ATL-BP: A Student Engagement Dataset and Model for Affect Transfer Learning for Behavior Prediction

Nataniel Ruiz, *Graduate Student Member, IEEE*, Hao Yu, *Graduate Student Member, IEEE*, Danielle A. Allessio, Mona Jalal, *Student Member, IEEE*, Ajjen Joshi, Tom Murray, John J. Magee, *Member, IEEE*, Kevin Manuel Delgado, Vitaly Ablavsky, *Senior Member, IEEE*, Stan Sclaroff, *Fellow, IEEE*, Ivon Arroyo, Beverly P. Woolf, Sarah Adel Bargal, and Margrit Betke, *Senior Member, IEEE*

***Abstract*—We propose a video-based transfer learning approach for predicting problem outcomes of students working with an intelligent tutoring system (ITS) by analyzing their faces and gestures. The ability to predict such outcomes enables tutoring systems to adjust interventions and ultimately yield improved student learning. We collected and released a labeled dataset of 2,749 problem-solving interaction samples of 54 students working with an intelligent online math tutor. Our transfer-learning challenge was then to design a representation in the source domain of images obtained from the Internet for facial expression analysis, and transfer this learned representation for human behavior prediction in the domain of webcam videos of students in a classroom environment. We developed a novel facial affect representation and a user-personalized training scheme that unlocks the potential of this representation. We designed several variants of a recurrent neural network that models the temporal structure of video sequences. Our final model, named ATL-BP for Affect Transfer Learning for Behavior Prediction, achieves a relative increase in the mean F-score of 50% over the state-of-the-art method on this new dataset. We also propose an additional set of annotations to predict students' engagement while solving a specific problem, and present models that can predict such engagement.**

***Index Terms*—Transfer learning, behavior prediction, engagement prediction, intelligent tutoring system, video classification.**

Manuscript received 10 March 2022; revised 6 August 2022; accepted 24 September 2022. Date of publication 30 September 2022; date of current version 7 August 2023. This article was recommended for publication by Associate Editor M. Vatsa upon evaluation of the reviewers' comments. *(Nataniel Ruiz and Hao Yu Contributed equally to this work.) (Corresponding author: Hao Yu.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the University of Massachusetts Amherst IRB, Federal Wide Assurance No. 00003909.

Nataniel Ruiz, Hao Yu, Mona Jalal, Kevin Manuel Delgado, Stan Sclaroff, and Margrit Betke are with the Department of Computer Science, Boston University, Boston, MA 02215 USA (e-mail: haoyu@bu.edu).

Danielle A. Allessio, Tom Murray, Ivon Arroyo, and Beverly P. Woolf are with the College of Information and Computer Sciences, University of Massachusetts Amherst, Amherst, MA 01003 USA.

Ajjen Joshi is with Affectiva, Boston, MA 02109 USA.

John J. Magee is with the Department of Mathematics and Computer Science, Clark University, Worcester, MA 01610 USA.

Vitaly Ablavsky is with the Applied Physics Laboratory, University of Washington, Seattle, WA 98195 USA.

Sarah Adel Bargal is with the Department of Computer Science, Georgetown University, Washington, DC 20057 USA.

Digital Object Identifier 10.1109/TBIOM.2022.3210479

I. Introduction

RESEARCH on developing intelligent tutoring systems (ITS) is a promising avenue for improving learning and education [1], [2], [3]. Previous work has shown that real-time signals from students can be used to improve their learning [4], [5], [6]. Predicting whether students are having trouble with problems can allow an ITS to provide interventions, such as providing hints or encouragement, which could help the students understand or solve the problem, thus improving learning outcomes.

MathSpring [1] is a popular online browser-based ITS that uses multimedia to encourage and support students as they solve math problems. Figure 2 shows the student interactive interface of MathSpring. Using the MathSpring ITS, a dataset named MathSpringSP [7] was collected, which includes 1,596 segmented videos of study sessions of students interacting with the ITS. Each problem tackled by a student has an associated outcome label automatically annotated by the ITS. Some example labels are *skipped*, *solved on first try*, *solved with hint*, among others. In this work we address the problem of predicting the outcome label from a video feed of the student while they are solving problems. As facial expressions and gestures are important cues for inferring problem outcomes, we propose to learn an affect representation using in-the-wild images for facial expression recognition, and transfer it to the task of predicting learning outcomes of students. Having a model that can successfully predict outcomes while a student completes a problem can help the ITS provide interventions such as hints or encouragement when the student is having difficulties.

Facial and gesture analysis are valuable tools for predicting emotions, but the question of how to use them for predicting student performance with an ITS remains challenging since cues can be very subtle or ambiguous. A smile, for example, does not necessarily mean that the student is happily solving an exercise. Instead, it could indicate a student's embarrassment for not knowing the answer to a question. Moreover, in our experience, trying to obtain valid ground truth labels of the student videos from human annotators is a futile experimental task because humans have a very low accuracy rate when predicting problem outcomes from video. Just like automated facial analysis tools, human annotators struggle with interpreting the ambiguity in and limited amount of information given by student gestures.

