ClassID: Enabling Student Behavior Attribution from Ambient Classroom Sensing Systems

PRASOON PATIDAR, Carnegie Mellon University, USA
TRICIA J. NGOON, Carnegie Mellon University, USA
JOHN ZIMMERMAN, Carnegie Mellon University, USA
AMY OGAN, Carnegie Mellon University, USA
YUVRAJ AGARWAL, Carnegie Mellon University, USA

Ambient classroom sensing systems offer a scalable and non-intrusive way to find connections between instructor actions and student behaviors, creating data that can improve teaching and learning. While these systems effectively provide aggregate data, getting reliable individual student-level information is difficult due to occlusion or movements. Individual data can help in understanding equitable student participation, but it requires identifiable data or individual instrumentation. We propose ClassID, a data attribution method for within a class session and across multiple sessions of a course without these constraints. For within-session, our approach assigns unique identifiers to 98% of students with 95% accuracy. It significantly reduces multiple ID assignments compared to the baseline approach (3 vs. 167) based on our testing on data from 15 classroom sessions. For across-session attributions, our approach, combined with student attendance, shows higher precision than the state-of-the-art approach (85% vs. 44%) on three courses. Finally, we present a set of four use cases to demonstrate how individual behavior attribution can enable a rich set of learning analytics, which is not possible with aggregate data alone.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing systems and tools**; • **Applied computing → Education**.

Additional Key Words and Phrases: Behavior Attribution, Classroom Sensing, Pedagogy, Computer Vision.

ACM Reference Format:
Prasoon Patidar, Tricia J. Ngoon, John Zimmerman, Amy Ogan, and Yuvraj Agarwal. 2024. ClassID: Enabling Student Behavior Attribution from Ambient Classroom Sensing Systems. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 2, Article 55 (June 2024), 28 pages. https://doi.org/10.1145/3659586

1 INTRODUCTION

Imagine, as an instructor, you had analytics that showed your class had frequent, active participation and regular in-class collaboration with their classmates. Imagine you then got an update showing that frequent, active participation typically only comes from 21% of the students in the class. Which of these reports would spur you to consider making changes to your teaching approach? Classroom observations with a professional development consultant can provide a snapshot of such information; valuable feedback about student behaviors, and suggestions for teaching strategies based on observations of one or two sessions. Several protocols for classroom observation provide complex measures about what instructors and students do in a classroom to give in-depth feedback about student engagement, some even at the individual student level [5, 33]. However,

Authors' addresses: Prasoon Patidar, prasoonpatidar@cmu.edu, Carnegie Mellon University, Pittsburgh, PA, USA; Tricia J. Ngoon, tngoon@andrew.cmu.edu, Carnegie Mellon University, Pittsburgh, PA, USA; John Zimmerman, johnz@cs.cmu.edu, Carnegie Mellon University, Pittsburgh, PA, USA; Amy Ogan, aeo@cs.cmu.edu, Carnegie Mellon University, Pittsburgh, PA, USA; Yuvraj Agarwal, yuvraj@cs.cmu.edu, Carnegie Mellon University, Pittsburgh, PA, USA.

ACM 2474-9567/2024/6-ART55
https://doi.org/10.1145/3659586

these protocols are highly labor-intensive both in training observers to code the protocol accurately and in the number of observers available [6]. Given the manual labor involved, repeated and longitudinal observations with associated analytics are not possible. Instructors do not have the opportunity to try various techniques and receive frequent observations to learn their own teaching strengths and weaknesses.

In recent years, techniques from the ubiquitous computing and sensing communities have offered a viable alternative to the scarcity and cost of highly trained teaching observers [3, 30, 46]. These systems, however, are either limited to research prototypes or small-scale, limited-time deployments due to the significant burden of maintaining them over the long term [97]. Ambient classroom sensing systems utilize sensors (camera, microphone, etc.) that are fixed in place to observe the entire classroom, and offer distinct advantages over instrumenting people or individual seats, including ease of deployment and simplified maintenance when compared to individualized instrumentation [97]. Notable examples include Edusense [1] and StuArt [94], which use video cameras and computer vision techniques to detect what is happening during class. These systems excel in the analysis of behavioral patterns in the aggregate, i.e., the number of students participating actively in a classroom session, the distribution of students across seating arrangements, and the extent of student movement within the class. However, there are still difficulties when it comes to disaggregating data and providing insights about individual students. These challenges arise from factors such as the ambient classroom sensing system's occasional inability to detect a particular student within a given frame accurately, instances of occlusion occurring when a student (or instructor) obstructs the view of another individual or the rapid and erratic movement of a student. These issues result in the erroneous assignment of more individual IDs than there are students in the classroom.

We evaluated Edusense [1], one of the state-of-the-art classroom sensing systems by running their open source implementation on our dataset, and found out that their ID assignment approach, which uses a combination of Euclidean distance and body inter-keypoint distance matching may sometimes assigns as many as $\tilde{2}00$ unique student identifiers for relatively small classes, ones with just 15 to 20 students. Consequently, the reliability of these systems to offer precise insights into individual behavioral patterns fails. This includes assessments of an individual's level of engagement and being able to differentiate between students who actively engage with the instructor, engage with their peers, or don't seem to engage with anyone. Furthermore, current ambient systems are constrained to single sessions and do not have the capability to track student behaviors consistently across multiple classroom sessions. Enabling such longitudinal analytics requires re-identifying anonymous students across non-overlapping sessions. The field of person re-identification (Re-ID) explores this challenge for security/surveillance applications [89]. However, these methods perform well only for individuals clearly visible in all frames, failing in dynamic crowded classroom environments [89]. Student behavior association across multiple sessions could be useful in understanding the consistency of student behaviors. Is a student always disengaged all session, or was their disengagement just for a particular session, maybe due to a particularly difficult topic covered on that day? While classroom observation protocols can help answer some of these questions, capturing the behaviors of all individual students is still impractical and difficult for a single human observer [25].

We introduce ClassID[1], a novel approach for longitudinal student tracking within and across classroom sessions without any student instrumentation. Our method leverages key attributes of the structured nature of classroom environments to make the problem more tractable. Specifically, we leverage the fact that the students in a course tend to remain generally consistent over the duration of the course (modulo adds/drops). Additionally, students predominantly face forward during most classroom activities [2, 9, 65]. In many types of classes, especially lecture-based, seats used by the students remain in the same physical location, and students often sit in the same or in a nearby seat across class sessions [29, 80]. These contextual cues provide continuity that facilitates persistent identification and re-identification. The consistency over sessions and directed focus together enable individual-level behavioral analysis. These semi-natural constraints help to link student actions to persistent

[1] www.github.com/edusense/ClassID

but anonymous digital identities, and we leverage these to improve the accuracy of assigning consistent student identifiers (IDs) in a computationally efficient manner.

ClassID utilizes videos captured from a 4K commodity camera placed at the front of the classroom looking at the students. This footage is input to an ensemble of five pre-trained neural networks for multi-object tracking [15], pose estimation [74], gaze estimation [2], finding of facial regions [23], and extraction of facial features [68]. These networks extract behavioral cues at the individual level at a video frame level. We propose a filtering and reconciliation technique to reduce erroneously assigned multiple identifiers into reliable within-session identifiers, which persist over time despite challenges posed by occlusions and student movements in a dynamic classroom environment. For re-identification across sessions, we generate per-individual representations integrating gaze patterns and facial features extracted from each session. A constrained matching process then associates these signatures across multiple sessions, enabling persistent tracking of individuals. Our approach goes beyond simply combining existing ML algorithms by adapting and integrating them in a context-aware manner and introducing new methods, contributing to the advancement of ambient classroom sensing and learning analytics. This context-aware approach differentiates our work from generic multi-object tracking and re-identification methods, establishing a foundation for ambient classroom sensing systems to support equitable learning experiences.

We evaluate ClassID on 15 classroom sessions spanning three distinct courses. Our evaluation show that for within-session attribution, ClassID can assign unique IDs to 98% students with an accuracy of 95% on average and significantly reduce erroneous assignment of multiple identifiers for the same individuals when compared with baseline approach (3 *vs.* 167 on average). Our cross-session matching also shows promise, attaining higher precision than a top re-identification technique (85% *vs.* 44% on average) when combined with attendance information at session level. Finally, we show four use cases to demonstrate how we can enable analytics using individual-level data that is not possible with aggregate-level data.

In summary, we make the following contributions:

- We propose ClassID, a novel method to (a) assign anonymous identifiers to students within classroom sessions for reliable behavior attribution without any student instrumentation, and (b) create session-level representation for individuals and match these representations to re-identify students across sessions.
- We evaluate ClassID on 15 classroom sessions across three courses, and show that it demonstrates reliable within-session ID assignment (achieving 95% accuracy on 98% students on average) and does significantly better than a baseline approach (Edusense [1]) in reducing erroneous multiple ID assignments (3 *vs.* 167 on average). For across-session ID matching, ClassID outperforms generic re-identification methods (85% *vs.* 44%) by developing session-level ID representations tailored to crowded classroom dynamics and leveraging instructor-provided attendance constraints.
- We present four use cases showcasing new insights into student participation, engagement, and interactions over long timescales, which is not possible from aggregated data. These analytics can enhance existing classroom observation protocols and multimodal learning analytics systems with more granular data about student learning behaviors.

2 BACKGROUND & RELATED WORK

In this section, we focus on the importance of understanding individual student behaviors instead of just looking at aggregate data. We also look at how current ambient classroom sensing systems assign IDs to students during a single session and recent studies on person re-identification across multiple cameras.

2.1 Importance of Individual Behavior Attribution in Classroom Setting

Classroom observation protocols can help instructors gain better awareness of their and students' behaviors toward better student engagement. These observations involve a professional development expert attending a class session and coding classroom behaviors in real-time according to a set protocol. For example, the Classroom Observation Protocol for Undergraduate STEM (COPUS) classifies active and passive teaching behaviors to give instructors feedback about active teaching implementation [73]. While some of these protocols capture data in the aggregate, some go further in capturing student engagement on group and individual levels. This can be important for understanding how instructors teach equitably across all students [63, 77]. Solely aggregate data may not fully capture the nuance of behaviors, particularly in understanding student engagement and learning. For example, Reinholz and Shah [13, 62] found that disaggregated data visualizations of individual student data drew teachers' attention to social markers of equity. For classroom observation protocols that capture individual student information, the Equity Quantified in Participation (EQUIP) observation tool consists of a teacher or observer marking each time each individual student participates in class discussion [63]. The VaNTH Observation System (VOS) includes capturing counts of which students are engaged and how [35]. The Student Resistance and Instructional Practices (StRIP) [71] measures how students respond to and behave in active learning activities. These observation protocols give a structure to classroom behaviors and engagement. However, training observers in any of these protocols is extremely labor-intensive [5]. Observations are also time-consuming and limited by the resources of the institution. Further, consistent observations over time are rare because of the limitations of the number of observers available.

In recent years, researchers used wearable sensors to understand students' attention and engagement using biomarkers [30, 40, 91], and show that anonymous engagement tracking can provide teachers with student engagement levels and help teachers understand the impact of different teaching contents on student engagement, thereby better-adjusting teaching speed and teaching methods [30]. A recent interview with an instructor shows that interpreting individual assessment is easier than aggregate assessment and provides more opportunities for actionable insights [61]. This shows that there is a growing need to provide disaggregated analysis across students in classroom settings, even if those students are not identified by name.

2.2 Student ID Assignment with Ambient Classroom Sensing Systems

Recent classroom video analysis research aims to study student behaviors by aggregating data from all people detected to characterize the overall dynamics of the scene rather than individuals [97]. Common approaches extract poses [1, 17, 37, 43], gaze [2, 34, 72], and facial features [76] from classroom video clips to identify behavioral events like hand raises, looking down or towards the instructor, engaging with peers, etc. [11, 12, 49, 72]. However, attribution of these behaviors to individual students remains challenging due to occlusion from low camera angles, sitting positions obscuring lower bodies, and close student proximity [37]. Thus, existing methods focus on snippet-based pattern detection rather than attributing behaviors to students over time. Recent systems like EduSense [1] and STUART [94] introduce interframe pose tracking to generate student IDs throughout sessions. However, pose estimation is susceptible to missed detection of people due to occlusions and produces multiple IDs when students pass each other as they enter/exit in the camera view. Prior research leverages deep learning-based multi-object tracking algorithms [21] to mitigate these missed detections in other learning settings [41]. Our work addresses these shortcomings by combining state-of-the-art multi-object tracking algorithms [15] with gaze estimation and analyzing facial features. This integrated framework leverages multiple cues for persistent identification, mitigating the impact of any single modality's failures.

