Classroom Digital Twins with Instrumentation-Free Gaze Tracking

Karan Ahuja*
Carnegie Mellon University
Pittsburgh, PA, USA
kahuja@cs.cmu.edu

Deval Shah*
Carnegie Mellon University
Pittsburgh, PA, USA
devals@andrew.cmu.edu

Sujeath Pareddy
Carnegie Mellon University
Pittsburgh, PA, USA
spareddy@andrew.cmu.edu

Franceska Xhakaj
Carnegie Mellon University
Pittsburgh, PA, USA
francesx@andrew.cmu.edu

Amy Ogan
Carnegie Mellon University
Pittsburgh, PA, USA
aeo@cs.cmu.edu

Yuvraj Agarwal
Carnegie Mellon University
Pittsburgh, PA, USA
yuvraj@cs.cmu.edu

Chris Harrison
Carnegie Mellon University
Pittsburgh, PA, USA
chris.harrison@cs.cmu.edu

ABSTRACT

Classroom sensing is an important and active area of research with great potential to improve instruction. Complementing professional observers – the current best practice – automated pedagogical professional development systems can attend every class and capture fine-grained details of all occupants. One particularly valuable facet to capture is class gaze behavior. For students, certain gaze patterns have been shown to correlate with interest in the material, while for instructors, student-centered gaze patterns have been shown to increase approachability and immediacy. Unfortunately, prior classroom gaze-sensing systems have limited accuracy and often require specialized external or worn sensors. In this work, we developed a new computer-vision-driven system that powers a 3D "digital twin" of the classroom and enables whole-class, 6DOF head gaze vector estimation without instrumenting any of the occupants. We describe our open source implementation, and results from both controlled studies and real-world classroom deployments.

CCS CONCEPTS

• **Human-centered computing → Interactive systems and tools**; *Ubiquitous and mobile computing*.

KEYWORDS

Classroom sensing, gaze tracking, digital twins.

ACM Reference Format:
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 *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3411764.3445711

*Both authors contributed equally to this research.

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05.
https://doi.org/10.1145/3411764.3445711

1 INTRODUCTION

Over the past decades, learning science research has identified many features of successful teacher-student interactions that lead to beneficial outcomes for students such as greater learning, higher self-efficacy, and increased student voice in the classroom. Yet changing one's classroom practice, even with awareness of this research, is not straightforward. For instance, in universities where we situate our work, professors are hired and promoted for their domain expertise, and they typically view themselves as domain experts and not teaching experts [9]. University faculty typically receive no training in instruction; instead, they *learn* how to teach on the job, often without much support [28].

One solution to support such instructors' growth is personalized and regular professional development. Today, this is partially achieved with professional observers, who attend one (or perhaps a few) lectures to observe and subsequently provide formative feedback to instructors. This approach is impossible to scale to every class and every instructor, and yet grounded, regular feedback on ones' current practice is an essential component of learning [14]. Teachers need to routinely reflect on how their practices (mis)align with effective pedagogy in order to change [30]. In short, the instructional feedback loop currently occurs at such large intervals as to have a negligible impact on the quality of higher education.

In response, researchers are investigating AI-augmented pedagogical professional development [3, 21]. Used together with professionals, such systems could support every instructor, attend every class, help instructors observe and reflect on trends across semesters, and capture fine-grained details for all occupants that would be impossible even with a team of in-situ human observers. There are several innovations and components required to achieve this vision, from low-level sensing and secure data storage, all the way to end-user interfaces providing instructors with actionable feedback, e.g., reflection opportunities following class [17, 20, 42, 44]. In this paper, we put forward the idea of a classroom "digital twin" – a concept borrowed from the Internet of Things (IoT) research (sometimes called *mirror models* or *mirror worlds*) [19, 25, 32] - which we believe can serve as an important contextual container for classroom sensor data, on top of which future end-user applications can be built.

More formally, digital twins are a *dynamic virtual representations of a physical system, using real-time data to enable understanding,*

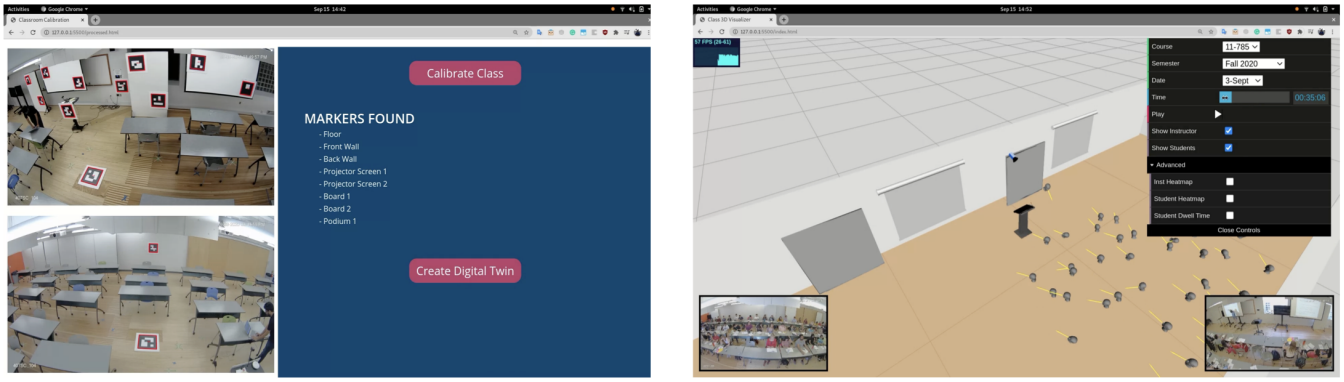

Figure 1: Left: Our web-based capture interface that detects ArUco markers and builds an inventory of important items (walls, whiteboards, etc.). Right: Example digital twin output, also a web-based application. Once a classroom is created, processed gaze data can be loaded into the scene and replayed.

learning and reasoning [35]. This representation can include everything from the precise dimensions of the space, the temperature of a room, the speed of the elevators in a building - essentially anything that can be measured about a physical location and digitized. The concept is akin to a simulation, but employs authentic sensed data from actual physical environments. Importantly, it allows this measured data to be better contextualized in a rich, three-dimensional scene that can be viewed and manipulated in space and time.