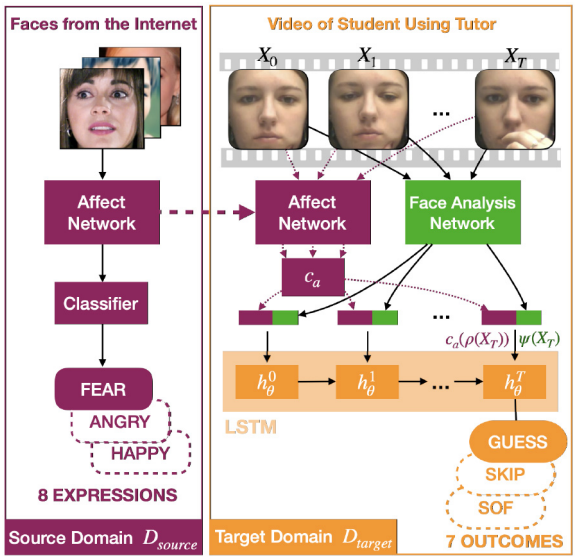

Fig. 1. Our proposed *Affect Transfer Learning for Behavior Prediction* (ATL-BP) model for predicting the behavior of students working with an intelligent tutoring system. For the source domain task of affect recognition (left), an affect network is trained to extract an affect representation from images of faces, which is used for classifying eight expressions. We employ this trained affect network for solving the target-domain problem of student outcome prediction. The target-domain ATL-BP model (right) consists of three components, the trained affect network, a facial analysis network, and an LSTM.

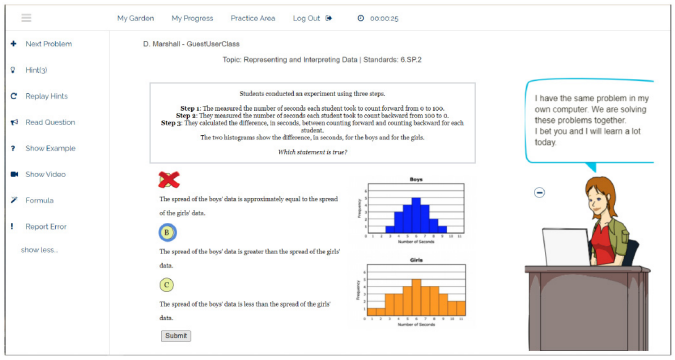

Fig. 2. MathSpring Interface. A sample MathSpring problem is presented in the figure. Problems are either multiple choice, short answer, or check-all-that-apply. When students are solving problems, hints, worked-out examples, tutorial videos, and formulas related to the problem are available from the corresponding buttons on the left. On the right, a learning companion (Jane) encourages and supports students when they make mistakes.

Prior research in transfer learning for facial analysis tasks mostly focuses on transfer learning for the same task in order to bridge domain gaps such as personalization of a prediction system to specific individuals [8], [9], [10], [11], [12], [13], improving results on a benchmark by fine-tuning neural networks that are pre-trained on external datasets for a similar prediction task [14], or improving results by pre-training on a related facial analysis task [15], [16]. In contrast, our work tackles the more challenging transfer learning across domains and tasks, which is a form of *transductive transfer learning* [17]. Specifically, we tackle the problem of learning a representation in the source domain of in-the-wild pictures for the task of facial expression analysis and transferring this learned representation to the task of human behavior prediction in the domain of webcam videos in a controlled environment (Figure 1). While prior work has explored transfer learning from facial analysis to behavior analysis, for example, using VGG-Face facial recognition embeddings to predict driver distraction [18], our work is, to the best of our knowledge, the first to propose leveraging an affect representation, learned using a deep neural network, for a behavior prediction task. Our learned affect representation is general and can be used not only for predicting problem outcomes on an ITS, but in any human behavior prediction problem where affect and expression are important cues.

The largest obstacle in training an end-to-end deep learning model for behavior analysis problems is the fact that data are relatively scarce, which increases the risk of overfitting. As a first step to alleviating the data problem, we present MathSpringSP+, an extended version of the MathSpringSP dataset, which is roughly double the size of the original dataset. Next, we propose a novel facial affect representation for behavior prediction problems that is learned from a large affect classification dataset. We show that, by incorporating this affect embedding, we can obtain improvements compared to more traditional deep face embeddings such as the VGG-Face facial recognition embedding [19]. We developed a two-layer *Long Short Term Memory* (LSTM) model [20] that takes into account the temporal structure of the problem and successfully leverages our affect embedding. We show that, by conducting user-personalized training where a small portion of a student's initial captured data is used to fine-tune the model, our method outperforms the previous state-of-the-art method [7] by 50%. We present a video dataset of problem-solving interactions of children and show that fine-tuning the ATL-BP affect network using children face images further improves the performance. Finally, we augment the set of annotations for the dataset to include the perceived engagement ('Looking at screen,' 'Looking at paper,' and 'Wandering') of the students while working on MathSpring. In this paper, we expand on our previous work [21] and summarize the comprehensive set of contributions as follows:

- We present MathSpringSP+: a large labeled dataset of student interactions with an intelligent online math tutor consisting of 68 sessions, where 54 students solved 2,749 problems in total. We make this dataset publicly available.[1]
- We present a novel affect transfer learning representation that can be used for behavior prediction tasks. We are the first to model the temporal structure of video sequences of students solving math problems using a recurrent neural network architecture.
- Our proposed *Affect Transfer Learning for Behavior Prediction* (ATL-BP) model outperforms the previous state-of-the-art method by 50%.
- We show that finetuning the ATL-BP affect network using children face images further improves the performance on MathSpring Children Dataset, a dataset of children problem-solving interactions collected in the same manner as MathSpringSP+.

[1]https://www.cs.bu.edu/faculty/betke/research/learning/

- We collected additional frame-wise labels of student engagement and trained models to demonstrate the possibility of successfully predicting engagement. This would enable future exploration of how affect, engagement, and learning outcomes correlate. We also make this additional set of labels publicly available.[1]

II. Related Work

Intelligent Tutoring Systems: ITSs have been evaluated and shown to produce learning gains [22], [23], [24], [25], [26], [27]. One meta-analysis shows test score improvements from the 50th to 75th percentile [28]. Some ITSs have been shown to match the success of one-on-one human tutoring and students using these tutors outperform students from conventional classes in 92% of the controlled evaluations and perform twice as high as for students using typical (non-intelligent) systems [28], [29], [30].