2.3 Individual Re-identification Using Deep Learning Methods

Prior classroom sensing systems do not support consistently identifying students across classroom sessions. Nevertheless, the concept of individual re-identification (Re-ID) has been extensively explored in other domains, such as security and surveillance [89]. Re-ID methods are designed to locate a specific individual across multiple non-overlapping cameras or even the same camera at distant time intervals [31]. Many of these techniques operate within the constraints of a closed-world scenario, assuming factors like a clear and complete view of the person's body, sufficient annotated training data, and a definite match of the person of interest within the available data [89]. Additionally, researchers have delved into unsupervised learning strategies, including dynamic graph matching [88] and clustering coupled with model training using pseudo labels [22]. Notably, these unsupervised approaches have not yet reached the level of performance achieved by supervised methods on established benchmark datasets [89]. Re-ID, in a generic setting, is still an open challenge due to factors like the presence of different viewpoints [38], illumination changes [36], and unconstrained poses [67]. However, some of these challenges are relaxed in a classroom context. The camera viewpoint remains relatively consistent, and students usually face toward the front of the classroom. Consequently, this setting presents a unique opportunity for the exploration and potential development of methodologies that might work well in this specific setting. In this paper, we propose an approach that uses gaze estimation, facial features, and seating preferences of students to facilitate the assignment of consistent student identifiers across different classroom sessions. Our results indicate that our method outperforms one of the leading Re-identification (Re-ID) methods that utilize a deep learning model trained on large-scale datasets in a more generic setting.

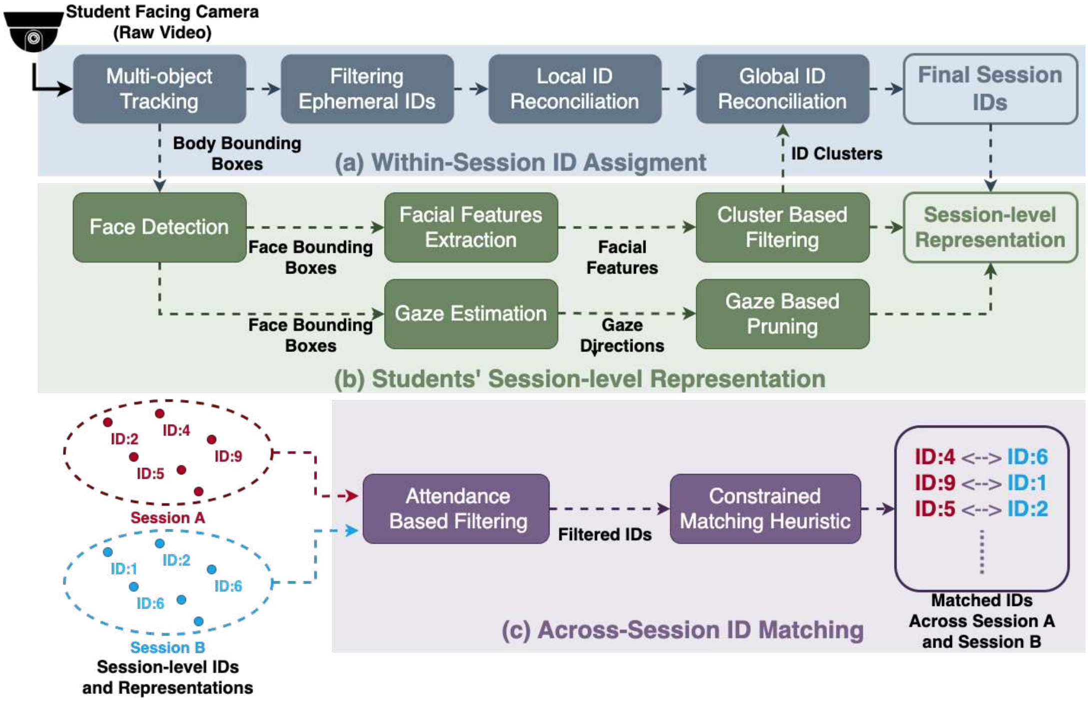


Fig. 1. A high-level overview of ClassID. (a) *Within-Session ID assignment* (§3.1) assigns unique identifiers (IDs) to students using a multi-object tracking approach (§3.1.1), followed by ID filtering and reconciliation (§3.1.2,§3.1.3 and §3.1.5), (b) *Students' Session-level Representation* extracts behavioral cues using face and gaze estimation (§3.1.4) for global ID reconciliation and generating session-level representation (§3.2.1) and (c) *Across-Session ID Matching* (§3.2.2) matches session level representation to match ID-pairs for same student across sessions.

3 SYSTEM DESIGN

We present a high-level overview of ClassID in Figure 1, comprising three key components: (a) *Within-Session ID assignment*, which processes raw video data from the classroom camera to generate within-session student identifiers using a multi-object tracking approach. Each unique identifier, or ID, is represented by a unique integer and set of bounding boxes across multiple frames that track the location of an individual across video frames. The initial ID assignment is followed by ID filtering and reconciliation approaches. (b) *Individual Session-level Representation*, which extracts behavioral information, including gaze and facial features, and processes them to generate session-level individual representations, and (c) *Across-Session ID Matching*, which uses session-level individual representations to match IDs belonging to the same individual across sessions. In Section 3.1 and 3.2, we discuss the design of each component in detail and present our end-to-end implementation in Section 3.3.

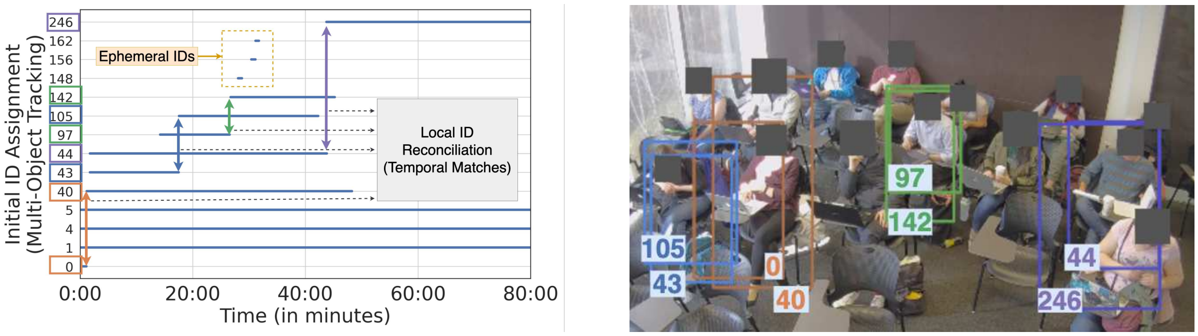
3.1 Assigning Student IDs within Classroom Session

In this section, we describe how ClassID assigns individual identifiers within classroom sessions. We start with using multi-object tracking (Section 3.1.1) to assign initial student IDs, followed by filtering ephemeral IDs (Section 3.1.2) and local ID reconciliation (Section 3.1.3). Then, we generate behavioral cues for all individuals at the frame level (Section 3.1.4), which are used for global ID reconciliation (Section 3.1.5).

3.1.1 Initial ID Assignment using Body Detection and Multi-Object Tracking: Prior approaches for student tracking rely primarily on pose estimation and inter-frame post-processing to assign session-level identifiers (IDs) to individuals in a classroom [1, 37]. However, unreliable pose detection often results in missed student detections due to occlusions and lighting conditions [37]. We instead perform whole-body detection as a more robust alternative to locate all individuals in the scene [28]. Next, we experimented with several state-of-the-art multi-object tracking techniques for persistent ID assignment. Numerous motion-based tracking methods exist, including classical algorithms like Kalman filtering [19] and SORT (Simple, Online, and Real-time) [83]. However, these techniques perform poorly in complex, nonlinear motions, which are commonly exhibited in lively classroom environments with students leaning, turning, and moving around. Recent methods like DeepSORT [84], ByteTrack [92], QDTrack [57] and OC-SORT [15] help overcome these limitations by learning robust appearance and motion models when trained on large, diverse benchmark tracking datasets (MOTChallenge [81], CrowdHuman [70] and DanceTrack [75]). We performed an empirical evaluation of these state-of-the-art methods in a recorded classroom session containing students seated in close proximity. We observed that OC-SORT performs better than the alternatives, as it emphasizes improving robustness in scenarios involving complex nonlinear motions and frequent occlusions, two primary factors behind poor tracking performance in classroom contexts [37].

Although OC-SORT shows impressive precision in its initial ID assignments, with each student receiving a single identifier, we observe that redundant IDs proliferate for certain active individuals. This occurs when students turn their heads around to chat with peers or temporarily vanish from the camera's view. During group work sessions, complex back-and-forth motions also disrupt tracking, triggering new IDs for already detected students. Likewise, brief exits from the camera frame, such as using the pencil sharpener or trash can, cause students to be improperly re-identified upon returning to their seats. We develop a novel pipeline tailored to classroom settings to overcome these challenges of ephemeral occlusions, complex movements, and irregular exits and returns. Specifically, our pipeline filters and consolidates the initial set of OC-SORT ID assignments into final, consistent ID assignments.

3.1.2 Removing Ephemeral ID assignments: A majority of individual movement in and out of the classroom happens during either the start or the end of classroom sessions and near the classroom entrance. Thus, IDs generated during the middle of the session and away from the entry/exit areas likely indicate ID assignment errors or changes. We characterize potential spurious identities by recording every ID's first detection frame,



(a) Temporal View

(b) Spatial View

Fig. 2. An illustration for local ID reconciliation in a single session, (a) shows a (partial) temporal view of correcting student ID assignments using local ID reconciliation. Each line depicts the start and end points of student IDs assigned by the multi-object tracking system in a class session. We first remove ephemeral IDs (*i.e., ID:148, ID:156, ID:162, etc.*), suggesting incorrect matching of objects to IDs. We then match the start and end of IDs belonging to the same student but erroneously separated due to quick movements (*i.e., ID:0⟶ID:40, ID:43⟶ID:105, etc.*). A spatial view in (b) shows the overlap of bounding boxes for student IDs reconcilied. In (a), we see *ID:40* and *ID:44* all start when *ID:0* ends. However, only *ID:40* substantially overlaps, being the valid match for the same student. Similarly, overlapped bounding boxes of reconciled IDs are shown for other successful matches. Students' faces are removed due to institute review board restrictions. In summary, reconciliation combines temporal and spatial information to correct erroneous ID assignments produced by an imperfect tracking approach.

body bounding box location, and duration. We observed that ephemeral IDs arise when OC-SORT incorrectly detects an object to be an individual or when a single individual is assigned two IDs due to more animated movements (see Figure 2). We start the process of ID reconciliation by removing these ephemerally generated IDs. To do so, we mark the specific frames when these IDs first appear and when they stop appearing. Further, we also focus on the number of frames these IDs were present. Finally, we remove any IDs that are present for less than a given time period threshold (i.e., a minute based on our experiments) across an entire session.

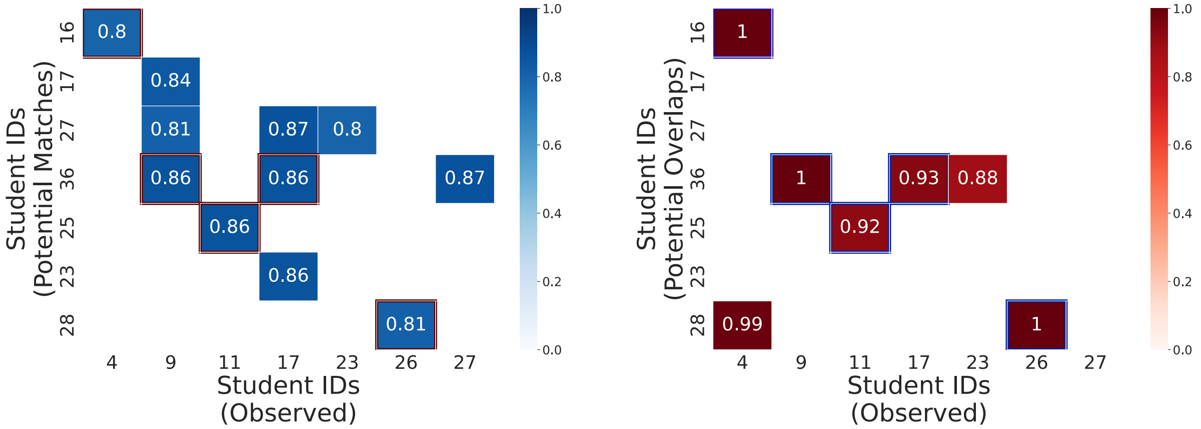
3.1.3 Local ID reconciliation via Spatial-Temporal Analysis: Once we remove ephemeral generated IDs, we reconcile the remaining duplicates using a spatio-temporal consistency heuristic. Classroom motions, even if abrupt and spurious, tend to be localized and not prolonged over long periods[29, 80]. As a next step, we reconcile multiple ID assignments to a given individual due to short-period events, i.e., significant occlusion, students bowing down to look for something in their bag, or students talking to each other. These new redundant IDs for an already tracked student tend to be temporally non-overlapping, with one identity ending shortly before the next begins (see Figure 2 (a)). We extract potential ID matches across all assigned IDs based on the temporal proximity of frames these IDs start and finish. Further, we refine these matches based on how much the bounding boxes of matched IDs overlap based on the ratio of frame area covered by the intersection of bounding boxes and union of bounding boxes (see Figure 2 (b)). Finally, we select the objectively best match for each ID by choosing the ID with maximum bounding box overlap. The two-step process ensures that we do not merge ID assignments for two students sitting in close vicinity. In summary, this filtering and reconciliation stage integrates knowledge of classroom layout and movement dynamics to transform initial OC-SORT ID assignments into more consistent ID assignments with better one-to-one fidelity with students.

