implementation** • **Detailed interim report**

Feb 1, 2025	Complete implementation of attentiveness calculation
Feb 10, 2025	Finish training the AI models
Feb 20, 2025	Acquire permission to use EduNet and Gaze360 datasets Prepare other datasets
Mar 1, 2025	Complete UI implementation
Mar 10, 2025	Finish testing
Apr 1, 2025	Complete final report
Apr 20, 2025	Finish presentation preparation and video
Apr 21, 2025	**Deliverables of Phase 3** • **Finalized tested implementation** • **Final report**
Apr 22-26, 2025	Final presentation
Apr 29, 2025	Prepare for project exhibition
Apr 30, 2025	Project exhibition • 3-min video

5.6 Challenges

Firstly, the datasets for the L2CS-Net and ResEmoteNets are not sufficiently large. Finding additional datasets with the same classes poses a challenge. Therefore, preprocessing will be necessary to ensure that all datasets have the same number of classes and formats. We will explore platforms like Kaggle and Roboflow to locate more suitable datasets.

Secondly, we encountered performance issues. During training and inference, we utilised an RTX 2080 Ti from the GPU farm phase 2. Real-time inference tests revealed that the FPS for the models was relatively low, indicating that the program may not be able to perform real-time inference for all three models simultaneously. To address this, we plan to convert the models to ONNX format. Additionally, we will optimise the code used to implement the AI models for improved performance.

5.7 Future plans

We will seek permission to use the EduNet and Gaze360 datasets for training the student behaviour detection model. Furthermore, we aim to find additional datasets for both the L2CS-Net and ResEmoteNets. With these datasets, we will train improved models. Lastly, we will develop a prototype for the application.

6. Conclusion

In response to the issue of students becoming distracted during online classes, our project aims to create an AI attention monitoring system to help students stay focused. We target the monitoring of student inattentiveness using the L2CS-Net, Facial Landmark Detection HRNet, Emotion Detection, and YOLOv8 models.

After completing our literature review, we developed our implementation for calculating attentiveness based on the outputs of these four models. We have commenced training two of the models and recognised the need to find additional datasets. Additionally, we have outlined potential challenges and proposed solutions as we continue to advance our project.

References

[1] C. B. Hodges, S. Moore, B. B. Lockee, T. Trust and M. A. Bond, "The difference between emergency remote teaching and online learning," 2020.

[2] K. A. Aivaz and D. Teodorescu, "College Students' Distractions from Learning Caused by Multitasking in Online vs. Face-to-Face Classes: A Case Study at a Public University in Romania," *Int J Environ Res Public Health,* vol. 18, 2022.

[3] R. Junco, "In-class multitasking and academic performance," *Computers & Education,* vol. 59, no. 2, pp. 505-514, 2012.

[4] A. J. Mendez, "The effects of online learning on student engagement: A study on sleepiness," *Journal of Online Learning Research,* vol. 6, no. 1, pp. 45-63, 2020.

[5] M. Elbawab and R. Henriques, "Machine Learning applied to student attentiveness detection: Using emotional and non-emotional measures," *Educ Inf Technol,* vol. 28, p. 15717–15737, 2023.

[6] M. K. Hossen and M. S. Uddin, "Attention monitoring of students during online classes using XGBoost classifier," *Computers and Education: Artificial Intelligence,* vol. 5, 2023.

[7] K. Vignesh, V. D. N. Gupta and H. Kumar, "Attentiveness Recognition System for Online Classes using OpenCV," in *2022 2nd International Conference on Intelligent Technologies (CONIT)*, Hubli, India, 2022.

[8] Z. Trabelsi, F. Alnajjar, M. M. A. Parambil, M. Gochoo and L. Ali, "Real-Time Attention Monitoring System for Classroom: A Deep Learning Approach for Student's Behavior Recognition," *Big Data and Cognitive Computing,* vol. 7, no. 1, p. 48, 2023.