A classroom is a great exemplar of such a complex physical environment, which contains objects of various functions (whiteboards, projection screens, podiums, seats, tables) and occupants in at least two different roles. There are strong contextual and spatial relationships between these physical elements that can be (re)played out and analyzed in a digital twin that rows in a database or lines on a chart cannot so easily provide.

In this work, we digitize classrooms and the people and objects within them (Figure 1). Then, as a specific proof-of-concept data source for investigation, we digitize classroom gaze within this room: a feature made richer by being contextualized in a dynamic 3D scene. Gaze from a particular actor in the scene emanates from a source location and lands on a target. It changes rapidly over time and moves dynamically in space. Apart from providing a rich data source for modeling, gaze also provides psychological signals of great importance for both studying and improving classroom teaching.

We are not the first to consider classroom gaze and its utility as a part of professional development for improving instructor-student interactions has been well motivated in prior work (discussed in the next section). However, we are the first to embody it in a 3D classroom digital twin, and furthermore, our six-degree-of-freedom (6-DOF) gaze tracking pipeline outperforms prior systems that track classroom gaze, cutting angular error by roughly half. Together, these dual advances form the technical contribution of our paper, to which we add two evaluations: a controlled study and results from a large-scale deployment in real-world classrooms. We conclude with avenues for future work, as this is very much an early step in a much larger trajectory of supporting instructor professional development via classroom sensing systems.

2 RELATED WORK

We first provide some additional background on digital twins. We then summarize key work that underscores the pedagogical value of gaze sensing in classrooms and review other systems that have captured classroom gaze patterns through a variety of alternative sensing means.

2.1 Digital Twins

Digital twins are generally considered to have emerged in the early 2000s, in parallel with complementary advances in wireless connectivity and the Internet of Things (IoT) [36, 39]. However, the concept has much earlier roots, going back at least to the 1960s, with NASA "twinning" physical systems at ground level to match those in space. This proved invaluable during the Apollo 13 crisis where ground-level twins were used to simulate various on-board events and conditions. Using this, engineers were able to identify problems, replay events, and gain comprehension of complex interdependencies, all of which informed real-world decisions with significant consequences.

Since then, digital twins have been proposed for use in many other complex environments, including factory floors [46], military vehicles [26] and wildlife sanctuaries [33]. As sensor networks and ambitions have grown, there are now efforts to twin whole buildings (e.g. as an approach to increase sustainability), and even countries to serve a variety of purposes [1]. Information that is proposed to be twinned includes spatial layouts of objects in the environment, electrical and other infrastructural maps, sound levels, and features of occupants in spaces. To the best of our knowledge, this concept has not yet been applied to classroom environments. However, other sensing systems, such as using video cameras to capture a 2D classroom scene, have long been a part of teacher professional development approaches such as video-stimulated recall [24].

2.2 Pedagogical Value of Sensing Gaze

As noted above, we test our concept using the particular classroom feature of gaze. For decades, education researchers have understood and investigated the importance of gaze and eye contact in teaching

(also called the visual focus of attention (VFoA) in the literature [49]). For example, direct eye contact can increase closeness and rapport between teachers and students, reducing the psychological distance that the authority structures of the classroom can impose [4, 6]. Teachers who look at their students are perceived as more interested and more approachable [34]. The absence of gaze is just as telling, making the warmest teachers seem cold and distant [5]. A teacher who rarely looks at a student when talking creates the perception that she or he is not very interested in that student [13]. Breed et al. [13] also found that the absence of eye contact between teachers and university students produces negative feelings about the class. Thus, the gaze is an important component in the development of immediacy, a construct that captures this positive sense of warmth and belonging between interlocutors [7]. In addition to immediacy, eye contact permits teachers to monitor and regulate their classes while student gaze can provide a strong signal of attentiveness on their part to the learning material [5].

Beyond self-reported perceptions, teacher gaze also has a direct impact on subsequent student behavior. High levels of gaze cause students to be more attentive to the teacher [13]. Students in high eye contact availability conditions are more likely to participate in class than those in low eye contact availability conditions. In a series of controlled studies, gaze has also been found to increase recall; students were better able to answer questions from verbal presentations of information when the speaker looked at them [43].

Importantly, teachers' abilities in employing effective classroom behaviors such as gaze are not fixed, but can be changed through intervention. For instance, receiving visual warnings alerting them to students not receiving enough eye gaze enabled teachers to spread their attention more equally among students than teachers without augmented perception [10].

Taken together, this extensive literature motivates the importance of gaze in the classroom, from both a research standpoint as well as motivating the need for professional development approaches to improve instructors' pedagogical skills. This provides a basis for its use as an initial feature of our digital classroom twin. Further applications are discussed in the Future Work section.

2.3 Prior Classroom Gaze Systems

There has been a plethora of research in the graphics and computer vision community on gaze estimation. Two main approaches have been explored. The first is to instrument the wearer with a mobile eye tracking headset [16, 42, 45]. These devices are very accurate, but are more invasive (socially, ergonomically and aesthetically) and require many expensive headsets to track all participants. The second approach is to instrument the environment with sensors such as depth cameras [11] and RGB cameras [3, 47]. These approaches are significantly less accurate compared to their wearable counterparts (1 vs. 25 degrees of gaze angular error) but offer a cheap and scalable 6-DOF gaze tracking (3 degrees of freedom of head rotation - yaw, pitch and roll - and 3 degrees of translation with respect to the classroom - X, Y and Z). See [15, 27] for an in-depth survey of gaze estimation systems.

Prior classroom sensing systems have also recognized the utility of gaze in tracking instructor-student interactions [37, 48], behavior analysis [13] and attention tracking [11] to name a few. As they make use of approaches from the gaze literature itself, they can broadly be categorized based on the sensors' placement, type, and fidelity. We compare this prior work in Table 1. Very related to our approach is Bidwell et al. [11], which makes use of 9 cameras placed across the classroom to capture students' gaze and model their attention. However, this system does not track the instructor and the hardware setup is comparatively heavyweight. In contrast, EduSense [3] provides a comparatively lightweight setup (2 off-the-shelf cameras), but does not capture the 3D classroom and only captures the 3-DOF gaze (head rotation) of the students and instructor. Our approach combines the best of both worlds, providing 6-DOF gaze capture of students and instructors, while also capturing the 3D scene, all the while making use of only two cameras.