3.1.4 Detecting Multi-Modal Behavioral Cues and Global ID Reconciliation. In the next step, we detect students' behavioral cues using video frames. Detailed characterization of student actions, attention, and responses during classroom activities offer invaluable insights into the learning process [97]. This information can be captured using behavioral cues such as body/hand poses, head movements, facial expressions, and gaze, thus providing the opportunity to understand student participation on an individual level. For example, visual markers of fatigue could reveal challenges with session pacing or content for specific students [66]. Disengaged patterns like frequent distraction may highlight the need for differentiated instruction [78]. Equitable participation across students could be assessed through the distribution of hand raising or on-task gaze direction [10, 14, 97]. To extract rich multi-modal behavioral cues from video frames, we leverage an ensemble of state-of-the-art vision techniques tailored to the classroom context.

To extract rich behavioral cues at an individual level, we leverage the personalized body bounding boxes generated during multi-object tracking with OC-SORT [15]. These detections focus on precise pose estimation of individual students. We employ a High-Resolution Representation Learning Network [74] trained on the diverse COCO datasets [42] to detect body keypoints. This provides robust skeletal poses for each individual. Next, for extracting precise facial regions, we use RetinaFace [23], a state-of-the-art single-shot face detector. RetinaFace is a single-stage face detector that performs pixel-wise face localization across different sizes of faces. This proves critical for classrooms, where student facial size varies significantly based on proximity to the front-facing camera. RetinaFace is trained to identify faces in the wild and surpasses other state-of-the-art algorithms in terms of the precision of face detection [48]. Notably, in many cases, the body bounding box of an individual encompasses more than one face due to the close proximity of students within the classroom setting. To mitigate the possibility of erroneously detecting additional faces within an individual's bounding box, we restrict our focus to identifying the topface as students face the top-mounted front camera among all faces detected within the bounding box. We trim the detected face regions and send them to ClassGaze [2], a 6-degree-of-freedom gaze estimator to obtain head pose (i.e., roll, pitch, and yaw with reference to camera viewpoint) as a proxy for visual attention. Finally, we generate semantic facial representations (encoded as a 512-dimensional vector), for each student using an InceptionResnetV1 model [68] trained on diverse facial datasets [16, 90]. This provides a robust representation invariant to visual variations [68]. By combining body pose, estimated gaze direction, and learned facial representations, we capture rich information for every individual in each frame. This characterization further enables both more precise ID assignment by pruning erroneous identities (see Section 3.2.1) and deeper latent pattern mining over long timescales (see Section 6).

3.1.5 Global ID reconciliation using individual behavioral cues: The local ID reconciliation corrects any inconsistent ID assignments that happen due to local spatio-temporal factors. However, it does not consider people moving in and out of the camera frame. This cannot be resolved by just using a spatio-temporal consistency heuristic. These kinds of movements are very common in cases when students take a break in the middle of the session or they move to the front of the classroom for presentations. In some cases, we also see instructors moving in an out-of-camera frame when they visit a student, and they are mistakenly identified as another student in the class. These cases are not captured very well using local features. We utilize individual behavioral cues (see Section 3.1.4) to reconcile IDs in these situations.

We start by extracting IDs that are not consistent for the majority of the classroom session. For each of the selected (or Observed) IDs, we match them with IDs that do not temporally overlap with them. We assess each of the matches to find potential cases when these IDs belong to the same person (or Potential Matches). To do so, we examine the spatial proximity of the bounding boxes associated with these two IDs. Next, we create an individual representation for each ID and quantify how close these individual representations are using the Cosine distance metric (see Figure 3 (a)). These individual representations are created by computing the median of facial feature vectors across frames for each ID. We determined the threshold for closeness empirically. Ultimately, potential matches that share spatial proximity (or Potential Overlaps) within the classroom area are consolidated into



(a) Visual Similarity Scores

(b) Spatial Overlapping Scores

Fig. 3. An illustration demonstrating global ID reconciliation to match individuals as they enter and exit camera view across a single session. X-axis represents IDs that are being observed or assigned first, and Y-axis represents IDs that are assigned later, which can be a potential match to observed IDs, and highlighted boxes show successful matches for reconciliation. The Figure on the left (a) displays the visual similarity between non-overlapping IDs in time. A single ID may match multiple others owing to visual commonalities (e.g. *ID:9* exhibits similarity to *ID:17*, *ID:27*, and *ID:36*). The Figure on the right (b), we analyze the spatial overlap between matched IDs, delineating location congruency between student IDs paired by visual semblance in (a). *ID:9* demonstrates significant overlap solely with *ID:36*, despite similarities with other IDs. This verifies *ID:9* and *ID:36* as positive matches. *ID:4* spatially overlaps both *ID:28* and *ID:16*, yet is only visually analogous to *ID:16*. This indicates *ID:28* and *ID:16* represents distinct students in close proximity. This additional step hinders erroneously pairing visually discrete individuals by cross-validating visual and spatial reliability across IDs.

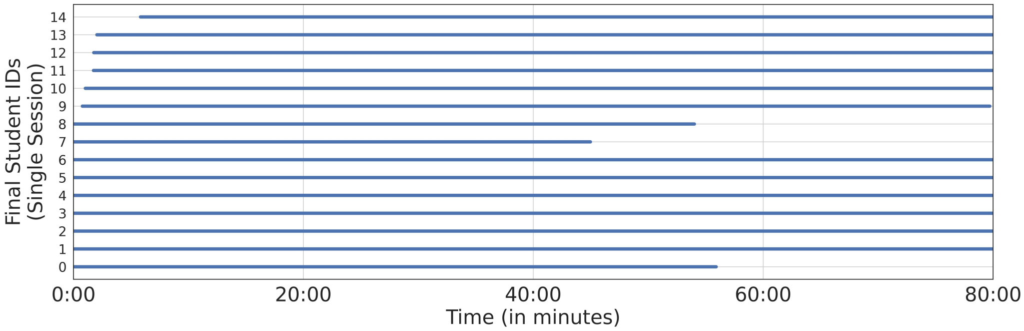


Fig. 4. Final student identification assignments after applying local and global reconciliation procedures on the original tracking output. The visualization shows consistent IDs maintained for all students throughout the class session, in contrast to the initial ID assignment in Figure 2. While multi-object tracking alone is insufficient, our system's automated reconciliations correct short-term mismatches and long-term gaps, resulting in persistent IDs and enabling the association of captured student behaviors to unique individuals across the entire classroom session.

a single student ID (see Figure 3 (b)). Finally, any remaining IDs appearing for under 10% of the session are removed, as they likely belong to instructors briefly entering and exiting the camera frame (see Figure 4).

3.2 Assigning Student IDs across Classroom Sessions

This section provides details on persistently associating student IDs across classroom sessions. We start by constructing session-level representations per individual for each session (Section 3.2.1). We then perform a bipartite matching between these individual representations across two sessions to reconcile IDs across them (Section 3.2.2). The session-level representations need to encapsulate the uniqueness of individuals' seating behavior and feature representation to allow matching well across separate sessions. Naively aggregating all frame-level representations per student proves to be ineffective, as we observed significant variance for the same individuals based on their head pose, expressions, etc.

3.2.1 Creating Session-level Representation for Individuals: A key challenge in representation learning is capturing individual uniqueness while enabling consistency across multiple sessions. We observed that naive aggregation of individual representations across all frames introduces variance from changing poses, expressions, etc. We propose two parallel filtering techniques to derive stable per-session signatures:

A. Gaze-based Pruning: To reduce variance in frame-level representations due to head movements, we remove frames where students look away from the camera. This excludes instances of writing, discussion, etc. The roll, pitch, and yaw of an individual's head estimate head rotation in 3 dimensions, and all frames where these values exceed angle thresholds are removed to focus on forward-facing instances.

B. Cluster-based Filtering: To filter erroneous facial feature representations due to errors in facial region mapping, we first extract facial features from each frame using a pre-trained model. These facial features are then projected into a latent space that captures the essential characteristics of the facial regions. We perform DBSCAN clustering [69], an unsupervised density-based spectral approach robust to outliers [53], on the latent space representations of the facial features across all frames. By clustering in the latent space, we can effectively group similar facial feature representations and identify outliers. The generated cluster centroids aggregate observation points solely using distance metrics without making any assumptions on the distribution of those points.

The gaze filtering and clustering stages yield two distinct session-level representations per individual, capturing frontal-facing and high-density facial features, respectively. To integrate these, we derive unified signatures through feature-wise averaging. The unified representation summarizes both focused visual attention and aggregated facial cues within a session. Additionally, prior work indicates students tend to choose similar sitting locations across different class sessions [39]. To incorporate this insight, the representation includes the face region area (in pixels) as a proxy for distance from the front of the room. This positional continuity further strengthens re-identification. The resulting session signatures encapsulate visual and spatial consistency to enable persistent matching.

3.2.2 Identity Association via a Constrained Matching Heuristic: Naively matching session representations by minimizing distance risks erroneous associations. Thus, we developed a custom matching pipeline, which only focuses on high-quality matches leveraging two insights:

- Large facial regions yield more reliable representations. We thus prioritize matches for identities with larger face areas (i.e., students sitting closer to the camera) first.
- True ID match associations likely rank highly (in terms of matching distance) for both sessions. We favor matches to candidate pairs appearing in mutual top-k lists, ensuring bi-directional consistency.

Based on these insights, the matching process proceeds iteratively as follows:

(1) Sort identities by descending order in terms of the face area (in the count of frame pixels) in both sessions.
(2) Match IDs where both appear in the other's k-nearest neighborhoods and are within a distance threshold.
(3) Remove matched pairs before repeating.
(4) For remaining candidates, match based on spatial proximity and representation closeness.

We associate likely similar students across sessions by constraining associations to mutually consistent candidates and then propagating outwards based on spatial proximity and representation closeness. The method prioritizes precision over recall to avoid false matches due to the challenging nature of re-identification. Overall, our approach integrates domain knowledge about classroom context into representation learning and matching for improving accuracy.

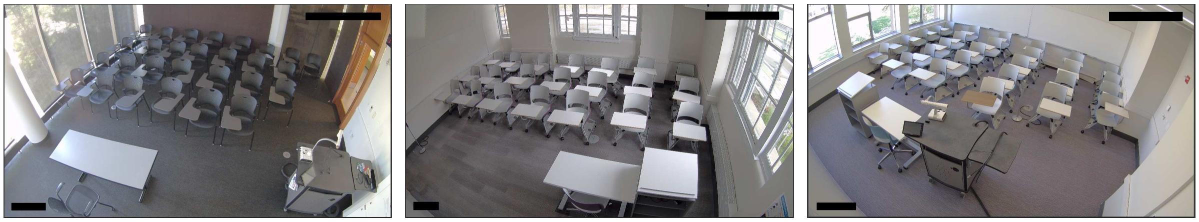
3.3 Implementation

Our entire pipeline is written in Python with over 5000 lines of code. For within-session ID assignment, we start from a raw video recorded from the classroom session as an input and output a serialized dictionary [59] consisting of session IDs, individual behavioral cues (i.e., body-pose, head-pose, and facial features) for all video frames, and session-level representation for each individual. For across-session ID matching, we take as input the session-level representation, and return the matching ID pairs for both sessions with match confidence. ClassID consists of four major components: (i) Individual ID assignment within a single session using a multi-object tracking method (OC-SORT), followed by ID filtering and local ID reconciliation; (ii) Estimating pose, gaze, and facial features from body bounding boxes; (iii) Global ID conciliation followed by building session level representation for all individuals and (iv) Matching IDs across all sessions in a pair-wise manner. All our deep learning models are implemented using PyTorch [58]. For our OC-SORT implementation, we use MMTrack [21], an open-source object tracking toolbox by OpenMMLab [20], and configured it with pre-trained weights from the original papers' trained model [15]. For pose estimation with the HRNet model, we use MMPose [50], an open-source 2D/3D pose estimation toolbox by OpenMMLab [20], and configured it with pre-trained weights from their benchmarks on the COCO Dataset [42]. For facial region detection and head pose estimation, we used an open-source implementation of ClassGaze [2]. For facial feature extraction using InceptionResnetV1, we used facenet_pytorch library [24], where the model weights are initialized using parameters ported from facenet implementation in tensorflow [68].

Our end-to-end pipeline, designed for multi-object tracking and feature extraction (including face detection, gaze estimation, and facial feature estimation), can process video recordings at approximately three frames per second. This means that for a 15 FPS camera, the processing time is around five times the recording duration. The preprocessing step for ID reconciliation takes an additional 10-15 minutes, depending on the length of the session and the number of individuals present. It is important to note that our pipeline prioritizes accuracy over speed, making it well-suited for post-processing recorded classroom sessions. However, there is potential for optimization of the feature extraction pipeline to enable real-time execution in the future. We have made ClassID openly available to the research community [60]. It can process recorded raw classroom video data and can also integrate with existing ambient classroom sensing systems to facilitate longitudinal, student-level analysis.

4 EVALUATION SETUP

Large lectures often limit student discussions due to their scale, while smaller recitations enable more engagement and demonstration of understanding [52]. Our system focuses on these interactive small groups (under 15-20 students), where closely monitoring each student is highly valuable. In recitations, instructors can provide tailored support and advice leveraging detailed behavioral data [45, 79]. Applying such analytics addresses a key need to objectively measure engagement linked to comprehension. While an active area involves developing systems for large classes via advanced sensing, we focus on the ubiquitous yet overlooked small-group setting using commodity hardware. Example use cases benefiting from our techniques include interactive seminar courses, workshops, and tutorials. By accurately assigning student IDs in small dynamic classes, our system can enrich these experiences to improve learning.