Prior research has analyzed user affect, emotions and expressions from interactions with educational games [31] or intelligent tutoring systems [22], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42]. In certain cases the predicted affect information is used to improve learning. For example, Strain and D'Mello [43] have studied the role of emotion in ITS engagement, task persistence, and learning gain. Gaze prediction has also been used in an effort to respond to students' boredom and to perform interventions [22]. Further, relationships between visual facial Action Unit (AU) factors and self-reported traits such as academic effort, study habits, and interest in the subject have been studied [39].

In contrast to this body of work, our work focuses on using predicted deep affect embeddings that are learned from a large facial affect dataset to improve behavior prediction in an ITS. Behavior prediction can be useful in improving learning by tailoring the interventions of the ITS to the predicted actions of the student. To the best of our knowledge, our work is the first to use an affect embedding for behavior prediction in an ITS.

Interventions in an Online Tutor: Prior research has examined the impact of several interventions in ITS to improve student outcome and affect, specifically, affective messages delivered by avatars and empathetic messages that responded to students' recent emotions [3]. Interventions in the MathSpring ITS led to improved grades in state standardized exams [44] as well as influenced students' perceptions of themselves as learners [45]. Empathetic characters that provide interventions generate superior results both to improve student interactions with the system, address negative student emotions, and in the overall learning experience [46]. Predicting outcomes of problems for students is a valuable source of information for planning and executing ITS interventions for improving learning [47], [48]. For example the ITS could provide hints when the system predicts that the student will not be able to successfully complete the problem.

Predicting Exercise Outcome: Joshi et al. [7] presented a first attempt at tackling the problem of exercise outcome prediction. They did not explore deep learning representations but used traditional facial analysis features such as head pose, gaze and facial action units (AUs). They also did not attempt to model the temporal component of the videos, which is a rich source of information, and instead opted to summarize features from a video into one single feature vector. The method by Joshi et al. [7] can be considered the previous state of the art in student outcome prediction, and, thus, our experimental results include a performance comparison between this method and our models.

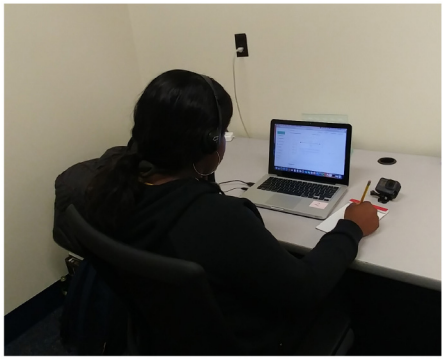

Fig. 3. Data capture setup for the MathSpringSP+ dataset. The student completes intelligent tutor problems on a laptop while being recorded with the laptop camera. The student may use a pad and pen to solve the problems. If the student writes with the right hand, as here, the pad is located to the right of the laptop, and the Go-Pro camera is also placed to the right so that the students' upper body and face can be recorded during the completion of the problems.

III. Datasets

In order to build an ITS capable of understanding student behavior and producing interventions, it is critical to build tailored datasets that allow development of behavior understanding techniques. To this end, we present datasets of students from different age groups interacting with the online tutor MathSpring [1]. Specifically, MathSpringSP+ Dataset includes videos of college students, and MathSpring Children Dataset comprises videos of sixth grade students. We make MathSpringSP+ Dataset publicly available.

The datasets were collected with informed consent by participants. Our institution's Internal Review Board (IRB) has approved the human subject research and the data collection process.

A. MathSpringSP+ Dataset

In this work, we expand the MathSpringSP dataset described by Joshi et al. [7], following the same data collection protocol. The extended dataset MathSpringSP+ is roughly double the size of the original MathSpringSP dataset.

Data Collection: MathSpringSP+ consists of Webcam and GoPro videos that are recorded while college students solve math problems using the online tutor MathSpring [1] on a laptop. The webcam is positioned on the laptop and films the student at a frontal angle. Figure 3 illustrates our data capturing setup (for right-handed students). For right-handed students, the GoPro video cameras are placed on the right above the students' pad of paper. When students look down to use their paper and pencil to work solving problems, the

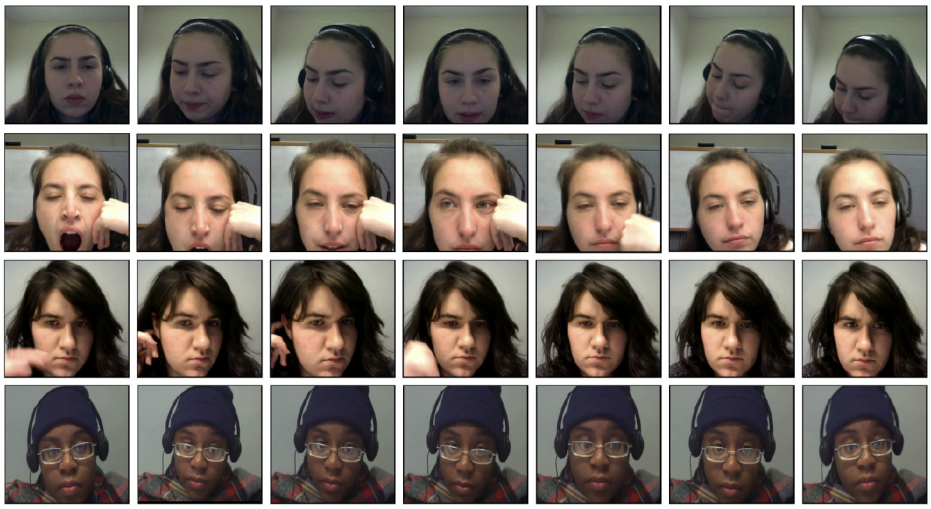

Fig. 4. Example face-cropped images from the MathSpringSP+ dataset showing the evolution of student expressions. In particular we notice changes in head pose, hand gestures, face occlusion and facial gestures throughout the videos. Expressions in videos can be very subtle, as well as ambiguous, making the prediction problem challenging.