3 IMPLEMENTATION

Our system is built upon key developments in computer vision and image processing that we utilize to provide a holistic sensing system. We now describe the main components of our approach.

3.1 Hardware

In order to run our system, classrooms must be outfitted with two cameras: one at the front of the room looking towards the students and another looking at the instructor. While more cameras can increase the field of view and sensing fidelity, we settled on two cameras, finding this to be a good balance between deployment practicality (hardware cost, available Ethernet ports, time to deploy, etc.) and classroom coverage. We make use of off-the-shelf Lorex LNE8950AB cameras that have a 112° field of view and cost ~$150 in single-unit retail prices. These use Power over Ethernet (POE) for power and connectivity, making installation simple and clean. These cameras are configured in software to transmit data 4K video at 5 FPS. Our processing backend is an Intel Core i9-7920X CPU running at 2.90GHz with a GeForce GTX 1080 Ti GPU.

3.2 Digital Twin Capture

First, we have to establish our camera's intrinsic parameters (e.g., focal length, distortion coefficients) using a checkboard pattern [29]. For this, we use OpenCV's camera calibration routines [12]. Once calibrated, the intrinsics can be used for all cameras of the same model. These parameters are later used to correct distortions like fish-eye and to estimate 3D distances in real-world units.

Our next step is to detect the location of various objects related to pedagogy in the classroom, which include items such as whiteboards used by the instructor, overhead projection screens, and the podium. The physical size of the classroom is also important, and so we also need to get the reference of the walls and floor of the classroom. To detect these, users place two ArUco [23] markers for each item on diagonal corners to establish their 6-DOF plane (3D position and rotation). We provide a library of pre-defined ArUco markers for different objects, allowing our pipeline to not only localize walls/objects in space, but also know the category. This one-time process takes only a few minutes per classroom. Note that items that have an irregular shape (such as a podium) are approximated as a rectangular plane (and captured using a single ArUco

Table 1: Comparison of our system vs. prior work on classroom gaze sensing.

	Sensor Type	3D Classroom Capture	6-DOF Gaze	Instructor Gaze	Student Gaze	Deployed at scale	Mean Gaze Error
Thomas et al.[47]	RGB Camera	✗	✓	✗	✓	✗	not reported
Cutumisu et al.[16]	Student-Worn Eye Tracker	✗	✗	✗	✓	✗	0.5°
Sumer et al. [45]	Instructor-Worn Eye Tracker	✗	✗	✓	✗	✗	not reported
Raca et al. [42]	Student-Worn Eye Tracker	✗	✗	✗	✓	✗	not reported
Bidwell et al. [11]	5 RGB + 4 Depth Cameras	✓	✓	✗	✓	✗	not reported
Aung et al. [8]	Dataset of Youtube Videos	✗	✗	✓	✓	✗	38.3°
Ahuja et al. [3]	2 RGB cameras	✗	✗	✓	✓	✓	34.6°
Our approach	2 RGB cameras	✓	✓	✓	✓	✓	21.3°

marker). We use OpenCV's [12] ArUco marker detection API to provide us with each marker's 3D pose.

We use the marker placed in the center of the classroom floor as a common origin. Both cameras use this marker to set their own 6-DOF position. Then, in turn, all other markers seen by the cameras can be appropriately located and oriented in space, creating a 3D classroom with walls and objects.

3.3 6-DOF Head Pose Estimation

To infer the direction of gaze, we need to estimate the head pose of each student and the instructor in the class using the two camera views. We start by first detecting all of the faces in the scene using RetinaFace [18], which outputs face bounding boxes. We then run 3DDFA [50] to extract facial landmarks (68 points) that correspond to features like eyes, nose, mouth, jawline, etc. The output of 3DDFA is 2D coordinates in the image space of the classroom and 3D coordinates of landmarks in the object space.

To convert the landmarks from the object coordinate space to the classroom world space, we need to solve for this translation and rotation with respect to our classroom origin. We make use of SolvePnP [22] to find the world position of the 3D face points by solving for its correspondences to the 2D points. At the end of this step, we have the 6-DOF head pose - encoding the 3D rotation (yaw, pitch, and roll) and 3D position - of all the people in the scene. This gives us a head gaze vector with an origin at the center of each head. Figure 4 offers an example scene with head gaze plotted as a 3D frustum.

The next step is to distinguish between the instructor and the students in the scene. We make the assumption that the instructor is the person that is closest to the podium, whiteboard(s), and/or projection screen(s). In the future, more advanced techniques might be employed, including who is standing vs. sitting, who is talking, and facial recognition. Once we identify the instructor, we then track them across frames using a standard centroid-based Euclidean-distance tracker [38].

It is important to reiterate that we estimate the head pose rather than the eye gaze of students and instructors in the classroom. This is because even with 4K cameras, there is insufficient resolution to estimate true eye-gaze even with state-of-the-art techniques at several meters range. Fortunately, prior research [3, 40, 41] has shown that head orientation is a good proxy for gaze attention in classrooms.

3.4 Foci Estimation & Heatmaps

To provide semantically useful information about where the students and the instructions are actually looking, in the final step we combine the 3D gaze with the locations of the walls, floor, and the different objects (projector screen, whiteboard) in the unified 3D classroom. For the students, we find their point of gaze by finding the point of intersection between the gaze direction vector and the 3D planes (such as whiteboard, podiums, etc.) in the digital twin. Upon aggregating these points of gaze over time, we can create a semantically meaningful gaze dwell map for different objects. Lack of dwell on items is also a useful metric. By tracking dwell over time, we can compute percentages of attention (Figure 2, left) and even derive detailed heatmaps (Figure 2, center).