(a) Course A

(b) Course B

(c) Course C

Fig. 5. Snapshots from 4K IP cameras positioned at the front of various classrooms where data was collected. The cameras are attached to walls near Ethernet ports, capturing views of the classrooms from slightly different distances and angles based on room layout. The views show a variety of perspectives across typical classrooms. The variation of front-facing camera placements demonstrates the system's flexibility to work effectively in multiple classroom recording setups without needing adjustment for specific camera locations or orientations.

Our evaluation spanned 15 sessions across three courses (i.e., five sessions per course) - one for an advanced-level, seminar-style interactive lecture on *research in learning sciences* (Course A) with 80-minute classes where the instructor uses a seminar-style teaching, and two recitation sub-sections (50 minutes each) accompanying a large introductory course on *building software systems at scale* (Courses B and Course C) which focuses on teaching assistant (TA) supported problem-solving. Course A is taught by a senior professor with more than 15 years of teaching experience in the HCII community. Courses B and C are led by teaching assistants with less than five years of teaching experience. We randomly selected five sessions across these courses during a semester to see how well our system tracked various classroom interactions. Each classroom is instrumented with front-facing 4K cameras that capture student activity at 5 FPS. The classrooms featured movable furniture as depicted in Figure 5, which introduced additional challenges since students often rearranged their seating, compared to the fixed setup in a typical lecture hall. Our tracking system had to deal with potential blockages from view and changing perspectives. The videos from the recitations also caught common classroom scenarios, such as students arriving late or stepping out for a while.

We manually annotate ground truth by assigning unique persistent IDs to all visible individuals across five sessions per course. For manual verification of within-session ID tracking, we annotate every detection's ID in randomly sampled 1-minute chunks from the beginning, middle, and end of each session. In total, 13,500 frames were hand-annotated across all sessions, with 1̃60,000 ID annotations, providing a dense set of labels for quantitative analysis. Power analysis provides guidance on the sample size required to detect a minimum performance difference of 5% for 80% tracking accuracy, at a 95% confidence level and 5% margin of error. Specifically, for an anticipated tracking precision of 80%, a sample of 900 frames per session is estimated to provide 95% power to establish through a one-sample proportion test [18] that the true precision exceeds 75%. However, it is important to note that this power analysis assumes the underlying processes are ergodic, meaning that the 1-minute sampling captures the variability in the data consistently over time. In reality, student behavior may not always exhibit such temporal consistency, and the variability in tracking performance may not be fully captured by the selected samples. Consequently, while the power analysis provides a useful guideline for sample size determination, it is not an absolute guarantee of the model's performance bounds. The sampling strategy aims to ensure the representation of varied activity patterns throughout sessions, but the inherent limitations of power analysis in this context should be considered when interpreting the results. Despite these limitations, the annotations allow for an assessment of both within-session and cross-session ID assignment capabilities, providing valuable insights into the model's performance.

5 PERFORMANCE EVALUATION

In this section, we evaluate our ID assignment and matching approaches to answer the following key questions:

1. (Within-Session) Comparison with Baseline Approaches: How well does our approach do in terms of consistency of assigning student IDs when compared to a state-of-the-art baseline system for classroom sensing, Edusense [1]? Additionally, how does our approach perform in comparison to the actual student count using ground truth annotations?

2. (Within-Session) Post-processing Contributions: How much does each post-processing step contribute in terms of filtering and reconciling student IDs following the initial ID assignments by the OC-Sort algorithm?

3. (Within-Session) Manual Verification of Interframe Tracking: What is the overall accuracy and detection rate achieved by our approach when compared with frame-level annotations across all individuals? This step involves manual verification and serves to assess the validity and effectiveness of our methodology.

4.(Across-Session) Performance of student ID matching: How effective is our ID matching heuristic to match student IDs across multiple sessions in comparison to a baseline Re-ID (Re-identification) approach [96]?

5.1 (Within-Session) Comparison with Baseline Approach

Figure 6 illustrates the comparative analysis of the total unique student IDs assigned by our proposed approach as opposed to a baseline approach, namely EduSense [1], across three courses with five sessions each (total 15 sessions). EduSense employs a centroid-based interframe tracking method, which takes pose estimations of students as input and returns consistent IDs across multiple frames. We observe that there is a notable increase (167 on an average, sd=50) in erroneous re-assignments occurring throughout the session for the baseline approach, which can be attributed to factors such as partial occlusions, inaccuracies in pose estimation resulting from student congestion and student movements. As a point of comparison, for Session S1 for Course A, Edusense detects 200 student IDs, while the actual ground truth is only 18 students. In contrast, our approach demonstrates a substantial reduction in errors arising from these aforementioned factors, resulting in a notably higher level of consistency in ID assignments (e.g., 18 unique IDs for Session S1 for Course A). Our ground truth annotation for actual student count in the session shows that our approach has significantly fewer erroneous multiple assignments (3 on an average, sd=2) when compared with the baseline approach (167 on an average, sd=50).

5.2 (Within-Session) Post-processing Contributions

Figure 7 shows the contribution of each algorithmic step of our algorithm to get persistent student IDs. The initial ID assignment using the OC-SORT approach assigns a high number of IDs, exceeding that of the Edusense (baseline) method. However, the ID filtering step reduces assigned IDs by approximately 95% on average (sd=2%) by removing short-duration assignments, leaving only eligible IDs. Post ID filtering, Local ID reconciliation using spatiotemporal consistency heuristics further consolidates multiple IDs arising from short-term occlusions, reducing unique IDs by 37% on average (sd=10%) compared to post-filtering. Post Local ID reconciliation, Global reconciliation exploiting individual cues and session-level ID appearance consistency additionally decreases unique IDs by 33% on average (sd=15%) versus local ID reconciliation. This indicates each algorithmic step is essential, and provides meaningful contributions toward consistent ID assignment for all students.

5.3 (Within-Session) Manual Verification of Interframe Tracking

Manual verification demonstrates a strong performance of our proposed approach for persistent ID assignment and individual detection over an extended duration. Figure 8 compares our accuracy and detection rate compared to ground truth annotations for 900 frames (300 consecutive frames from the beginning, middle, and end) per session. Accuracy is defined as the percentage of correctly identified body bounding boxes out of total ID assignments across all frames. The detection rate is the ratio of the total detected body boxes to the total

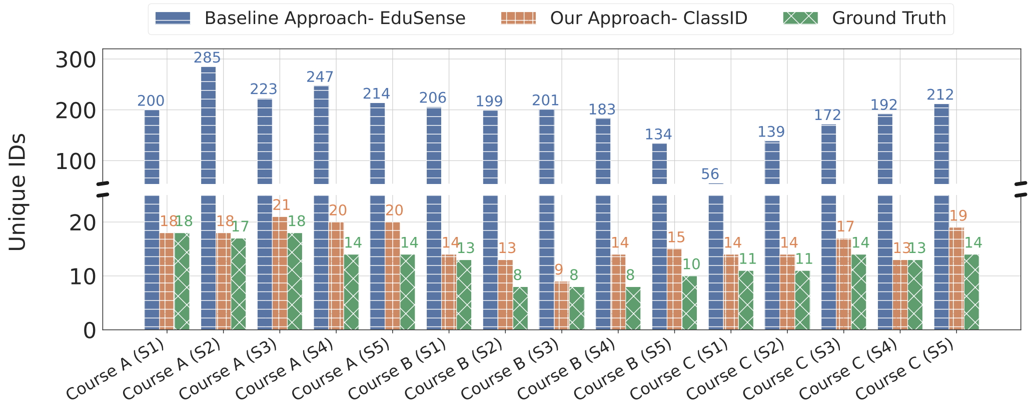


Fig. 6. Comparing the count of unique IDs assigned across all sessions between the baseline approach (EduSense), our approach, and ground truth annotation. The amount of unique IDs assigned by the baseline approach is significantly overestimated (167 on average, sd=50) as compared to our approach (3 on average, sd=2) with respect to ground truth.

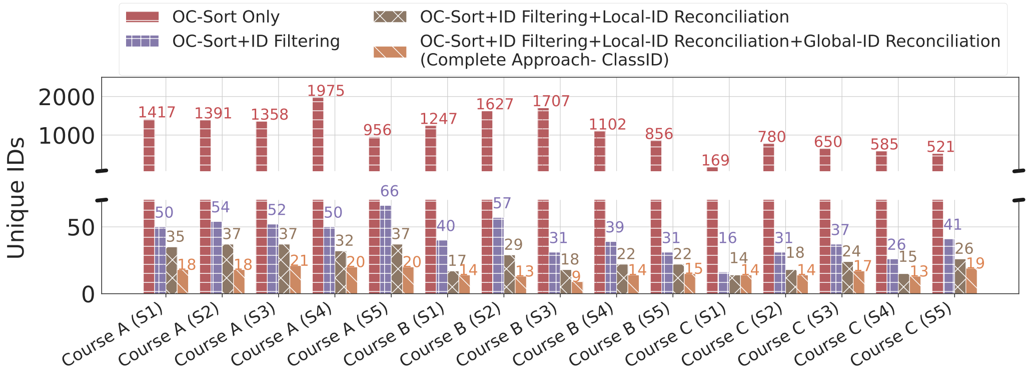


Fig. 7. Count of unique IDs assigned after every filtering and reconciliation step of our approach. In comparison to OC-SORT only, OC-SORT+ID Filtering has 95% fewer unique identifiers on average (sd=2%). Similarly, the next consecutive steps reduced the count of unique identifiers assigned by 37% and 33% on average (sd=10% and 15%) when compared with the previous step, showing the meaningful contribution of every step towards consistent ID assignment within a session.

individuals present, summed across all frames. Our approach achieves an average accuracy of 95% (sd=5%) in assigning persistent student IDs, with an average of 98% (sd=1.3%) individual detection rate. This indicates consistent detection and tracking of nearly all individuals throughout the session duration. In comparison, a prior state-of-the-art approach by Hur and Bosch [37], which utilizes pose detection followed by spatiotemporal postprocessing for interframe tracking attained 93% accuracy over just 40 pairs of consecutive frames across six sessions, with 77% detection rate. This shows that our approach significantly advances persistent identification and localization over the duration of a classroom session.

5.4 (Across-Session) Performance of student ID matching

To evaluate cross-session student ID matching performance, we developed a baseline approach using one of the top-5 re-identification methods pre-trained on large datasets [96]. Since none of the existing techniques

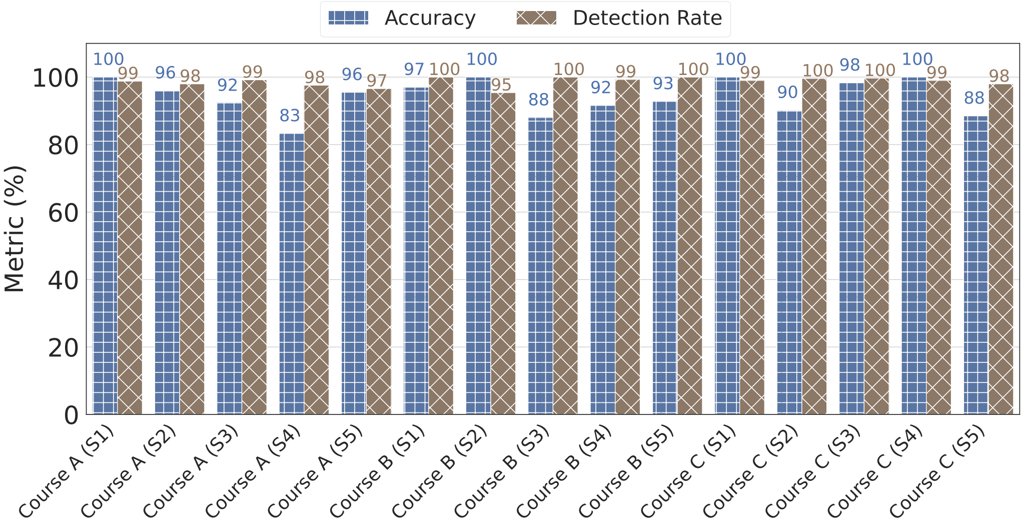
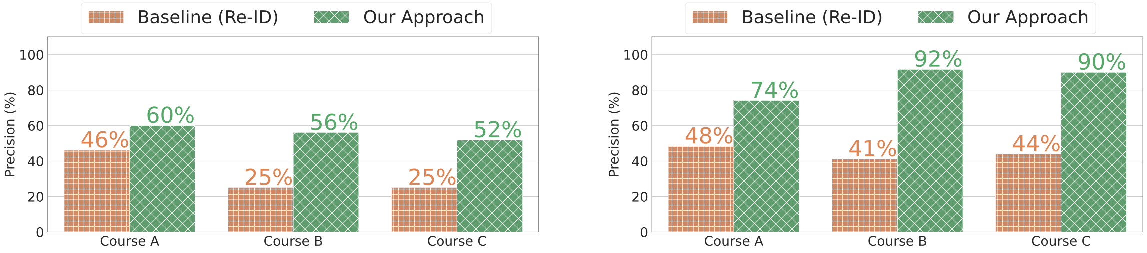


Fig. 8. Accuracy and Detection Rate for within-session ID assignment based on manual verification across all individuals and all frames. Our approach can accurately detect 98% of students on average (sd=1.3%), out of which 95% of students on average (sd=5%) are assigned unique identifiers accurately when compared with ground truth annotations at the frame level.

were specifically developed for classroom settings, we adapted the top-performing general re-identification algorithm [95], which achieves 85% accuracy on an average across large datasets [64, 82, 93]. To create a compelling classroom-targeted baseline, the baseline algorithm extracts features from body bounding boxes from our within-session approach and aggregates feature representations across frames to build session-level individual representations. Our matching heuristic is tuned to emphasize precision in matching over recall, as correctly identifying even a subset of students can enable useful applications [7]. We conducted analysis in two different settings.