GoPros capture the view of the students' faces. For left-handed students, the GoPro cameras are placed on the left. The GoPro and webcam cameras are synchronized. At the beginning of each session, a few seconds of the desktop clock are recorded by the GoPro camera and then a movie clapper is clapped in front of the webcam camera allowing both the cameras to hear and record the clap. Before students start to solve math problems, they are asked to finish an oral expressiveness baseline survey and a pre-survey. The baseline expressiveness survey contains questions such as 'What is your least favorite school subject? Why do you dislike it so much?'. Students read the questions out loud and then they click submit to move through questions. For the pre-survey, students need to give the Likert rating ('This is VERY MUCH/MOSTLY/SOMEWHAT/NOT MUCH/NOT LIKE me') for each question. An example of a pre-survey question is 'I have overcome setbacks (obstacles on the way) to conquer an important challenge.'

Students work on solving math problems for 30–40 minutes or approximately 50 problems. The number of problems solved is variable between sessions depending on the rate at which each student solves problems. We divide each student's video session into shorter video segments, where each segment is associated with an individual math problem. Each math problem video clip has an associated problem outcome y, recorded in the log files of the ITS [7]. This problem outcome is automatically labeled by the software using a rule-based algorithm that chooses from the following seven possible student outcomes:

- ATT (attempted): Student did not see any hints and solved the problem after one incorrect attempt
- GIVEUP: Student tried to answer the problem or asked for a hint but ultimately skipped the problem
- GUESS: Student did not see hints, but solved the problem after more than one incorrect attempts
- NOTR (not read): Student performed some action, but the first action was too fast for the student to have read the problem
- SHINT (solved with hint): Student eventually submitted the correct answer after seeing one or more hints
- SKIP: Student skipped the problem without asking for a hint or attempting to answer the problem
- SOF (solved on the first attempt): Student answered correctly on the first attempt, without seeing any hints

Dataset Details: Examples of the variation in student facial expression throughout the process of answering problems in the math tutor are shown in Figure 4. We note that expressions can be very subtle. Expressions can also be ambiguous: a frown can mean that the student is very focused and will solve the problem correctly or that they are having difficulties with the problem. Expression intensities and variance depend on the individual, and it is challenging to generalize to different identities. Finally, our method has to deal with hand gestures, face occlusions and extreme pose changes, some of which are shown in Figure 4. A total of 24 students participated in the extended study, compared to 30 in the original study. We note that the dataset only includes individuals who have provided written consent that their data may be used publicly for research purposes. Several students participated in multiple sessions. Each session lasted approximately one hour. In total, 30 student sessions were recorded, which yielded 1,153 problem samples. Thus, the extended MathSpringSP+ dataset contains videos of a total of 54 unique students, 68 student sessions and 2,749 problem samples. This amount of data almost doubles the original MathSpringSP dataset, which contains 38 student sessions and 1,596 problem samples. A detailed breakdown of the relative sizes of MathSpringSP and MathSpringSP+ is shown in Table I.

B. MathSpring Children Dataset

Besides expanding the previous MathSpringSP dataset, we further collected a dataset of sixth grade students who used MathSpring in Latin America following the same data collection protocol. MathSpring Children Dataset presents videos of

TABLE I
SIZE COMPARISON OF OUR EXTENDED MATHSPRINGSP+ DATASET COMPARED TO MATHSPRINGSP

	MathSpringSP	MathSpringSP+
Individual Students	30	54
Student Sessions	38	68
Problem Samples	1,596	2,749

Fig. 5. Sixth grade students are working on MathSpring in their classroom. The figure shows the classroom layout and data capture setup of our MathSpring Children Dataset.

students from a younger age group, allowing us to explore the generalizability of our behavior prediction models to different age groups.

Fifty-one sixth grade students and their teachers used a version of MathSpring translated into Spanish for 2 months in their daily classes (Figure 5). With their corresponding parental consent, students who used MathSpring in three different schools in Argentina were videorecorded. This dataset contains 58 sessions, over 35 hours of facial expressions of 11-year-old children using MathSpring to practice math problem solving as part of their regular mathematics classes in either Spanish-speaking or bilingual schools. Following the same data processing and annotation steps, 968 recorded problem-solving interaction samples as well as the seven problem outcome annotations have been collected.

IV. METHOD

A. Problem Formulation

The dataset consists of labeled video pairs $(\boldsymbol{X}, y)$, where the video $\boldsymbol{X}$ is a time series of RGB frames $\boldsymbol{X} = \{\boldsymbol{X}_t \mid t = 1, \ldots, T\}$ of a student solving a problem, and the scalar label y indicates the outcome class for that problem. The task is a 7-label classification problem, i.e., $y \in \{1, \ldots, C\}$, for $C = 7$.

B. The Proposed ATL-BP Framework

The proposed ATL-BP model consists of three main components (Figure 1), the affect network trained for the source domain problem of affect recognition, a facial analysis network, and an LSTM. We also study variants of our model by either removing the affect network or replacing it with a face recognition network.