For instructors, there often does not exist a single plane of focus. Instructors move around and look at different areas of the classroom and at different students. We thus find the intersection of the instructor's gaze with different student planes (based on the students' 3D facial bounding boxes). Note that since the instructor gaze can intersect with multiple planes (e.g. students sitting behind one another) it is challenging to positively identify the instructor's true gaze target. Instead, we record all possible intersections along the gaze ray as possible targets and them project this information down onto a 2D heatmap (accumulated over time), which we render on the floor of the classroom (Figure 2 right). In the future, we could rely on cues such as hand gaze or triangulation of active speakers using microphone arrays to held resolve this 3D ambiguity.

3.5 User Interface

We created two proof-of-concept user interfaces to synthesize and explore classroom digital twins, and render gaze data that we processed (see Figure 1). This was a web app created in javascript using the three.js [2] library for 3D rendering. In addition to connecting via a desktop or mobile web browser, we also allow users to enter classroom digital twins via a VR headset. This allows for highly embodied exploration of the 3D space and experience different perspectives (see Video Figure and Figures 2 and 3). We believe this new modality could drive new and interesting opportunities for reflecting on pedagogical practice — an area we hope to explore in future work. We note that the current instantiation of these interfaces is not yet intended for direct use by instructors (i.e., the pedagogical value is currently low), and is chiefly meant for illustrative purposes and to aid us with debugging. Our current research is

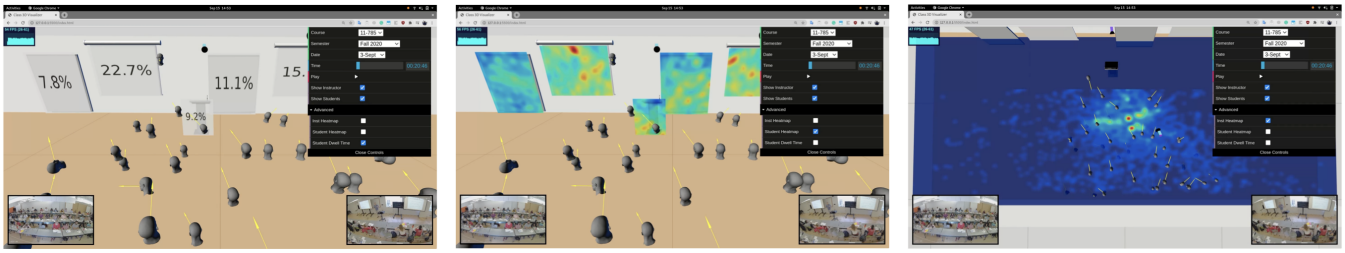

Figure 2: Left: Percentage of student gaze across various classroom foci (whiteboards, projector screens, lectern) at the end of a class session. Center: Heatmaps of students gaze across the same foci. Right: Heatmap of the instructor gaze aggregated across a class session.

meant as a vehicle and important technical stepping stone to future applications, which we discuss in Future Work.

3.6 Privacy Preservation

Even though our study was reviewed and approved by our university's IRB, any system that captures images and video from classrooms naturally evokes potential privacy concerns. Images and videos of students in classrooms, if stored, can lead to concerns for both students and instructors about being tracked. Left unaddressed, these concerns could lead to such systems not being widely adopted by universities. We took these privacy challenges head-on. In particular, similar to the EduSense system [3], we address these privacy challenges by only storing processed data (e.g. facial landmarks/keypoints, facial bounding boxes, location of walls, floors, and objects) and discarding the raw video frames immediately after being processed by our pipeline (in our Video Figure, we include reference footage for illustration). We believe most of the concerns around privacy in classrooms pertain to raw data (audio, images, video) and much less so around processed facial keypoints (which are not tied to any person). Notably, once we process the summarized views, such as the heatmaps of the students and instructor gaze, we can even discard facial keypoints to further alleviate privacy concerns.

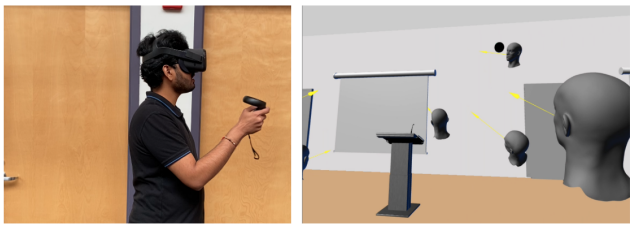

Figure 3: In additional to conventional web browsers, users can enter classroom digital twins via a VR headset, and then move around to replay data from different perspectives, offering an interesting new modality for reflecting on pedagogical practice.

4 OPEN SOURCE MODEL AND DATA

To enable other researchers and practitioners to build upon our system, we have open-sourced the code for our 6-DOF gaze tracking module at https://github.com/edusense/edusense. The code and a sample demo for a classroom digital twin can be found at https://github.com/edusense/ClassroomDigitialTwins.

5 CONTROLLED STUDY

To assess the geometric accuracy of our classroom digital twins and the angular accuracy of our gaze tracking pipeline, we devised a controlled study, which used a series of known targets. This evaluation naturally complements our uncontrolled, in-the-wild study (i.e., real classrooms) discussed later. The latter is more ecologically valid, but because ground-truth gaze angles are unknown, it precludes assessing fine-grained metrics such as angular error. However, taken together, the two studies provide a holistic assessment of our system's feasibility.

5.1 Procedure

We ran the controlled study in an exemplary classroom to test the spatial accuracy of our 6-DOF gaze detection modules. To estimate the gaze accuracy of students we placed 17 AruCo markers in total across the classroom including 4 markers each on two two writing boards and two projectors, and one AruCo marker on the podium. This setup can be seen in Figure 1. We recruited 8 participants (2 female) with a mean age of 26.8 years. We conducted the study across 6 rounds. In each round, each participant chose one of the 21 different seating locations available in the classroom and then looked at each of the 17 markers one by one. This resulted in a total of 6 rounds $\times$ 8 participants $\times$ 17 gaze targets = 816 trails. For the instructor study, we recruited 5 participants (1 female) with a mean age of 24.8 years. 10 AruCo markers were placed to simulate the position of students in the classrooms. Each participant changed their position 5 times to simulate different positions for the instructors resulting in a total of 5 participants $\times$ 5 positions 10 gaze targets = 250 trials.