1. No attendance information: In this setting, our system only receives raw anonymous session-level ID representations, without any metadata or assumptions. However, in real-world classrooms, attendance fluctuates each day, sometimes substantially, with students missing sessions sporadically. Non-student personnel like teaching assistants may also enter and exit the classroom. Combined with differences in clothing, lighting, and viewpoint, these issues introduce significant appearance changes across sessions. Without constraints, a student only attending one session could be incorrectly matched to a different random student of similar appearance attending the other session. As attendance variation increases across sessions, these types of invalid matches accumulate and accuracy greatly suffers.

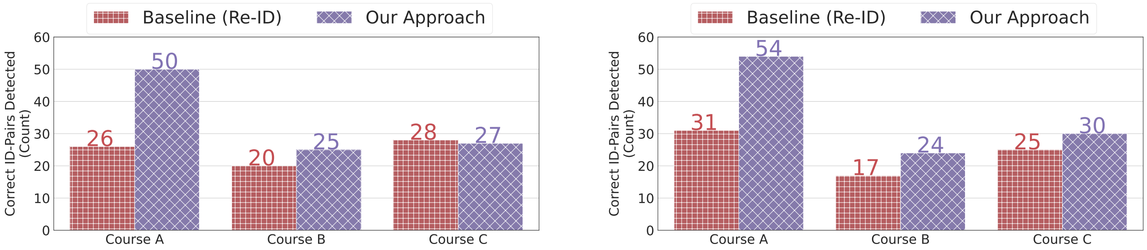
2. With attendance information: Here, we limit match candidates to students confirmed as attending both sessions through manual instructor attendance records. This prevents wasting computation on impossible cross-session pairs with students missing either session. Reliably tracking attendance does increase instructor effort and the risk of identity leakage. Instructors must log the student ID and session for each student, allowing leakage if both session IDs and attendance sheets are acquired. However, the risk is reasonable if attendance sheets are properly secured. Meanwhile, the constraints significantly reduce invalid match attempts, thus boosting precision. Further research could explore privacy-preserving attendance constraints to obtain accuracy gains while safeguarding student identities.



(a) No attendance information.

(b) With attendance information.

Fig. 9. Comparing precision of ID matching across all session-pairs in a course in (a) No attendance information and (b) With attendance information setting. Overall performance with attendance information is significantly improved (85% *Vs.* 56% on average, sd=8% *Vs.* 3%) as students absent from one of the sessions in the pair are not considered for matches. Also, our approach has better precision when compared with the baseline approach [95] in both no attendance information (56% *Vs.* 32% on average, sd=3% *Vs.* 10%) and with attendance information (85% *Vs.* 44% on average, sd=8% *Vs.* 3%) settings.



(a) No attendance information.

(b) With attendance information.

Fig. 10. Comparing count of correct ID-Pairs matched across all session-pairs in a course in (a) No attendance information and (b) With attendance information setting. Our approach has a higher count of ID-Pairs detected for all the courses when compared with the baseline [95] approach in both settings. This information, combined with higher precision (Figure 9) shows that our approach is conclusively better than baseline approaches for matching student IDs across sessions.

Figures 9 and 10 show the precision and count of correctly matched student IDs by our approach vs. the baseline across five sessions (ten session pairs) for three courses. We observe higher precision for our approach compared to the baseline in both with attendance information (85% *Vs.* 44% on average, sd=8% *Vs.* 3%) and no attendance information (56% *Vs.* 32% on average, sd=3% *Vs.* 10%) settings. Precision also greatly improves with attendance information (85% *Vs.* 56% on average, sd=8% *Vs.* 3%) as removing students attending only one session from consideration, we eliminate erroneous matches. Meanwhile, the per-session attendance logs provide reliable ground truth on which students could possibly match. Our superior performance proves our session-level representations and matching better capture identifying traits versus generic re-identification. The setting with no attendance information remains challenging due to fluctuating attendance and other variability. While our results show promise for re-identifying classroom students, further improvements are necessary for reliable application in real-world scenarios. It is important to acknowledge that these accuracies may compound and worsen when the system is used across a range of classes, potentially dropping as more sessions have imperfect accuracies. Additionally, the results reported here are limited by the variability of classrooms and student demographics in the sample. Therefore, to ensure engagement statistics are properly tracked across a quarter/semester, some manual matching from the instructor may be warranted to reduce the effects of imperfect student re-identification across classes. This manual matching may only need to be completed for a subset of students that the model is uncertain about. Further research could build upon these constrained session representations and matches to improve the robustness and generalizability of the system.

6 USE CASES FOR CLASSROOM ANALYTICS

Our system for persistent student identification can augment existing video-based classroom analytics platforms like EduSense [1], ClassroomDigitalTwins [2], and StuArt [94]. While these ambient intelligence systems generate rich multipart behavior models - tracking metrics like body poses, hand gestures, location, and speech - analytics remain restricted to frame-level aggregation across all students. Such summarized class participation averages overlook critical distributional equity issues [62]. This section presents novel use cases integrating our identifier with engagement analytics from sensing systems to showcase the importance of individual-level longitudinal measurement. By combining these frame-based metrics with consistent student identifiers over multiple sessions, we can uncover variance trends and equity gaps not evident in conventional summarized views. We focus analysis on two pedagogically relevant dimensions of participation and engagement well-established across classroom observation protocols [35, 63, 71]:

1. Active Classroom Participation: Active participation in a classroom is characterized by students taking an interactive role, such as posing questions to the instructor. Building on prior research establishes the act of raising a hand as a reliable indicator of active engagement between students and instructors during lecture sessions [14], we made use of a pre-trained hand raise detection classifier. This classifier has demonstrated a high degree of effectiveness, accurately identifying hand raises in 90% of cases within the context of a real-world classroom setting [1]. While behavioral cues like raising eyebrows or opening mouth for speaking can also represent finer details on student participation, estimating these cues at a distance is not reliable due to the decrease in pixel area for faces [55]. Thus, for the purpose of our study, we considered instances where one or more hand raises were detected in a single video frame—and subsequently compiled this data on a per-minute basis—to signify moments of active student participation.

2. Gaze-based Attention: The direction of a student's gaze serves as an insightful indicator of where their attention lies and is recognized as a proxy for gauging engagement [10, 97]. By observing the orientation of a student's head and gaze, instructors can infer the focus of the student's attention—whether they are engaged with the instructor's activities when looking upwards and towards the front, or concentrated on individual tasks such as note-taking when their gaze is directed downwards. We use a state-of-the-art gaze tracking methodology that exhibits an average deviation of 20.7° for yaw (horizontal head movement) and 17.6° for pitch (vertical head movement) in real-world classroom settings [2]. With the camera's strategic placement and using these yaw and pitch thresholds, we categorize the students' gaze into two key attention categories: instructor-focused (upward gaze) and self-focused (downward gaze). This classification system simplifies complex behavioral data, which we further aggregate in 10-second intervals to make it more interpretable.

6.1 Student-level Participation in a Single Session

Equitable participation in the classroom is important for ensuring all students can meaningfully engage in learning. Learners may have different levels of participation, but keeping track of these students is challenging. Some class observation protocols such as EQUIP [63], StRIP [71], and VOS [35] require that the observer note engagement levels or activities of individual students. With ClassID, individual student attribution can show this information without requiring the labor and time required of in-person observations. Figure 11 shows an exploded view of a single class session, illustrating individual student participation through a single class

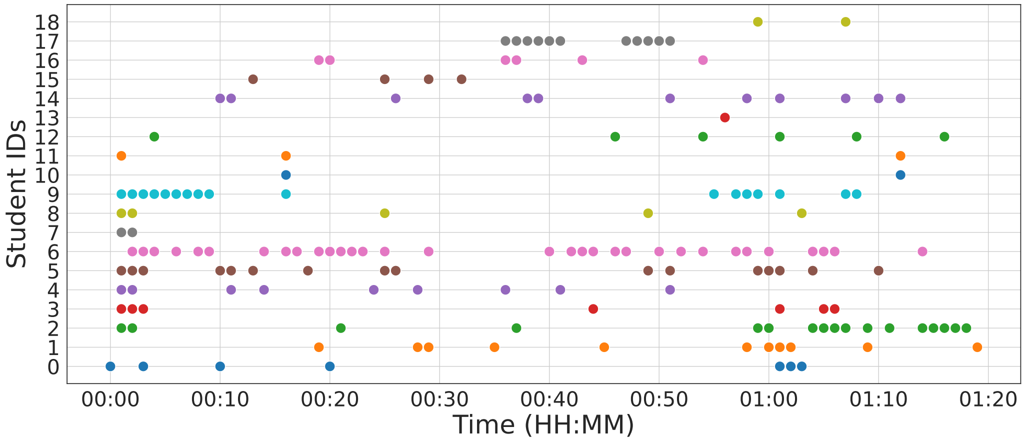


Fig. 11. Active participation across all students in a classroom session over time. Each dot represents the interval (in minutes) during which a particular student (on the y-axis) was actively participating in the class. Variations between the level of participation, high (ID:2,4,6,9,17) *vs.* low (ID:1,3,8,10,11) can be observed. For students with a high level of participation, variations between sporadic participation (ID:2,4,9,17) *vs.* sustained participation (ID:6) can be observed.

session enabled by persistent within-session tracking. Student-level data reveals how participation varies across students, which could assist in powering analytics, including the level of participation (high vs. low) per individual and consistency over time (sporadic vs. sustained). For this session, it can be observed that certain students (ID:2,4,6,9,17) exhibit higher participation than others (ID:1,3,8,10,11), suggesting they actively engage more frequently. Further, students with high levels of participation can be categorized as sporadic (ID:2,4,9,17), with few periods of very high engagement, or consistent (ID:6), with sustained participation over time. This sporadic behavior may indicate diverse class activities occurring, which may lead to fluctuating engagement. Additionally, certain periods in the session promote broad participation (00:50-01:00 engages 12-13 students) while others are more narrow (01:10-01:20 engages 7-8 students). This data could supplement a PD professional consultation with an instructor or serve as a way to observe classroom engagement according to a classroom observation protocol. A PD professional or an automated system that supports instructors in comparing these segments could reveal activity types and instructional techniques that drive equitable participation across more students.

6.2 Student-level Participation across Multiple Sessions

Our first case study showed how our system could enable a richer analysis of individual participation in a single class session. Seeing class engagement across course sessions can provide more information about the persistence of student engagement over time. Prior work in tracking student engagement, especially across class sessions, often takes place in online learning environments where engagement can be measured through log data or online interactions [51]. For in-person classes, longitudinal measurement of engagement often relies on observation or self-reporting [4], which can be unreliable and inaccurate. In this case study, we show how our system can track student participation across class sessions. Figure 12 illustrates student-level participation distribution across multiple sessions. Each data point represents an individual's duration of active participation for that session. While aggregate data only quantifies total participation, this finer-grained perspective reveals nuanced variations. For instance, analytics could highlight that Session 1 exhibits overall lower engagement. Sessions 2, 3, 4, and 5 have equal aggregate participation, but Sessions 2 and 3 derive from one highly engaged student, while

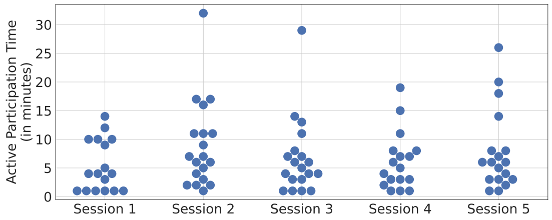


Fig. 12. Distribution of student active participation over multiple sessions. Each dot represents a single student's total active participation for a given session. A higher points' crowding in the bottom for Session 1 suggests low engagement when compared with other sessions. No qualitative difference is present between Sessions 2, 3, 4, and 5 from the aggregate level, but individual-level distribution reveals more equitable participation for Sessions 4 and 5 as compared to Sessions 2 and 3.