[9] A. A. Abdelrahman, T. Hempel, A. Khalifa, A. Al-Hamadi and L. Dinges, "L2cs-net: Fine-grained gaze estimation in unconstrained environments," in *2023 8th International Conference on Frontiers of Signal Processing (ICFSP)*, 2023.

[10] G. Ying, "facial-landmark-detection-hrnet," [Online]. Available: https://github.com/yinguobing/facial-landmark-detection-hrnet. [Accessed 25 January 2025].

[11] A. K. Roy, H. K. Kathania, A. Sharma, A. Dey and M. S. A. Ansari, "ResEmoteNet: bridging accuracy and loss reduction in facial emotion recognition," 2024.

[12] D. Reis, J. Kupec, J. Hong and A. Daoudi, "Real-Time Flying Object Detection with YOLOv8," 2023.

[13] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou and M. Pantic, "300 Faces In-The-Wild Challenge: database and results," *Image and Vision Computing,* vol. 47, pp. 3-18, 2016.

[14] P. Korshunov and S. Marcel, "Speaker inconsistency detection in tampered video," in *2018 26th European signal processing conference (EUSIPCO)*, Rome, Italy, 2018.

[15] N. A. Shah, K. Meenakshi, A. Agarwal and S. Sivasubramanian, "Assessment of Student Attentiveness to E-Learning by Monitoring Behavioural Elements," in *2021 International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India, 2021.

[16] K. Altuwairqi, S. K. Jarraya, A. Allinjawi and M. Hammam, "A new emotion–based affective model to detect student's engagement," *Journal of King Saud University - Computer and Information Sciences,* vol. 33, no. 1, pp. 99-109, 2021.

[17] L. Q. Thao, D. T. Kien, N. C. Bach, D. T. T. Thuy, L. T. M. Thuy, D. D. Cuong, N. H. M. Hieu, N. H. T. Dang, P. X. Bach and L. P. M. Hieu, "Monitoring and improving student attention using deep learning and wireless sensor networks," *Sensors and Actuators A: Physical,* vol. 367, 2024.

[18] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick and P. Dollár, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference*, Zurich, Switzerland, 2014.

[19] V. Sharma and M. Gupta, "Classroom-Monitoring-Action-Dataset," [Online]. Available: https://github.com/vijetait/Classroom-Monitoring-Action-Dataset?tab=readme-ov-file. [Accessed 30 September 2024].

[20] G. Ying, "facial-landmark-dataset," [Online]. Available: https://github.com/yinguobing/facial-landmark-dataset. [Accessed 26 January 2025].

[21] J. Shen, S. Zafeirious, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos and M. Pantic, "The first facial landmark tracking in-the-wild challenge: Benchmark and results," in *ICCVW*, 2015.

[22] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *CVPR*, 2012.

[23] X. Zhu, Z. Lei, X. Liu, H. Shi and S. Z. Li, "Face alignment across large poses: A 3d solution," in *CVPR*, 2016.

[24] V. Le, J. Brandt, Z. Lin, L. Bourdev and T. S. Huang, "Interactive facial feature localization," in *ECCV*, 2012.

[25] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou and M. Pantic, "A semi-automatic methodology for facial landmark annotation," in *CVPR*, 2013.

[26] P. Belhumeur, D. Jacobs, D. Kriegman and N. Kumar, "Localizing parts of faces using a consensus of exemplars," in *CVPR*, 2011.

[27] X. Zhang, Y. Sugano, M. Fritz and A. Bulling, "MPIIGaze: Real-World Dataset and Deep Appearance-Based Gaze Estimation," *IEEE transactions on pattern analysis and machine intelligence,* vol. 41, no. 1, pp. 162-175, 2017.

[28] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik and A. Torralba, "Gaze360: Physically unconstrained gaze estimation in the wild," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[29] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, et al., "Challenges in representation learning: A report on three machine learning contests," in *INCONIP*, 2013.

[30] J. Fredricks, P. C. Blumenfeld and A. Paris, "School engagement: Potential of the concept, state of the evidence.," *Review of Educational Research,* vol. 74, no. 1, pp. 59-109, 2004.