All participants gave written consent to their data and video being recorded. For each trial, we captured a random representative video frame and ran our analysis on that. Using this corpus of data we were able to calculate our angular gaze accuracy for students and instructors respectively.

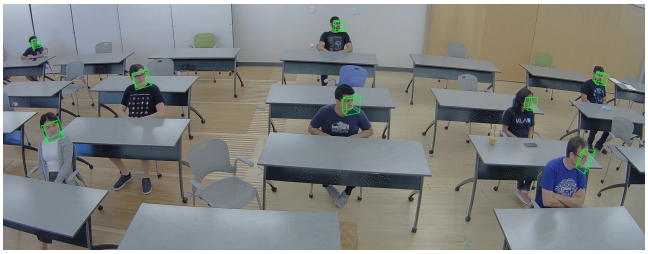

Figure 4: Sample scene from our controlled study with 3D gaze frustums overlaid in green.

5.2 Results

Given the known 3D locations of the gaze targets (captured via the AruCo markers), our pipeline can estimate the 6-DOF gaze for each user and therefore calculate the gaze accuracy in an automated manner. We found that our gaze estimation module has an average yaw error of 21.7° (S.D. = 2.6°) and an average pitch error of 20.9° (S.D. = 5.1°). For students, our yaw and pitch errors were 20.7° and 17.6° respectively. For instructors, our yaw and pitch errors were 24.8° and 31.7° respectively. The larger instructor gaze errors can be attributed to the fact that the instructors were standing, while the students were seated. Hence, this resulted in a more oblique viewpoint and a larger pitch error for this particular classroom.

On average, across all trials and occupants, our system had an angular gaze error of 21.3°. This compares favorably to prior systems in Table 1. Aung et al. [8] reports an angular error of 38.3°; EduSense [3] uses the same number of cameras as our system, and demonstrates a gaze error of 34.6°. Prior work also did not calculate higher-level semantics, such as dwell times across objects of interest. The higher accuracy afforded by our system enables us to explore these fine-grained uses and provides a robust platform for future researchers to build upon (Figure 2).

6 IN-THE-WILD EVALUATION

In addition to our controlled study, we also ran our system in five real classrooms (see Figure 1, right and Figure 5), capturing data for one semester. These classrooms varied in physical size and shape, as did the number of enrolled students (and thus occupant density). We used this deployment to not only test our system's stability and performance, but also capture data for a real-world evaluation. Before any video recordings were made, a researcher visited the class to explain the research project and the types of data collected. All instructors and all students had to consent to take part in the research, or the class was dropped from the study.

6.1 Procedure

To generate images for annotations, we first pulled 2400 frames at random for students and instructors each. We only took frames that contained at least one person's face in them for both the student and instructor views. These images were then annotated for three tasks, namely: 1) the number of false positive faces detected, 2) the number of missed faces, and 3) testing the accuracy of our gaze pipeline for the faces that were correctly detected.

All images were annotated by a team of privately-hired crowdworkers who were experienced in body bounding box and face annotation tasks. All data remained on university-controlled machines and encrypted over HTTPS. Additionally, all images were water-marked with overlays and machine annotations to significantly deteriorate value to third parties.

For the first task, the workers were asked to mark all the bounding boxes of faces that did not contain a face. These included incorrect "ghost" bounding boxes, boxes on the neck and hands of people. These annotations were used to calculate the false positive rate of our face detection module in a classroom setting. The next task was to annotate all the faces that were missed by our model. For this, the annotators were asked to label all the faces that were visible (even partially), but missed by our face detection system. The last task was to help compute the gaze error of our system. Here, the annotators were shown a face bounding box with a gaze vector superimposed. For all the faces that were detected correctly, they were asked to evaluate whether the prediction was correct or not. An incorrect prediction meant a gaze arrow that was off by more than 15° in either yaw or pitch. Workers were provided exemplary images of correct and incorrect gaze detections to calibrate. Workers were also provided a detailed document containing edge cases. Workers were encouraged to mark images with the 'I'm not sure' tag to discourage guessing. Each image was labelled independently by two crowdworkers.

6.2 Results

We now break down the results for our face detection and gaze estimation modules across all three tasks listed above. Across our experiments, our inter-reviewer reliability was 92.54%. In general, our face detection model had a false positive rate (faces that were detected incorrectly) of 4.37% (4.04% for students and 4.74% for instructors respectively). The low discrepancy between student and instructor frames is due to the common errors occurring in both cases - such as incorrect detections on chairs or hands. As such errors are agnostic to student or instructor viewpoints, we see a common rate of errors for both conditions.

On average, our face detection module missed 4.5% of students faces and 0.83% of instructor faces. The lower miss rate for instructors can be chiefly attributed to rare occlusion of the instructor's face (in contrast to students, who look down, partially cover their face with their hands, faces blocked by students in other rows, etc.).

Of the faces that were correctly detected by our model, our gaze estimation accuracy did not vary much across students and instructors, having an accuracy of 90.03% and 91.09% respectively. This suggests that once a face is detected reliably, our gaze estimation module is reasonably robust to viewpoints and partial faces.

7 LIMITATIONS

While the results of our system look promising, there are several technical limitations that should be addressed. First is that our model does not track eye gaze directly, but rather makes use of head pose as a proxy for gaze. Furthermore, estimating point of gaze from the gaze vector still has some ambiguity in our system. A single gaze vector can intersect with multiple planes or objects, thus having multiple candidate focal points. In such cases, a cone