1) Source Domain Learning: Our challenge was to determine how to leverage state-of-the-art affect recognition techniques to compute an output label y from the input video $\boldsymbol{X}$. Affect recognition models provide affect estimates from images of faces that typically show strong emotions, e.g., the fear expressed in the women's face on the left in Figure 1. We used a ResNet-50 network [49] and the AffectNet dataset [50], which contains more than one million facial images collected "in the wild" from the Internet, to solve the source domain problem of predicting eight emotions (neutral, happiness, sadness, surprise, fear, disgust, anger, and contempt), plus the two classes (uncertain, and non-face). We employ this trained affect network to solve the target-domain problem of student outcome prediction.

2) Feature Extraction: First, from the last layer of the trained affect network, ATL-BP extracts a fixed-size embedding of size 8,192, computed for each frame $\boldsymbol{X}_t$, and compresses it into a lower-dimensional vector $\rho(\boldsymbol{X}_t)$ by learning the weights for a fully-connected neural network layer c_a (Figure 1, magenta). The intuition behind having this learnable linear layer that compresses the representation is that the LSTM can struggle with very large representations ($>1,000$ in this case), especially since it has to also learn the temporal relationship between these vectors. To make the task easier for the LSTM (which has 200 hidden units per layer), we reduced the representation. We found that this improved performance and that training convergence was faster in early iterations.

Second, ATL-BP uses a facial analysis model to extract facial Action Unit (AU) presence and intensity, gaze direction, and head pose for each frame $\boldsymbol{X}_t$. We note these traditional facial analysis features as $\psi(\boldsymbol{X}_t)$ (Figure 1, green). We chose the OpenFace 2.0 model [51] to compute student head position, head pose, gaze, facial AU presence, and facial AU intensity from individual frames in each video segment.

For our main ATL-BP model we devised a feature representation that is based on concatenating the outputs of our proposed affect representation and the facial analysis components:

$$\phi(\boldsymbol{X}_t) = c_a(\rho(\boldsymbol{X}_t)) \oplus \psi(\boldsymbol{X}_t),$$

where $\oplus$ is the concatenation operation. The compressed embedding $c_a(\rho(\boldsymbol{X}_t))$ is 100-dimensional. The full feature vector $\phi(\boldsymbol{X}_t)$ is 149-dimensional for every frame $\boldsymbol{X}_t$.

3) Temporal Modeling: Finally, in order to model the temporal nature of the videos, we designed a unidirectional 2-layer LSTM classifier h with 200 hidden units that processes the feature vector $\phi(\boldsymbol{X}_t)$ frame by frame and produces the final estimate of student outcome y (Figure 1, orange).

4) Model Variants: We designed and studied two variants of our model (Figure 6). The first variant is *ATL-BP without transfer learning.* In this model, we removed the affect network, and the LSTM directly interprets the output ψ of the facial analysis network. For the second variant *ATL-BP with VGG-Face embedding*, we replaced the affect network by a face recognition model in order to extract face related features. We selected the pre-trained VGG-Face network [19], which computes an embedding ξ of dimension 2,622. ATL-BP compresses the feature representation $\xi(\boldsymbol{X}_t)$, computed by this network for each video frame $\boldsymbol{X}_t$, using another fully-connected layer c_v, into $c_v(\xi(\boldsymbol{X}_t))$. The LSTM then interprets the output $c_v(\xi(\boldsymbol{X}_t))$ concatenated with the output ψ of the facial analysis network.

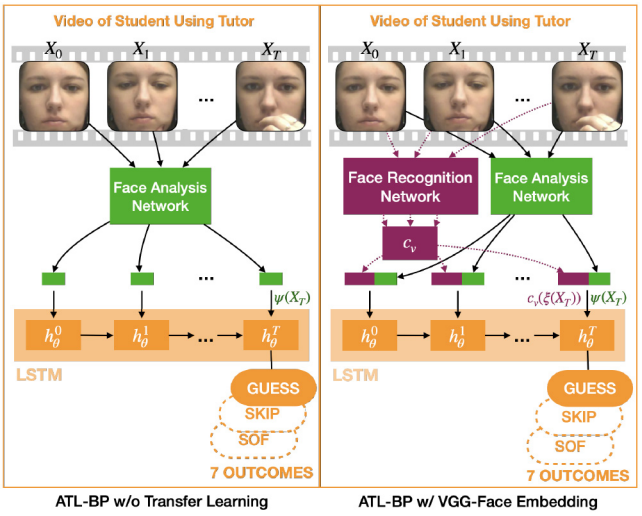

Fig. 6. Model variants. *ATL-BP without transfer learning* removes the affect network. *ATL-BP with VGG-Face embedding* replaces the affect network by a face recognition model.

V. Experiments

We present experiments on problem outcome prediction on the MathSpringSP+ dataset and MathSpring Children Dataset. These experiments study our contributions, which include incorporating temporal information from video streams by using an LSTM and using our affect transfer learning representation. The experiments also show how user-personalized training unlocks the effectiveness of our affect representation. We also study early prediction as well as present ablation studies for the dimensionality reduction that is accomplished by the proposed fully-connected layer. In this work we limit ourselves to the webcam video of the student. Finally, we include additional frame-wise labels of student engagement and present experiments showing the possibility of successfully predicting such engagement.

A. Implementation Details

We implemented all our models in PyTorch. All the experiments were conducted on an NVIDIA GeForce GTX TITAN X GPU. For facial analysis model, we used the official implementation[2] of OpenFace. We used its command line interface to extract head pose (three-dimensional location and rotation), three-dimensional eye gaze, and facial action units (the presence and the intensity of 18 pre-defined facial action units) for each video. We used default values for all the parameters of the facial analysis toolkit. Overall, a 49-dimensional feature vector is extracted for each video. We then standardized all the features by removing the mean and scaling to unit variance to obtain the final feature vector ψ. We also share the details of training the affect network and training ATL-BP for outcome prediction as follows.