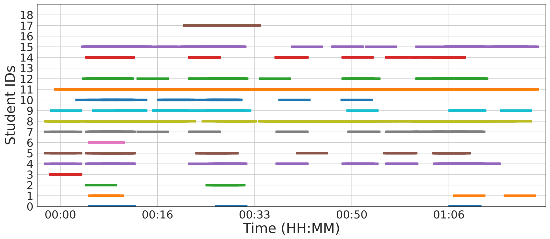
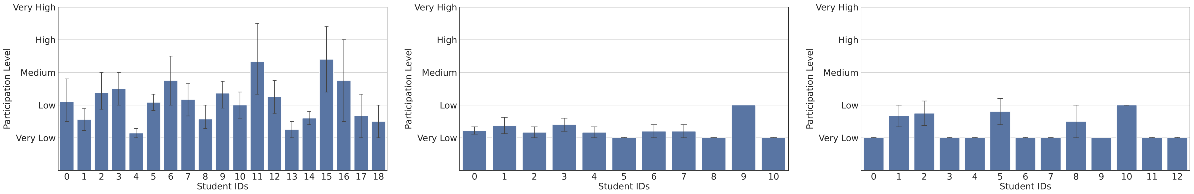


Fig. 13. Periods of self-focus (based on downward gaze) across all students in a classroom session over time. Extended duration of self-focused work(ID: 11, 8, 17, 15) suggest low engagement. In comparison, short spurts of self-focused work (ID: 4, 9, 14) suggest thinking about content while consistently engaging overall.



(a) Course A

(b) Course B

(c) Course C

Fig. 14. Comparing active participation for students across multiple sessions in Course A, Course B, and Course C respectively. Each bar plot shows the average participation for all individuals across different courses based on our cross-session student ID matching.

Sessions 4 and 5 reflect more individuals participating moderately. Interfaces that promote comparison of these sessions could suggest that activities and techniques in 4 and 5 better promote broad engagement, which would be invisible in aggregate data.

In another example, Figure 14 visualizes average student active engagement with an 80% confidence interval for cross-session id-matching results. This visualization becomes viable only through cross-session individual matching. This engagement distribution can provide insights into what actions instructors can take for more equitable participation. For example, Course A (Figure 14a) shows overall moderate student engagement but varied engagement between students. In this instance, an instructor is likely using active learning strategies already but may use other strategies such as moving around more in the classroom, using forms of participation that are less reliant on hand-raising or speaking up, or cold-calling to try to achieve more equitable participation. Separately, Course B (Figure 14b) and Course C (Figure 14c) show low participation levels for all students with lower amounts of variability. Instructors for these courses might instead engage in more active learning strategies

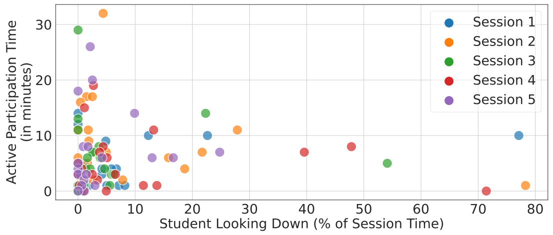


Fig. 15. Summarizing interaction between two engagement proxies across multiple sessions in the same course. Sessions 2 and 4 show the broader distribution of students doing self-focused work, with both high and low active participation, whereas Sessions 1 and 5 show most students spending little time on self-focused work.

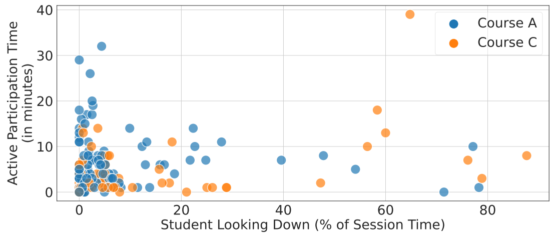


Fig. 16. Summarizing interaction between two engagement proxies across multiple courses. Course A has a higher level of active participation when compared with Course C, whereas Course C has a higher spread of students doing focused work when compared with Course A.

in general. Though noise and inaccuracies might exist, our system can provide instructors with data about the distribution of student engagement and participation.

6.3 Student-level Gaze-based Attention Patterns in a Single Session

Several systems measure gaze as a proxy for student engagement. For example, Glancee [44] detects student gaze in online courses. Bidwell and Fuchs [8] also explore the use of gaze in classroom analytics to determine students' engagement behaviors. our system could integrate with these types of systems to provide nuanced data about individual students' gaze behaviors. Figure 13 illustrates when individual students are engaged in self-focused work within a session, enabled by persistent tracking of gaze features. Extended duration of self-focused work (ID: 11, 8, 17, 15) might be indicative of low engagement, particularly when this is limited to a small number of individuals [86]. Short spurts of self-focused work (ID: 4, 9, 14) likely represent thinking about content while remaining engaged overall. Segments where most students briefly self-focus (05:00-05:30) reveal class-wide reflection, which might be mistaken as poor engagement in aggregate data. Patterns of consistent disengagement (ID: 11, 8) might suggest the need for changes in instructors' teaching strategies in order to engage all students. Interestingly, these same students show low active participation in Figure 11. Thus, individual gaze patterns across students show subtle insights into student attention, which can assist in the accurate interpretation of classroom events versus relying solely on aggregate data. Analytics that use this information to demonstrate sustained downward gaze could provide finer-grained visibility into disengaged students over time, suggesting the need for revised engagement strategies in future sessions.

6.4 Student-Level Interaction Between Different Engagement Modalities

Finally, we introduce a more complex case that provides a starting point for learning analytics engines to display more fine-grained data than solely aggregate data. Figures 15 and 16 show the interaction between two proxies for student engagement (active participation and gaze). In these figures, each data point represents an individual student's pattern seen in a single session (Figure 15), or averaged across all sessions in that course (Figure 16).

While these graphs themselves would be unlikely to help an instructor, an analytics engine could use this data as a starting point for creating interpretable data visualizations and providing detailed feedback to instructors. As seen in Figure 15, Sessions 1 and 5 show visual engagement with the instructor from almost all students. In Sessions 2 and 4, on the other hand, fewer students are engaged, which could lead to feedback or reflection about the content of each session. Importantly, however, not everyone in Session 2 or 4 was self-focused; this indicates that the class content was not an exam or another type of individual work session. Class activity (such as exam vs. lecture) is an important feature for analytics engines to detect. For instance, if we had observed a session in which most students were mostly self-focused, an analytics engine that detects activity type could classify that session as a non-discussion-based (or exam) activity and remove that session from feedback given on engagement.

Looking across courses, in Figure 16, we observe that Course A has a higher level of active participation, and most students focus on instructor activities rather than self-work, whereas in Course C, we see a wider spread of students engaged in self-work and lower participation across the board. A PD professional or teaching observer might use this data to explore whether Course A has better strategies for overall student engagement than Course C. This data could integrate with multimodal analytics systems such as Edusense [1] to provide more nuanced information about instructor and student behaviors.

7 DISCUSSION AND FUTURE WORK

Understanding individual student behaviors is critical for improving engagement and outcomes but has faced technology barriers [56, 85]. By enabling persistent tracking, this work unlocks novel longitudinal insights that are impossible with only summarized aggregate data. Our approach helps address the need for individual-level measurements emerging across observation protocols [35, 63, 71], and can provide utility for various educational stakeholders. In this section, we discuss the usability of ClassID for various stakeholders, potential extensions to other learning environments, and privacy considerations while deploying the system for recording classroom data. Finally, we discuss some of the limitations of our system and further advances needed for even better student behavior attribution in more generalized settings.

7.1 Utility for Educational Stakeholders

Instructors & Consultants: With these analytics, a teaching professional development consultant observes and reviews an instructor's data with them without needing to be physically present in the class. These fine-grained analytics can help instructors receive feedback on their teaching and develop data-driven teaching strategies. For example, instructors might see that the same five students never raise their hands but are active in group work. This might lead to the instructor incorporating more peer activity within the class.

Pedagogy Researchers: While current techniques focus analysis on understanding student behavior, associating teacher, peer, and whole-class events could substantively enrich passively quantified observations. For instance, explicitly linking student attention levels with specific content delivery modes and transitory activities may reveal activation triggers personalized to teaching style and topic. Sessions focusing on learning mathematical concepts could correlate engagement rise with the shift from lecture to collaborative problem demonstration on the whiteboard. Similarly, mapping speech and movement profiles against in-class assessments could expose optimal evaluation methods for fostering participation.

Classroom Sensing Systems: At scale, emergent experiential patterns within and across cohorts could also inform instructor development. These analytics could integrate into existing systems that measure student engagement behaviors, such as Sync Class, in which student engagement is measured by how many students are behaviorally in sync at a given point [27]. Another example is from Giannokos et al. [32], which used wearable devices and video to track student learning. Our system could supplement those additional devices by providing the same granularity of data with video alone. Ultimately, combining student behavior association with precise instructor logs also promotes the development of specialized multimodal solutions from instruments like time-synced sensor rigs and speech analysis pipelines.

Administrators: Administrators can also utilize engagement distributions across students to inform space design and optimization, potentially leading to the design of better learning spaces. Broader future directions include large-scale deployment to study emerging behavioral patterns across individuals related to timing, curriculum, and other variables. Longitudinal tracking can reveal how engagement and learning evolve across the semester and courses.

Overall, persistent student analytics creates opportunities to enhance instruction, participation, and equitable outcomes. This establishes an important foundation for ambient sensing to unlock both individual and collective insights to shape the next generation of responsive, learner-centered education.

7.2 Extending to Diverse Learning Environments

Classroom Types and Scales: While evaluated in small classrooms, demonstrating robustness across formats is key for real-world adoption. We initially focused on smaller groups enabling rich interaction analytics impossible in large impersonal lectures. However, broad applicability to lectures, recitations, discussions, and open spaces remains vital through flexible multi-camera tracking. Current constraint to smaller groups maximizes accurate, granular analytics. However, with adequate architectural instrumentation, larger spaces are addressable by methodology. Smaller classes best promote interactivity and ambient sensing insights. Rather than overwhelm giant lectures, cost-efficient hardware could permeate many overlooked intimate rooms otherwise unaddressed – prioritizing pervasive analytics over sheer size. Strategic, high-density camera deployment mitigates scaling challenges like occlusion and visibility across big, uncontrolled indoor expanses. Modern lecture capture systems show even huge classes can be instrumented given resources. In summary, current evaluations intentionally optimize tracking in typical room-scale spaces. However, further assessments across all learning formats are still needed to confirm adaptable multi-camera analytics can reliably expand given careful architectural planning.

Beyond Classrooms: Importantly, the developed multi-camera tracking methodology readily applies to diverse environments exhibiting consistent front-facing views and people over time beyond just classrooms. Examples include labs, libraries, tutoring, and community classrooms. Analyzing behavioral movement patterns can provide admins insights to improve space design and learning experiences. The technique could also unobtrusively track conference session choices, networking, engagement, and more without intrusive devices. Enabled analytics would help organizers enhance diversity and inclusion year-over-year. Further, the approach generalizes to any consistent group gathering - work meetings, community events, social gatherings, and more. Longitudinal quantification of granular interpersonal intricacies presents numerous research opportunities. While classrooms provide an ideal testbed, the core tracking concepts are versatile. Applying such ambient intelligence innovations across contexts represents an exciting frontier with diverse potential benefits. Reliably capturing subtle human dynamics unobtrusively over time provides a powerful tool widely applicable given similar camera perspectives.

7.3 Potential Privacy Concerns and Mitigation Strategies

De-anonymization attacks: Despite anonymized student ID assignments, employing techniques like extracting facial features and gaze estimation for student ID reconciliation and matching might raise legitimate privacy issues that could impede real-world adoption. Educators, students, and institutions may resist usage over apprehensions regarding the potential misuse of sensitive information. One of the primary issues is de-anonymization - using embeddings to reveal student identities either directly or by reverse engineering. Recent work shows the feasibility of such de-anonymization from facial data using inference attacks [26]. In some cases, risks also arise from instructors' ability to de-anonymize based on familiarity, analogous to current classrooms. This is somewhat mitigated by restricting tracking to particular student actions or instrumentation. Overall, while immediate risks seem limited under reasonable assumptions of instructor capability [54] and data handling, techniques like differential privacy [87] and restricted collection represent proactive future steps for robust privacy preservation.

Misuse of behavioral analytics: Without diligent protections, there is a potential for misuse of behavioral analytics. Tracking student attendance, attention levels, class participation, etc. could allow inferences about disabilities, mental health issues, or other sensitive student attributes that they may not wish to disclose. Beyond aggregate statistics, restrict access to any potentially sensitive analytic outputs only to vetted researchers under strict data handling policies. Allowing sensitivity profiling around disabilities, mental health, or other protected characteristics poses another threat. Tightly restricting and vetting analytics access attempts mitigation alongside purpose-limited data collection. However, policy and consent alone cannot prevent all algorithmic bias. Developing robust technical systems for fairness and representation will prove increasingly vital.

Capturing compromising situations: Another risk can come from simply capturing students on video, thus exposing them to compromising situations if adequate protections are not engineered. Access policies for footage review are thus critical, as are restricted collection and camera operation protocols aligned with minimal necessary capture. Redaction techniques can also automatically obscure sensitive video segments. In some cases, recording student work or speech likely entails capturing intellectual property they wish to protect. Allowing review and contesting alongside speech redaction technology enables recourse. Data retention minimization and expiration likewise limit vulnerability windows.

Re-identification at small classes: For small classes, even aggregate statistics could implicitly reveal identifiable individual information. Differential privacy and setting minimum cohort sizes before reporting seek to close this loophole. However, techniques must be validated specifically for small groups. Utilizing student data beyond the local context, such as in external training datasets, must require explicit consent and rights management. Local systems must facilitate compliant data sharing where permissible or legally mandated.