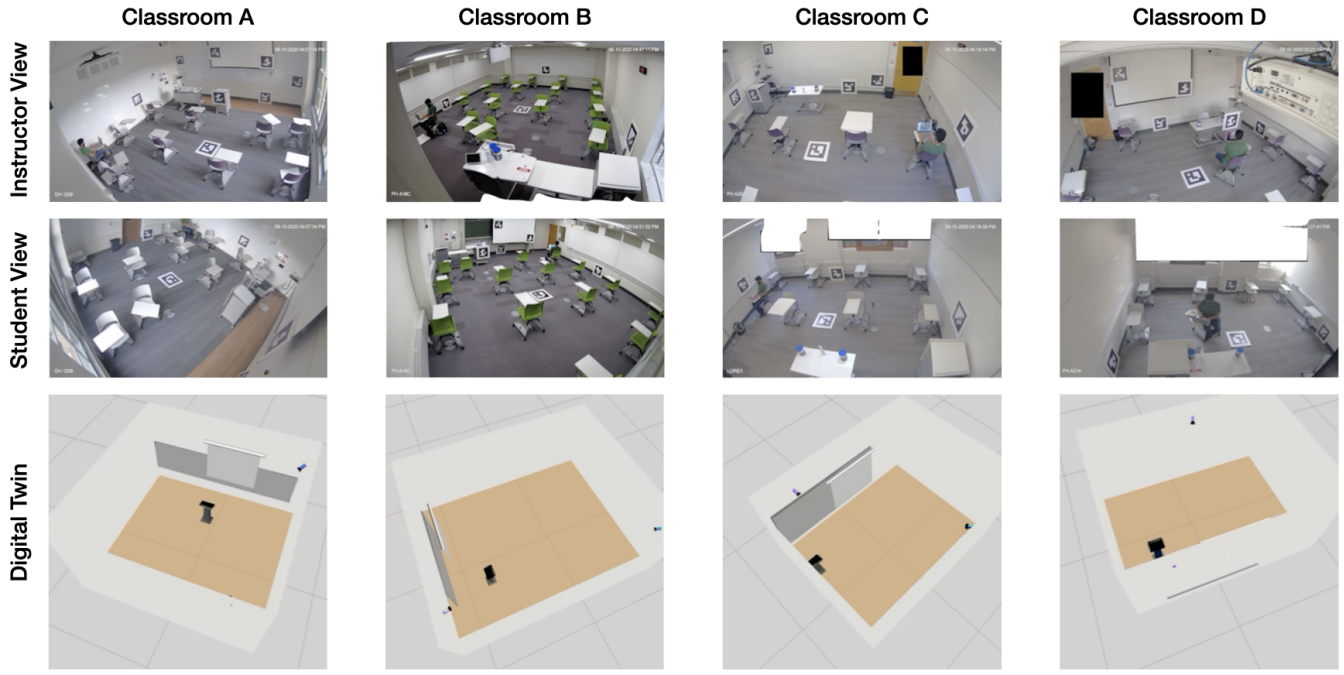

Figure 5: Exemplary digital twins (bottom row) of 4 classrooms created from combining the Instructor (top row) and Student (middle row) views.

of gaze, rather than a point of gaze may make more semantic sense. Lastly, our algorithm can suffer from occlusion and lack of field of view for bigger classrooms. As we only make use of two cameras to digitize whole classrooms, there are some cases wherein the students do not lie within the field of view of the camera or are occluded by other students seated in front of them.

Beyond gaze-specific concerns, the concept of digital classroom twins as an approach to education research and professional development also has limitations. For instance, simple capturing of sensor data is unlikely to suffice in order to make the data of use to instructors, as noted in our User Interface section. Instead, further processing likely has to be done. This means such a concept will need an ecosystem of applications that analyze or format the data in interpretable ways to make it useful to teachers or researchers, an area we hope to explore in future work.

8 DISCUSSION AND FUTURE WORK

The work in this paper contributes to a long line of research into technology to support professional development through replay. For instance, prior work has used video recall as a stimulus for teacher professional development (see e.g., [24]), allowing teachers to watch and reflect on a 2D version of their own practice. Another promising new approach for professional development that has been explored recently is the idea of virtual classroom simulations (see e.g., [31]). Similar to medical or aviation simulations, such simulations allow participants to test out difficult, rare, or risky behaviors without taking action in the real world. In this approach a virtual classroom environment is created with simulated students. Then, student behaviors, dialog, and other types of interactions can be programmed into the simulation. Classroom simulations have been explored particularly with novice or pre-service instructors who have had little experience in the classroom, allowing them to practice before they stand in front of a room of skeptical students.

The digital twin approach holds the possibility to combine the power of these two approaches. The twinned room could be replayed over time, allowing it to act as a video recall, but in three dimensions and providing the opportunity to move around in the space and take alternate perspectives – such as in a virtual reality interface like that described above. A teacher could therefore really experience the class from any students' perspective with greater immersion than a video can provide. On the other hand, it could also take on characteristics of a classroom simulation, improved by seeding that simulation with one's own data in a model of one's very own classroom. This could allow teachers to reflect on and make new choices stimulated by a moment of their own teaching rather than hypothetical or invented situations. This approach has the potential to be much more powerful with greater relevance to even expert teachers.

The work described in this paper therefore can help to open a new avenue for the future of professional development systems using digital twins. While we specifically focus on gaze as a feature of interest, such a system could enable instructor professional development across a range of classroom features. For instance, other contributing components of teacher immediacy include movement around the classroom, open and welcoming posture, facial expressions, and more. Introducing audio features of the classroom would allow for exploration of student and teacher dialog situated in time

and space (e.g., are the students participating only in the front row of the class?).

Beyond professional development, this concept is also one that holds potential for researchers. The ability to automatically detect a broad range of classroom features like gaze that are now situated more richly in their 3D context could facilitate the study of many open questions in the learning sciences.

In our own work, we expect to next develop a teacher-facing interface for use as a professional development tool as described above. This will introduce a number of interesting challenges regarding the presentation of the digital twin environment, as well as its integration with other professional development supports such as trained human observers. We also intend to explore additional features beyond gaze in our digital twin environment, some of which are described above.

9 CONCLUSION

In this paper, we introduced the concept of a classroom "digital twin" to aid in both research and professional development. We describe our generalizable approach to capturing the physical environment needed for such a twin, and the sensing of a particular feature of interest: instructor and student gaze. With this sensing approach, we ran two studies that have demonstrated the accuracy of our system. The first, a controlled study using known targets, demonstrates that this system can reduce the error of prior non-worn classroom gaze systems by roughly half. The second, an in-the-wild study conducted in multiple and varied classrooms over the course of a semester, demonstrates the ecological validity of our approach. This work advances the literature on classroom gaze systems while simultaneously opening up new avenues for classroom research and professional development through digital twins, i.e., high-fidelity simulation environments that employ real data streams.