1) Training the Affect Representation Network: For source domain affect training, we selected a ResNet-50 network. We pre-trained the affect network on a subset of 50,000 randomly sampled images from the AffectNet dataset and validated the network on 5,000 randomly selected images. We limited ourselves to a subset since the dataset contains more than one million examples. Note that our training and validation data subsets are not the same as used by [50]. On our subset, our network achieves a mean accuracy of 47.3%, which is close to the accuracy reported by [50] on their skew-normalized validation set of 54%, and much higher than the random baseline of 9.0%. The relatively low accuracy scores can be accounted for by data that is unbalanced, noisy, and overall challenging.

We used CNN based face detector from dlib [52] for both the source domain pre-training and feature extraction. We detected and cropped the face in the image and fed it into the affect network. We extracted the target domain affect features from our videos by performing inference of the affect network on every frame. We chose a granularity of three frames per second, down from 30 frames per second in our videos, in order to save on processing time and storage space. We found that this granularity was a good compromise between performance and cost. The affect network uses each frame as an input and the last-layer features are extracted as a vector of dimension 8, 192.

We trained the affect network with the Adam optimizer with a learning rate of 3×10^{-4}, β_1 of 0.9, and β_2 of 0.999. The standard batch normalization layers of the ResNet-50 were used and fixed throughout training.

2) Training ATL-BP to Predict Exercise Outcome: For each frame used, the feature vector computed is $\phi(\boldsymbol{X}_t) = \psi(\boldsymbol{X}_t) \oplus c_a(\rho(\boldsymbol{X}_t))$. The original dimension of $\rho(\boldsymbol{X}_t)$ is 8,192, and we further reduced it to 100 by a linear compression layer c_a. We observed that the dimensionality reduction due to the compression layer stabilizes training and improves performance. The feature vector ϕ is used to train the LSTM with two stacked layers. We adopted a 2-layer LSTM because we found that it provided the most appropriate model complexity to learn reasonable complex temporal features and achieved the best performance for our video dataset. One layer LSTM is too simple to capture the complex features, while more layers lead to overfitting issues. Moreover, recent success of Transformers [53] in computer vision [54], [55] has demonstrated the potential of applying Transformers in behavior prediction tasks. We believe that Transformers would provide performance benefits when applied to our task of problem outcome prediction, especially due to the long range dependencies that they are able to capture. Therefore, a natural next step, which we leave for future work, will be to replace the LSTM with Transformers in our model.

Specifically, at each instant t, features $\phi(\boldsymbol{X}_t)$ are fed to the 2-layer LSTM. The LSTM is trained on all the video segments. It outputs a class probability for each problem outcome. We used the standard implementation of a unidirectional 2-layer LSTM with 200 hidden units from PyTorch. The LSTM is trained using the cross-entropy loss function. The Adam optimizer is used for training. We used a learning rate of 3×10^{-5} for 30 epochs, and a batch size of 1. We used a batch size of 1 because we found this improved generalization compared to any other batch size. We found large batch sizes degraded performance a great deal. This is related to findings that stochastic gradient descent (SGD) with smaller

[2]https://github.com/TadasBaltrusaitis/OpenFace

batch sizes finds flatter local minima that generalize better (at least when training is not extremely long) [56]. Specifically here this is more important because the dataset is medium sized and not very large. The learning rate chosen allows the model to converge in a fair number of epochs. Any higher learning rate that we explored either led to early divergence of the model or lesser generalization.

B. Experimental Setup

Model Variations: In addition to our main proposed ATL-BP, shown in Figure 1 and which we call "ATL-BP with affect embedding" for clarity, we implemented and tested two variants of ATL-BP, *ATL-BP without transfer learning* and *ATL-BP with VGG-Face embedding*, as described in Section IV-B4. Furthermore, for comparison baselines, we reproduced the method described by Joshi et al. [7] and show results for a majority vote classifier. The majority vote classifier simply selects the most prevalent class in our dataset, "Solved on First Try," for every video.

Random Dataset Split: Following the experimental setup in [7], we performed five-fold cross validation by randomly shuffling video segments and constructing five different train and test splits. The train splits contain 80% of the data while the test splits contain the rest.

Experiments conducted using this random splitting experimental setup cannot reliably measure generalization to new users since videos of problems from the same student can be present in both the training and test set. This means that the network does not have to learn how to generalize to a completely new identity. We propose an improved experimental setup next.

User Generalization Split: In order to test generalization to new users we propose a leave-users-out experimental setup where users are exclusively split into either the training or test set. In other words, we enforce the rule that no video clips of the same user can be in both the test and training sets. In this manner we can measure how the system performs when applied to an unseen user. This is a substantially more challenging task since the network has to generalize to new identities and features. We suggest that all future research on this dataset use this type of setup. We created five leave-users-out splits for five-fold cross-validation and train different model variations for each split.

C. Results and Discussion

We present results and discussion on predicting seven problem outcomes on the MathSpringSP+ dataset and the MathSpring Children Dataset.

ATL-BP Results for Random Splits: Using the experimental protocol of a random dataset split, our ATL-BP for problem outcome prediction on MathSpringSP+ achieves an accuracy of 60.2% (Table II). Compared to the previous state-of-the-art method [7], this is an increase of 14 percent points (pp) in accuracy. ATL-BP also achieves a 44% relative increase in mean F-score improving from 0.238 to 0.330. The mean F-score is computed by first computing the individual F-score for all classes and averaging over all classes. By comparing the results for ATL-BP without transfer learning and those by Joshi et al. [7], we can see that by integrating an LSTM architecture that allows for modeling the temporal component in the videos we can achieve a marked increase in performance (5.6 pp). We achieve a further increase in performance by using deep embeddings (8.6 pp for using the VGG-Face embedding ξ), and especially our proposed affect embedding ψ (as mentioned, 14 pp).