Moving forward, a priority must be advancing privacy-aware learning analytics and ambient intelligence. Technological and policy advances that balance analytical utility with ethical protections are imperative as classrooms progress toward ambient intelligence. Learning enhancement and privacy objectives can and must be met in tandem. However, doing so necessitates continued research innovation and an unwavering focus on respecting student protections to the fullest degree.

7.4 Limitations and Future Work

Technical limitations: While showing promise, several clear system limitations necessitate ongoing work. Within-session tracking remains susceptible to momentary ID swaps when students pass closely, fully occluding one another too briefly to trigger new IDs. Analyzing frame-level representation anomalies or multi-pass assignments comparing sequence extremes could strengthen detection. Cross-session matching also proves challenging with lower resolution inhibiting distinctive representations, though emerging camera and inference advances may close this gap. More broadly, large-scale validation across diverse environments is imperative to confirm adaptable real-world performance given adequate instrumentation. While optimized for small interactive cohorts, applying methodologies in complex spaces could enable campus-wide ambient intelligence. Careful evaluation must determine sufficient camera densities and ideal placement configurations to track numerous participants through uncontrolled occlusion and visibility barriers.

Strengthening privacy protections: Current techniques preclude obvious data abuse, but enhancing protections remains vital. Expanding differential privacy integration, federated on-device learning, and vulnerability assessments could significantly bolster ethics. Exploring restrictive policies and access controls provides complementary standards guidance. Unambiguous consent currently limits some analyses, but accumulated classroom data could fuel semi-supervised learning uniquely suited to these spaces, unlocking less label-reliant capabilities. Human-AI collaboration also holds promise, with instructor expertise overseeing automated analytics.

Equity Considerations: The current data collection methodology operates under the assumption that no personally identifiable information is revealed by students, thus providing assurances that the collected data cannot be used to impact individual students negatively. However, this also means that it is difficult to determine if the system is biased or if it benefits all groups of students equally. Without identifiable information or objective (or self-reported) demographic data, it is challenging to see how the system affects different groups of students differently. To address this, future studies could use anonymous surveys given to a diverse group of students. These studies could also be designed in a more controlled way to help identify any biases in the system. In doing so, researchers could improve the system itself to ensure fairness towards all students, regardless of their background, which is important to create an inclusive and equitable learning environment.

Future Work: One promising direction of future work can be towards data-centric approaches, such as Trackformer [47], which could enable the development of models specifically tailored to the classroom environment. By learning from classroom-specific challenges, such as occlusion, these models could potentially lead to improved ID assignment results. Additionally, future work could explore the use of finer ground truth annotations, not only to minimize incorrect assignments but also to provide deeper insights into the nature and causes of inaccuracies. Moreover, our research could be extended to incorporate a more detailed analysis of emerging student engagement patterns, investigating the interplay between individual behaviors and their collective impact on overall classroom dynamics. As camera technology advances, future studies could also integrate a broader range of engagement metrics to gain a more comprehensive understanding of student participation behaviors. It is important to note that our work focuses on capturing patterns among regular students in typical lecture scenarios, and its findings may not be generalizable to systematically different populations, such as younger children, students with disabilities, or those in non-traditional educational settings. Future research should aim to validate and expand upon these findings across diverse student populations and learning contexts to ensure the development of equitable and inclusive educational technologies.

8 CONCLUSION

This work presents a novel approach for persistent student identification within and across classroom sessions without any student instrumentation. We used off-the-shelf deep learning methods, and combined them with a series of reconciliation techniques to assign consistent IDs within sessions, overcoming errors from occlusions and movements. We present a cross-session ID matching heuristic that leverages session-level behavioral representations to re-identify students. Our evaluation results demonstrate state-of-the-art performance. Within-session tracking achieves 95% assignment accuracy with 98% detection over multiple sessions, with significantly less erroneous IDs when compared to prior pose-based methods (3 *vs.* 167 on average). Cross-session matching outperforms top re-identification algorithms (Precision: 85% *vs.* 44% on average). Detailed use cases highlight how individual attribution of behaviors can provide more granular insights into student behaviors than aggregate data as well as how these analytics can supplement class observation practices and existing multimodal learning analytics systems. Overall, this contributes an important technical foundation for ambient classroom intelligence systems to scale existing classroom observation practices and extend current multimodal learning analytics systems.

ACKNOWLEDGMENTS

We gratefully acknowledge the generous support of the James S. McDonnell Foundation and the National Science Foundation (NSF awards SaTC-1801472, IIS-1822813, IIS-1747997, CSR-1526237, and STEM Education Postdoctoral Fellowship #2222530). We also thank Carnegie Mellon University's CyLab Security and Privacy Institute and JP Morgan Chase for their support on smart buildings research. We would like to thank Prof. Chris Harrison's, and other original contributors of Edusense [1] for guidance during the ideation process and development support for initial versions of the ClassID system. We acknowledge Yash Govilkar for his contributions to building early versions of the system. We are grateful to Franceska Xhakaj and all the students and instructors who participated in the data collection. Finally, we thank the anonymous reviewers for their constructive feedback on our paper.

REFERENCES

[1] Karan Ahuja, Dohyun Kim, Franceska Xhakaj, Virag Varga, Anne Xie, Stanley Zhang, Jay Eric Townsend, Chris Harrison, Amy Ogan, and Yuvraj Agarwal. 2019. EduSense: Practical Classroom Sensing at Scale. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 71 (Sept. 2019), 26 pages. https://doi.org/10.1145/3351229

[2] Karan Ahuja, Deval Shah, Sujeath Pareddy, Franceska Xhakaj, Amy Ogan, Yuvraj Agarwal, and Chris Harrison. 2021. Classroom digital twins with instrumentation-free gaze tracking. In *Proceedings of the 2021 chi conference on human factors in computing systems*. 1–9.

[3] Islam Alkabbany, Asem M. Ali, Chris Foreman, Thomas Tretter, Nicholas Hindy, and Aly Farag. 2023. An Experimental Platform for Real-Time Students Engagement Measurements from Video in STEM Classrooms. *Sensors* 23, 3 (Feb 2023), 1614. https://doi.org/10.3390/s23031614

[4] Janis F. Andersen, Peter A. Andersen, and Arthur D. Jensen. 2009. The measurement of nonverbal immediacy. *Journal of applied communication research* 7, 2 (nov 2009), 153–180. https://doi.org/10.1080/00909887909365204

[5] Saira Anwar and Muhsin Menekse. 2021. A systematic review of observation protocols used in postsecondary STEM classrooms. *Review of Education* 9, 1 (2021), 81–120. https://doi.org/10.1002/rev3.3235 arXiv:https://bera-journals.onlinelibrary.wiley.com/doi/pdf/10.1002/rev3.3235

[6] Saira Anwar and Muhsin Menekse. 2021. A systematic review of observation protocols used in postsecondary STEM classrooms. *Review of Education* 9, 1 (2021), 81–120.

[7] Irina Arhipova, Gatis Vitols, and Inga Meirane. 2020. Long Period Re-identification Approach to Improving the Quality of Education: A Preliminary Study. In *Future of Information and Communication Conference*. Springer, 157–168.

[8] Jonathan Bidwell and Henry Fuchs. 2011. Classroom analytics: Measuring student engagement with automated gaze tracking. *Behav Res Methods* 49, 113 (2011).

[9] Anja J Boevé, Rob R Meijer, Roel J Bosker, Jorien Vugteveen, Rink Hoekstra, and Casper J Albers. 2017. Implementing the flipped classroom: an exploration of study behaviour and student performance. *Higher Education* 74 (2017), 1015–1032.

[10] Nigel Bosch, Sidney K D'mello, Jaclyn Ocumpaugh, Ryan S Baker, Valerie Shute, and ; S D'mello. 2016. Using video to automatically detect learner affect in computer-enabled classrooms. *ACM Transactions on Interactive Intelligent Systems* 6, 2 (2016), 17. https://doi.org/10.1145/2946837

[11] Babette Bühler, Ruikun Hou, Efe Bozkir, Patricia Goldberg, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. 2023. Automated Hand-Raising Detection in Classroom Videos: A View-Invariant and Occlusion-Robust Machine Learning Approach. In *International Conference on Artificial Intelligence in Education*. Springer, 102–113.

[12] Sean Anthony Byrne, Nora Castner, Ard Kastrati, Martyna Beata Płomecka, William Schaefer, Enkelejda Kasneci, and Zoya Bylinskii. 2023. Leveraging Eye Tracking in Digital Classrooms: A Step Towards Multimodal Model for Learning Assistance. In *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications* (Tubingen, Germany) *(ETRA '23)*. Association for Computing Machinery, New York, NY, USA, Article 80, 6 pages. https://doi.org/10.1145/3588015.3589197

[13] Sunghwan Byun, Niral Shah, and Daniel Reinholz. 2023. When Only White Students Talk: EQUIP-ing Prospective Teachers to Notice Inequitable Participation. *Mathematics Teacher Educator* 11, 3 (2023), 155–168.

[14] Ricardo Böheim, Tim Urdan, Maximilian Knogler, and Tina Seidel. 2020. Student hand-raising as an indicator of behavioral engagement and its role in classroom learning. *Contemporary Educational Psychology* 62 (2020), 101894. https://doi.org/10.1016/j.cedpsych.2020.101894

[15] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. 2023. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9686–9696.

[16] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. 2018. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 67–74.

[17] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).

[18] Stephane Champely, Claus Ekstrom, Peter Dalgaard, Jeffrey Gill, Stephan Weibelzahl, Aditya Anandkumar, Clay Ford, Robert Volcic, and Helios De Rosario. 2017. pwr: Basic functions for power analysis. (2017).

[19] Badong Chen, Lujuan Dang, Nanning Zheng, and Jose C Principe. 2023. Kalman filtering. In *Kalman Filtering Under Information Theoretic Criteria*. Springer, 11–51.

[20] MMCV Contributors. 2018. MMCV: OpenMMLab Computer Vision Foundation. https://github.com/open-mmlab/mmcv.

[21] MMTracking Contributors. 2020. MMTracking: OpenMMLab video perception toolbox and benchmark. https://github.com/open-mmlab/mmtracking.

[22] Zuozhuo Dai, Guangyuan Wang, Weihao Yuan, Siyu Zhu, and Ping Tan. 2022. Cluster contrast for unsupervised person re-identification. In *Proceedings of the Asian Conference on Computer Vision*. 1142–1160.

[23] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. 2019. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641* (2019).

[24] Tim Esler. 2020. Face Recognition Using Pytorch. https://github.com/timesler/facenet-pytorch.

[25] Joan Esson, Paul Wendel, Anna Young, Meredith Frey, and Kathryn Plank. 2022. Recent developments in classroom observation protocols for undergraduate STEM. *Journal of College Science Teaching* 52, 1 (2022), 72–82.

[26] István Fábián and Gábor György Gulyás. 2020. De-anonymizing facial recognition embeddings. *Infocommunications Journal* 12, 2 (2020), 50–56.

[27] Katsuya Fujii, Plivelic Marian, Dav Clark, Yoshi Okamoto, and Jun Rekimoto. 2018. Sync class: Visualization system for in-class student synchronization. In *Proceedings of the 9th augmented human international conference*. 1–8.

[28] Fei Gao, Changxin Cai, Ruohui Jia, and Xinzhong Hu. 2023. Improved YOLOX for pedestrian detection in crowded scenes. *Journal of Real-Time Image Processing* 20, 2 (2023), 24.

[29] Nan Gao, Mohammad Saiedur Rahaman, Wei Shao, Kaixin Ji, and Flora D. Salim. 2022. Individual and Group-wise Classroom Seating Experience: Effects on Student Engagement in Different Courses. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3, Article 115 (sep 2022), 23 pages. https://doi.org/10.1145/3550335

[30] GaoNan, ShaoWei, RahamanMohammad Saiedur, and SalimFlora D. 2020. n-Gage: Predicting In-Class Emotional, Behavioural, and Cognitive Engagement in the Wild. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (sep 2020). https://doi.org/10.1145/3411813

[31] Niloofar Gheissari, Thomas B Sebastian, and Richard Hartley. 2006. Person reidentification using spatiotemporal appearance. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, Vol. 2. IEEE, 1528–1535.

[32] Michail N Giannakos, Kshitij Sharma, Ilias O Pappas, Vassilis Kostakos, and Eduardo Velloso. 2019. Multimodal data as a means to understand the learning experience. *International Journal of Information Management* 48 (2019), 108–119.

[33] Drew Gitomer, Courtney Bell, Yi Qi, Daniel McCaffrey, Bridget K Hamre, and Robert C Pianta. 2014. The instructional challenge in improving teaching quality: Lessons from a classroom observation protocol. *Teachers College Record* 116, 6 (2014), 1–32.

[34] Patricia Goldberg, Ömer Sümer, Kathleen Stürmer, Wolfgang Wagner, Richard Göllner, Peter Gerjets, Enkelejda Kasneci, and Ulrich Trautwein. 2021. Attentive or not?: Toward a machine learning approach to assessing students' visible engagement in classroom instruction. *Educational Psychology Review* 33, 1 (2021).