ACKNOWLEDGMENTS

This research was generously supported with funds from the CONIX Research Center, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA. Our sincere thanks to the McDonnell Foundation and National Science Foundation for supporting this work (IIS-1822813, IIS-1747997 and CSR-1526237) for their generous support of fostering new research in education technologies. We also appreciate all of the help from the staff of CMU Media Services, particularly Brian Fitzgerald and Dan Noulett. Special thanks also to Murdar Memari, Ketaki Rao and Pranav Dheer for their efforts on this project. We are also grateful to Zensors, Inc. for letting us utilize their private crowd for labeling. Lastly, we would like to sincerely thank all the students and instructors who participated in this study and granted us permission to collect data in their classrooms.

REFERENCES

[1] 2017. National Infrastructure Commission (2017), "Data for the public good", National Infrastructure Commission report, London, December 14, p. 76. http://www.nic.org.uk/publications/data-public-good/.

[2] 2020. Three.js is a cross-browser JavaScript library and application programming interface used to create and display animated 3D computer graphics in a web browser using WebGL. https://threejs.org/.

[3] 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

[4] Janis F. Andersen. 1979. Teacher Immediacy as a Predictor of Teaching Effectiveness. *Annals of the International Communication Association* 3, 1 (1979), 543–559. https://doi.org/10.1080/23808985.1979.11923782 arXiv:https://doi.org/10.1080/23808985.1979.11923782

[5] Janis F. Andersen and Peter A. Andersen. 1987. Never Smile Until Christmas? Casting Doubt on an Old Myth. *Journal of Thought* 22, 4 (1987), 57–61. http://www.jstor.org/stable/42589246

[6] Janis F. Andersen, Peter A. Andersen, and Arthur D. Jensen. 1979. The measurement of nonverbal immediacy. *Journal of Applied Communication Research* 7, 2 (1979), 153–180. https://doi.org/10.1080/00909887909365204 arXiv:https://doi.org/10.1080/00909887909365204

[7] Janis F Andersen, Peter A Andersen, and Arthur D Jensen. 1979. The measurement of nonverbal immediacy. *Journal of applied communication research* 7, 2 (1979), 153–180.

[8] Arkar Min Aung, Anand Ramakrishnan, and Jacob Whitehill. 2018. Who are they looking at? Automatic Eye Gaze Following for Classroom Observation Video Analysis. In *EDM*.

[9] Ann E. Austin. 2002. Preparing the Next Generation of Faculty. *The Journal of Higher Education* 73, 1 (2002), 94–122. https://doi.org/10.1080/00221546.2002.11777132 arXiv:https://doi.org/10.1080/00221546.2002.11777132

[10] Jeremy N Bailenson, Nick Yee, Jim Blascovich, Andrew C Beall, Nicole Lundblad, and Michael Jin. 2008. The use of immersive virtual reality in the learning sciences: Digital transformations of teachers, students, and social context. *The Journal of the Learning Sciences* 17, 1 (2008), 102–141.

[11] Jon Bidwell and Henry Fuchs. 2011. Classroom Analytics: Measuring Student Engagement with Automated Gaze Tracking. (11 2011). https://doi.org/10.13140/RG.2.1.4865.6242

[12] G. Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools* (2000).

[13] George Breed. 1971. Nonverbal Behavior and Teaching Effectiveness. Final Report. (1971).

[14] Kathleen T. Brinko. 1993. The Practice of Giving Feedback to Improve Teaching. *The Journal of Higher Education* 64, 5 (1993), 574–593. https://doi.org/10.1080/00221546.1993.11778449 arXiv:https://doi.org/10.1080/00221546.1993.11778449

[15] Dario Cazzato, Marco Leo, Cosimo Distante, and Holger Voos. 2020. When I Look into Your Eyes: A Survey on Computer Vision Contributions for Human Gaze Estimation and Tracking. *Sensors* 20, 13 (2020), 3739.

[16] Maria Cutumisu, Krystle-Lee Turgeon, Tasbire Saiyera, Steven Chuong, Lydia Marion González Esparza, Rob MacDonald, and Vasyl Kokhan. 2019. Eye Tracking the Feedback Assigned to Undergraduate Students in a Digital Assessment Game. *Frontiers in Psychology* 10 (2019), 1931. https://doi.org/10.3389/fpsyg.2019.01931

[17] Elise J Dallimore, Julie H Hertenstein, and Marjorie B Platt. 2004. Classroom participation and discussion effectiveness: Student-generated strategies. *Communication Education* 53, 1 (2004).

[18] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. 2019. RetinaFace: Single-stage Dense Face Localisation in the Wild. *CoRR* abs/1905.00641 (2019). arXiv:1905.00641 http://arxiv.org/abs/1905.00641

[19] A. El Saddik. 2018. Digital Twins: The Convergence of Multimedia Technologies. *IEEE MultiMedia* 25, 2 (2018), 87–92.

[20] Michael Eraut. 1995. Schon Shock: a case for refraining reflection-in-action? *Teachers and teaching* 1, 1 (1995), 9–22.

[21] Nan Gao, Wei Shao, Mohammad Saiedur Rahaman, and Flora D Salim. 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 (2020), 1–26.

[22] X. Gao, Xiaorong Hou, Jianliang Tang, and H. Cheng. 2003. Complete Solution Classification for the Perspective-Three-Point Problem. *IEEE Trans. Pattern Anal. Mach. Intell.* 25 (2003), 930–943.

[23] S. Garrido-Jurado, R. Mu noz Salinas, F.J. Madrid-Cuevas, and M.J. Marín-Jiménez. 2014. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition* 47, 6 (2014), 2280 – 2292. https://doi.org/10.1016/j.patcog.2014.01.005

[24] Vince Geiger, Tracey Muir, and Janeen Lamb. 2016. Video-stimulated recall as a catalyst for teacher professional learning. *Journal of Mathematics Teacher Education* 19, 5 (2016), 457–475.