TABLE II
RESULTS FOR PROBLEM OUTCOME PREDICTION ON THE MATHSPRINGSP+ DATASET USING FIVE-FOLD CROSS-VALIDATION AND RANDOM DATA SPLITS

Method	Mean F-Score	Accuracy
Majority Vote Classifier	0.103	56.1%
Joshi et al. [7]	0.228	46.2%
ATL-BP w/o transfer learning	0.295	51.8%
ATL-BP w/ VGG-Face embedding	0.304	54.8%
ATL-BP w/ affect embedding	**0.330**	**60.2%**

TABLE III
RESULTS FOR PROBLEM OUTCOME PREDICTION ON THE ORIGINAL MATHSPRINGSP FOR ATL-BP FOLLOWING THE DATA SETUP FROM JOSHI et al. [7]

Method	Mean F-Score	Accuracy
Joshi et al. [7]	0.270	54.0%
ATL-BP w/ affect embedding	**0.362**	**61.0%**

TABLE IV
RESULTS FOR EARLY PREDICTION OF PROBLEM OUTCOME USING ONLY THE FIRST FIVE SECONDS OF VIDEO FOOTAGE ON THE MATHSPRINGSP+ DATASET (FIVE-FOLD CROSS-VALIDATION, RANDOM DATA SPLITS)

Method	Mean F-Score	Accuracy
Majority Vote Classifier	0.103	56.1%
Joshi et al. [7]	0.173	46.7%
ATL-BP w/o transfer learning	**0.295**	51.8%
ATL-BP w/ VGG-Face embedding	0.239	47.0%
ATL-BP w/ affect embedding	0.270	**53.4%**

MathSpringSP Results: We conducted experiments on the original MathSpringSP dataset in order to verify that our ATL-BP model with affect embeddings achieves improved results in the same testing environment presented by Joshi et al. [7]. Our results show a consistent improvement in mean F-score and accuracy of our method (Table III).

Early Prediction of Problem Outcome: We experimented with obtaining prediction using only the five first seconds of each video clip (Table IV). Early outcome prediction is important since the ITS should have time to react and deliver the intervention should it be decided to do so. It turns out that to do early prediction is straightforward using an LSTM since it outputs a prediction at every time step, as opposed to the method proposed by Joshi et al. [7], where each video has to be summarized into a fixed-sized vector before being fed into a multilayer perceptron. We observe that ATL-BP achieves a large increase (6.7 pp) in performance over [7]. ATL-BP without transfer learning obtains the best F-score (0.295) in this experimental setup.

TABLE V
EMBEDDING DIMENSIONALITY REDUCTION ABLATION STUDY. WE SHOW RESULTS FOR PROBLEM OUTCOME PREDICTION ON THE MATHSPRINGSP+ DATASET USING FIVE-FOLD CROSS-VALIDATION AND RANDOM DATA SPLITS

Method	Mean F-Score	Accuracy
ATL-BP w/ VGG-Face	0.304	51.3%
ATL-BP w/ VGG-Face & dim. reduction	0.304	**54.8%**
ATL-BP w/ affect	0.330	58.7%
ATL-BP w/ affect & dim. reduction	0.330	**60.2%**

TABLE VI
GENERALIZING TO UNSEEN USERS FROM THE MATHSPRINGSP+ DATASET. RESULTS FOR PROBLEM OUTCOME PREDICTION ON THE MATHSPRINGSP+ DATASET USING FIVE-FOLD CROSS-VALIDATION AND THE MORE CHALLENGING LEAVE-USERS-OUT SPLITS

Method	Mean F-Score	Accuracy
Majority Vote Classifier	0.102	55.9%
Joshi et al. [7]	0.182	41.9%
ATL-BP w/o transfer learning	**0.270**	50.3%
ATL-BP w/ VGG-Face embedding	0.246	51.8%
ATL-BP w/ affect embedding	0.251	**54.0%**

TABLE VII
RESULTS FOR PROBLEM OUTCOME PREDICTION (7-CLASSES) ON THE MATHSPRINGSP+ DATASET AFTER USER PERSONALIZATION (FIVE-FOLD CROSS-VALIDATION AND LEAVE-USER-OUT SPLITS)

Method	Mean F-Score	Accuracy
Majority Vote Classifier	0.090	45.3%
Joshi et al. [7]	0.206	43.8%
ATL-BP w/o transfer learning	0.278	48.4%
ATL-BP w/ VGG-Face embedding	0.262	48.7%
ATL-BP w/ affect embedding	**0.308**	**55.1%**

TABLE VIII
GENERALIZING TO UNSEEN PROBLEMS. RESULTS FOR PROBLEM OUTCOME PREDICTION (7-CLASSES) ON THE MATHSPRING CHILDREN DATASET (FIVE-FOLD CROSS-VALIDATION, RANDOM DATA SPLITS)

Method	Mean F-Score	Accuracy
Majority Vote Classifier	0.070	32.3%
Joshi et al. [7]	0.202	32.0%
ATL-BP w/o transfer learning	0.238	33.4%
ATL-BP w/ affect embedding	0.260	39.6%
ATL-BP w/ LIRIS children affect embedding	0.272	45.2%
ATL-BP w/ CAFE children affect embedding	0.273	44.4%
ATL-BP w/ LIRIS+CAFE affect embedding	**0.278**	**45.2%**

Deep Embedding Dimensionality Reduction: We performed an ablation study on the fully-connected layer that is used for reducing the dimensionality of the deep embeddings that are used as inputs for our LSTM architecture (Table V). While the mean F-score does not change on both the VGG-Face and proposed affect embedding ATL-BP variants, dimensionality reduction does improve the accuracy of the models by 3.5 pp and 1.5 pp, respectively.