[35] Alene H Harris and Monica Farmer Cox. 2003. Developing an observation system to capture instructional differences in engineering classrooms. *Journal of Engineering Education* 92, 4 (2003), 329–336.

[36] Yukun Huang, Zheng-Jun Zha, Xueyang Fu, and Wei Zhang. 2019. Illumination-invariant person re-identification. In *Proceedings of the 27th ACM international conference on multimedia*. 365–373.

[37] Paul Hur and Nigel Bosch. 2022. Tracking individuals in classroom videos via post-processing OpenPose data. In *LAK22: 12th International Learning Analytics and Knowledge Conference*. 465–471.

[38] Srikrishna Karanam, Yang Li, and Richard J Radke. 2015. Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *Proceedings of the IEEE international conference on computer vision*. 4516–4524.

[39] Naz Kaya and Brigitte Burgess. 2007. Territoriality: Seat preferences in different types of classroom arrangements. *Environment and Behavior* 39, 6 (2007), 859–876.

[40] Sara Khosravi, Stuart G Bailey, Hadi Parvizi, and Rami Ghannam. 2022. Wearable sensors for learning enhancement in higher education. *Sensors* 22, 19 (2022), 7633.

[41] Xinyu Li, Lixiang Yan, Linxuan Zhao, Roberto Martinez-Maldonado, and Dragan Gasevic. 2023. CVPE: A Computer Vision Approach for Scalable and Privacy-Preserving Socio-spatial, Multimodal Learning Analytics. In *LAK23: 13th International Learning Analytics and Knowledge Conference*. 175–185.

[42] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 740–755.

[43] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. MediaPipe: A Framework for Building Perception Pipelines. arXiv:1906.08172 [cs.DC]

[44] Shuai Ma, Taichang Zhou, Fei Nie, and Xiaojuan Ma. 2022. Glancee: An adaptable system for instructors to grasp student learning status in synchronous online classes. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–25.

[45] Madhu Mahalingam, Fred Schaefer, and Elisabeth Morlino. 2008. Promoting student learning through group problem solving in general chemistry recitations. *Journal of chemical education* 85, 11 (2008), 1577.

[46] Roberto Martinez-Maldonado, Vanessa Echeverria, Jurgen Schulte, Antonette Shibani, Katerina Mangaroska, and Simon Buckingham Shum. 2020. Moodoo: Indoor Positioning Analytics for Characterising Classroom Teaching. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12163 LNAI (jul 2020), 360–373. https://doi.org/10.1007/978-3-030-52237-7_29

[47] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixé, and Christoph Feichtenhofer. 2022. TrackFormer: Multi-Object Tracking With Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8844–8854.

[48] Shervin Minaee, Ping Luo, Zhe Lin, and Kevin Bowyer. 2021. Going Deeper Into Face Detection: A Survey. arXiv:2103.14983 [cs.CV]

[49] Jennalyn N. Mindoro, Niño U. Pilueta, Yolanda D. Austria, Luisito Lolong Lacatan, and Rhowel M. Dellosa. 2020. Capturing Students' Attention Through Visible Behavior: A Prediction Utilizing YOLOv3 Approach. In *2020 11th IEEE Control and System Graduate Research Colloquium (ICSGRC)*. 328–333. https://doi.org/10.1109/ICSGRC49013.2020.9232659

[50] MMPose. 2020. OpenMMLab Pose Estimation Toolbox and Benchmark. https://github.com/open-mmlab/mmpose.

[51] Libby V Morris, Catherine Finnegan, and Sz-Shyan Wu. 2005. Tracking student behavior, persistence, and achievement in online courses. *The Internet and Higher Education* 8, 3 (2005), 221–231.

[52] Catherine Mulryan-Kyne. 2010. Teaching large classes at college and university level: Challenges and opportunities. *Teaching in higher Education* 15, 2 (2010), 175–185.

[53] Nivedha Murugesan, Irene Cho, and Cristina Tortora. 2021. Benchmarking in cluster analysis: a study on spectral clustering, DBSCAN, and K-Means. In *Data Analysis and Rationality in a Complex World 16*. Springer, 175–185.

[54] Tricia J Ngoon, David Kovalev, Prasoon Patidar, Chris Harrison, Yuvraj Agarwal, John Zimmerman, and Amy Ogan. 2023. " An Instructor is [already] able to keep track of 30 students": Students' Perceptions of Smart Classrooms for Improving Teaching & Their Emergent Understandings of Teaching and Learning. (2023).

[55] Eilidh Noyes and Rob Jenkins. 2017. Camera-to-subject distance affects face configuration and perceived identity. *Cognition* 165 (2017), 97–104.

[56] Sharon Oviatt. 2018. Ten opportunities and challenges for advancing student-centered multimodal learning analytics. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 87–94.

[57] Jiangmiao Pang, Linlu Qiu, Xia Li, Haofeng Chen, Qi Li, Trevor Darrell, and Fisher Yu. 2021. Quasi-dense similarity learning for multiple object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 164–173.

[58] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

[59] Mark Pilgrim. 2009. Serializing Python Objects. In *Dive Into Python 3*. Springer, 205–223.

[60] Prasoon Patidar, Tricia Ngoon, John Zimmerman, Amy Ogan, Yuvraj Agarwal. 2024. ClassID: Open-source repository for ClassID. https://github.com/edusense/ClassID.

[61] DL Reinholz. 2021. Equity and equality: Data visualizations as mediating artifacts for teacher sensemaking about racial and gender inequity. *Contemporary issues in technology and teacher education* 21, 3 (2021).

[62] Daniel L Reinholz, Kevin Pelaez, and Niral Shah. 2021. Capturing who participates and how: the stability of classroom observations using EQUIP. *SN Social Sciences* 1 (2021), 1–18.

[63] Daniel L Reinholz and Niral Shah. 2018. Equity analytics: A methodological approach for quantifying participation patterns in mathematics classroom discourse. *Journal for Research in Mathematics Education* 49, 2 (2018), 140–177.

[64] Ergys Ristani, Francesco Solera, Roger S. Zou, R. Cucchiara, and Carlo Tomasi. 2016. Performance Measures and a Data Set for Multi-target, Multi-camera Tracking. In *ECCV Workshops*.

[65] David Rosengrant, Doug Hearrington, Kerriann Alvarado, and Danielle Keeble. 2012. Following student gaze patterns in physical science lectures. *AIP Conference Proceedings* 1413, 1 (02 2012), 323–326. https://doi.org/10.1063/1.3680060 arXiv:https://pubs.aip.org/aip/acp/article-pdf/1413/1/323/12210821/323_1_online.pdf

[66] Mukul Lata Roy, D Malathi, and JD Dorathi Jayaseeli. 2022. Facial Recognition Techniques and Their Applicability to Student Concentration Assessment: A Survey. In *Proceedings of International Conference on Deep Learning, Computing and Intelligence: ICDCI 2021*. Springer, 213–225.

[67] M Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. 2018. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 420–429.

[68] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. https://doi.org/10.1109/cvpr.2015.7298682

[69] Erich Schubert, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. 2017. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS)* 42, 3 (2017), 1–21.

[70] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. 2018. CrowdHuman: A Benchmark for Detecting Human in a Crowd. *arXiv preprint arXiv:1805.00123* (2018).

[71] Prateek Shekhar, Matt Demonbrun, Maura Borrego, Cynthia Finelli, Michael Prince, Charles Henderson, and Cynthia Waters. 2015. Development of an observation protocol to study undergraduate engineering student resistance to active learning. *International Journal of Engineering Education* 31, 2 (2015), 597–609.

[72] Tripti Singh, Mohan Mohadikar, Shilpa Gite, Shruti Patil, Biswajeet Pradhan, and Abdullah Alamri. 2021. Attention Span Prediction Using Head-Pose Estimation With Deep Neural Networks. *IEEE Access* 9 (2021), 142632–142643. https://doi.org/10.1109/ACCESS.2021.3120098

[73] Michelle K Smith, Francis HM Jones, Sarah L Gilbert, and Carl E Wieman. 2013. The Classroom Observation Protocol for Undergraduate STEM (COPUS): A new instrument to characterize university STEM classroom practices. *CBE—Life Sciences Education* 12, 4 (2013), 618–627.

[74] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep High-Resolution Representation Learning for Human Pose Estimation. arXiv:1902.09212 [cs.CV]

[75] Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo. 2022. DanceTrack: Multi-Object Tracking in Uniform Appearance and Diverse Motion. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20961–20970. https://doi.org/10.1109/CVPR52688.2022.02032

[76] Ömer Sümer, Patricia Goldberg, Sidney D'Mello, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. 2023. Multimodal Engagement Analysis From Facial Videos in the Classroom. *IEEE Transactions on Affective Computing* 14, 2 (2023), 1012–1027. https://doi.org/10.1109/TAFFC.2021.3127692

[77] Kimberly D Tanner. 2013. Structure matters: twenty-one teaching strategies to promote student engagement and cultivate classroom equity. *CBE—Life Sciences Education* 12, 3 (2013), 322–331.

[78] Chinchu Thomas and Dinesh Babu Jayagopi. 2017. Predicting student engagement in classrooms using facial behavioral cues. In *Proceedings of the 1st ACM SIGCHI International Workshop on Multimodal Interaction for Education* (Glasgow, UK) *(MIE 2017)*. Association for Computing Machinery, New York, NY, USA, 33–40. https://doi.org/10.1145/3139513.3139514

[79] Lisa Tsui and Eleanor Gao. 2006. The efficacy of seminar courses. *Journal of College Student Retention: Research, Theory & Practice* 8, 2 (2006), 149–170.

[80] Marcella Ucci, Stephen Law, Richard Andrews, Abi Fisher, Lee Smith, Alexia Sawyer, and Alexi Marmot. 2015. Indoor school environments, physical activity, sitting behaviour and pedagogy: a scoping review. *Building Research & Information* 43, 5 (2015), 566–581. https://doi.org/10.1080/09613218.2015.1004275 arXiv:https://doi.org/10.1080/09613218.2015.1004275

[81] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. 2019. MOTS: Multi-Object Tracking and Segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7934–7943. https://doi.org/10.1109/CVPR.2019.00813

[82] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. 2018. Person Transfer GAN to Bridge Domain Gap for Person Re-Identification. arXiv:1711.08565 [cs.CV]

[83] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. 2017. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*. 3645–3649. https://doi.org/10.1109/ICIP.2017.8296962

[84] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. 2017. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*. 3645–3649. https://doi.org/10.1109/ICIP.2017.8296962

[85] Marcelo Worsley. 2018. Multimodal learning analytics' past, present, and, potential futures. *CEUR Workshop Proceedings* 2163 (2018). Publisher Copyright: © 2018 CEUR-WS. All Rights Reserved.; 2nd Multimodal Learning Analytics Across (Physical and Digital) Spaces, CrossMMLA 2018 ; Conference date: 06-03-2018.

[86] Hui Xu, Junjie Zhang, Hui Sun, Miao Qi, and Jun Kong. 2023. Analyzing students' attention by gaze tracking and object detection in classroom teaching. *Data Technologies and Applications* 57, 5 (2023), 643–667.

[87] Zheng Xu, Maxwell Collins, Yuxiao Wang, Liviu Panait, Sewoong Oh, Sean Augenstein, Ting Liu, Florian Schroff, and H. Brendan McMahan. 2023. Learning to Generate Image Embeddings with User-level Differential Privacy. arXiv:2211.10844 [cs.LG]

[88] Mang Ye, Jiawei Li, Andy J Ma, Liang Zheng, and Pong C Yuen. 2019. Dynamic graph co-matching for unsupervised video-based person re-identification. *IEEE Transactions on Image Processing* 28, 6 (2019), 2976–2990.

[89] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. 2021. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence* 44, 6 (2021), 2872–2893.

[90] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. 2014. Learning Face Representation from Scratch. arXiv:1411.7923 [cs.CV]

[91] Xin Zhang, Cheng-Wei Wu, Philippe Fournier-Viger, Lan-Da Van, and Yu-Chee Tseng. 2017. Analyzing students' attention in class using wearable devices. In *2017 IEEE 18th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM)*. 1–9. https://doi.org/10.1109/WoWMoM.2017.7974306

[92] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. 2021. ByteTrack: Multi-Object Tracking by Associating Every Detection Box. *arXiv preprint arXiv:2110.06864*.

[93] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable Person Re-identification: A Benchmark. In *Computer Vision, IEEE International Conference on*.

[94] Huayi Zhou, Fei Jiang, Jiaxin Si, Lili Xiong, and Hongtao Lu. 2023. StuArt: Individualized Classroom Observation of Students with Automatic Behavior Recognition and Tracking. arXiv:2211.03127 [cs.HC]

[95] Kaiyang Zhou and Tao Xiang. 2019. Torchreid: A Library for Deep Learning Person Re-Identification in Pytorch. *arXiv preprint arXiv:1910.10093* (2019).

[96] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. 2019. Omni-Scale Feature Learning for Person Re-Identification. *CoRR* abs/1905.00953 (2019). arXiv:1905.00953 http://arxiv.org/abs/1905.00953

[97] Ömer Sümer, Patricia Goldberg, Sidney D'Mello, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. 2021. Multimodal Engagement Analysis from Facial Videos in the Classroom. arXiv:2101.04215 [cs.CV]