[25] David Gelernter. 1991. *Mirror worlds, or, The day software puts the universe in a shoebox– : how it will happen and what it will mean.* Oxford University Press, New York.

[26] Edward Glaessgen and David Stargel. 2012. The digital twin paradigm for future NASA and U.S. air force vehicles. https://doi.org/10.2514/6.2012-1818

[27] Dan Witzner Hansen and Qiang Ji. 2009. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE transactions on pattern analysis and machine intelligence* 32, 3 (2009), 478–500.

[28] Patricia Hardré and Alicia Burris. 2012. What contributes to teaching assistant development: Differential responses to key design features. *Instructional Science - INSTR SCI* 40 (12 2012). https://doi.org/10.1007/s11251-010-9163-0

[29] Richard Hartley. 2004. *Multiple view geometry in computer vision.* Cambridge University Press, Cambridge, UK New York.

[30] Charles Henderson, Andrea Beach, and Noah Finkelstein. 2011. Facilitating change in undergraduate STEM instructional practices: An analytic review of the literature. *Journal of Research in Science Teaching* 48, 8 (2011), 952–984. https://doi.org/10.1002/tea.20439 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/tea.20439

[31] Pavlos Kallonis and D Sampson. 2010. Implementing a 3D virtual classroom simulation for teachers' continuing professional development. In *Proceedings of the 18th International Conference on Computers in Education*. 36–44.

[32] Seng Loke, Sucha Smanchat, Sea Ling, and Maria Indrawan-Santiago. 2008. Formal Mirror Models: an Approach to Just-in-Time Reasoning for Device Ecologies. *International Journal of Smart Home* 2 (02 2008).

[33] S. W. Loke, B. S. Thai, T. Torabi, K. Chan, D. Deng, W. Rahayu, and A. Stocker. 2015. The La Trobe E-Sanctuary: Building a Cross-Reality Wildlife Sanctuary. In *2015 International Conference on Intelligent Environments*. 168–171.

[34] Linda McCroskey, Virginia Richmond, and James McCroskey. 2002. The scholarship of teaching and learning: Contributions from the discipline of communication. *Communication Education* 51, 4 (2002), 383–391.

[35] Bolton Ruth N., McColl-Kennedy Janet R., Cheung Lilliemay, Gallan Andrew, Orsingher Chiara, Witell Lars, and Zaki Mohamed. 2018. Customer experience challenges: bringing together digital, physical and social realms. *Journal of Service Management* 29, 5 (01 Jan 2018), 776–808. https://doi.org/10.1108/JOSM-04-2018-0113

[36] Elisa Negri, Luca Fumagalli, and Marco Macchi. 2017. A Review of the Roles of Digital Twin in CPS-based Production Systems. *Procedia Manufacturing* 11 (12 2017), 939–948. https://doi.org/10.1016/j.promfg.2017.07.198

[37] Amy Ogan. 2019. Reframing classroom sensing: promise and peril. *interactions* 26, 6 (2019), 26–32.

[38] Zenon W. Pylyshyn and Ron W. Storm. 01 Jan. 1988. Tracking multiple independent targets: Evidence for a parallel tracking mechanism*. *Spatial Vision* 3, 3 (01 Jan. 1988), 179 – 197. https://doi.org/10.1163/156856888X00122

[39] Qinglin Qi, Fei Tao, Tianliang Hu, Nabil Anwer, Ang Liu, Yongli Wei, Lihui Wang, and Andrew Nee. 2019. Enabling technologies and tools for digital twin. *Journal of Manufacturing Systems* (10 2019). https://doi.org/10.1016/j.jmsy.2019.10.001

[40] Mirko Raca. 2015. *Camera-based estimation of student's attention in class.* Technical Report. EPFL.

[41] Mirko Raca and Pierre Dillenbourg. 2013. System for assessing classroom attention. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*. 265–269.

[42] Mirko Raca and Pierre Dillenbourg. 2014. Holistic Analysis of the Classroom. In *Proceedings of the 2014 ACM Workshop on Multimodal Learning Analytics Workshop and Grand Challenge* (Istanbul, Turkey) *(MLA '14)*. Association for Computing Machinery, New York, NY, USA, 13–20. https://doi.org/10.1145/2666633.2666636

[43] James V Sherwood. 1987. Facilitative effects of gaze upon learning. *Perceptual and Motor Skills* 64, 3_suppl (1987), 1275–1278.

[44] Ben Shneiderman, Ellen Yu Borkowski, Maryam Alavi, and Kent Norman. 1998. Emergent patterns of teaching/learning in electronic classrooms. *Educational Technology Research and Development* 46, 4 (1998), 23–42.

[45] Omer Sumer, Patricia Goldberg, Kathleen Sturmer, Tina Seidel, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. 2018. Teachers' Perception in the Classroom. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[46] Fei Tao, Qinglin Qi, Lihui Wang, and AYC Nee. 2019. Digital twins and cyber–physical systems toward smart manufacturing and industry 4.0: correlation and comparison. *Engineering* 5, 4 (2019), 653–661.

[47] 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

[48] Narayanan Veliyath, Pradipta De, Andrew A Allen, Charles B Hodges, and Aniruddha Mitra. 2019. Modeling Students' Attention in the Classroom using Eyetrackers. In *Proceedings of the 2019 ACM Southeast Conference*. 2–9.

[49] Lingyu Zhang, Mallory Morgan, Indrani Bhattacharya, Michael Foley, Jonas Braasch, Christoph Riedl, Brooke Foucault Welles, and Richard J. Radke. 2019. Improved Visual Focus of Attention Estimation and Prosodic Features for Analyzing Group Interactions. In *2019 International Conference on Multimodal Interaction* (Suzhou, China) *(ICMI '19)*. Association for Computing Machinery, New York, NY, USA, 385–394. https://doi.org/10.1145/3340555.3353761

[50] Xiangyu Zhu, Xiaoming Liu, Zhen Lei, and Stan Z. Li. 2018. Face Alignment in Full Pose Range: A 3D Total Solution. *CoRR* abs/1804.01005 (2018). arXiv:1804.01005 http://arxiv.org/abs/1804.01005