ATL-BP Results for User Generalization: For the user generalization split of the training and testing data, we report the mean F-score and mean accuracy in Table VI for the "Majority Vote Classifier" benchmark, Joshi et al. [7] and our proposed model with different combinations of embeddings. We observe that the temporal modeling improves results from Joshi et al. [7] substantially (12.1 pp in accuracy). We observe that ATL-BP without transfer learning outperforms the ATL-BP version with our proposed affect embedding with regards to the F1 score. We hypothesize that leveraging affect embeddings is more difficult in this setup since the model does not have access to baseline levels of expression for each user.

Personalization of Prediction: An effective real-time tutoring system would benefit from personalizing its prediction using initial data captured from a specific user stream. People have different emotional and expression baselines that can be learned using data collected in a trial run of the system. Specifically, we want the model to act on the variations of our affect embedding compared to the mean affect embedding, since each person will have a different baseline expression and thus a different baseline affect embedding. The model does not have any way to integrate this information without it being personalized for each user.

We propose a personalization scheme in which our system can be tailored to individual users and can fully utilize our proposed affect embedding. In this scheme, the network is fine-tuned on the initial problems corresponding to 20% of the session for users in the test set for 30 epochs. Our experiments show that user personalization unlocks the potential of the affect features (Table VII). ATL-BP with affect embedding achieves the highest F-score of 0.308 and the highest accuracy of 55.1% compared to the other methods. Our full method achieves a relative increase of 50% in mean F-score as well as an absolute increase in accuracy of more than 11 pp compared to the previous state of the art [7]. Our full method also outperforms variants of ATL-BP, which do not use our proposed affect representation.

Outcome Prediction for Children: As a final experiment we tested our method on a new dataset of children working on math problems. Results on this Children Dataset show that our model consistently outperforms the baseline and previous state-of-the-art method (Table VIII).

Since the AffectNet dataset mainly captures facial expressions of adults, we further finetuned the affect representation network using two datasets of children facial expressions, LIRIS [57] and CAFE [58], in order to tailor the model specifically for children. LIRIS contains 208 video clips of 6-to-12-year-old children showing six basic spontaneous facial expressions, while CAFE dataset contains 1,192 images of 2-to-8-year-old children posing for seven facial expressions. For the LIRIS dataset, we used extracted frames for training and validation. For both datasets, 90% of images were used for training and 10% of images were used for validation. The ResNet affect model achieves 99.1% accuracy and 85.8% accuracy on the validation set of LIRIS and CAFE dataset respectively. The trained affect network is then applied on MathSpring Children videos.

We trained three variants of models using LIRIS only (frames), CAFE only, and a combination of both datasets. For random data splits, the best model among the three achieves the highest accuracy (45.2%) and mean F-score (0.278),

Fig. 7. Example face-cropped images showing the evolution of student expressions and gestures, with the corresponding problem outcomes. Top two rows present the student solved the problem on the first attempt (SOF), 3-4th rows present the student solved the problem with hints (SHINT), and the bottom two rows present the student tried but ultimately skipped the problem (GIVEUP).

TABLE IX
GENERALIZING TO UNSEEN USERS FROM THE MATHSPRING CHILDREN DATASET. RESULTS FOR PROBLEM OUTCOME PREDICTION (7-CLASSES) ON THE MATHSPRING CHILDREN DATASET (FIVE-FOLD CROSS-VALIDATION, LEAVE-USER-OUT SPLITS)

Method	Mean F-Score	Accuracy
Majority Vote Classifier	0.071	31.8%
Joshi et al. [7]	0.181	29.4%
ATL-BP w/o transfer learning	0.251	34.0%
ATL-BP w/ affect embedding	**0.270**	37.7%
ATL-BP w/ LIRIS children affect embedding	0.269	41.4%
ATL-BP w/ CAFE children affect embedding	0.268	40.6%
ATL-BP w/ LIRIS+CAFE affect embedding	0.263	**42.4%**

improving on the previous state-of-the-art [7] (13.2 pp absolute increase in accuracy and 38% relative increase in mean F-score). Leveraging extra children data further improves the mean F-score and accuracy over our original transfer learning model pretrained on AffectNet (5.6 pp increase in accuracy and 6.9% relative increase in mean F-score). For leave-user-out splits, the results also demonstrate that our model achieves an increase of 13 percent points in accuracy and 49% relative increase in mean F-score on the challenging task of predicting problem outcome using only student face movements and gestures. The prediction task has 7 classes which contributes to the difficulty.

D. Visual Examples

To visually illustrate our prediction of problem outcomes and understand student behavior, we present visual examples of an eighth grade student using MathSpring.

The student used MathSpring for one session of around 20 minutes and consented to have his face and screen recorded. Figure 7 shows the evolution of student expressions and gestures, and their corresponding problem outcomes. When the student successfully solves the problem on the first attempt (SOF), we can observe that he focus tightly on the problem during the period (first row). When he finally solve the problem correctly, he clenches his fist which indicates his excitement and passion (second row). When asking for hints, the student looks confused scratching his head but still engaged and actively attempts to solve the problem (rows 3–4). For the last problem (GIVEUP), the student gradually gets distracted and presents frustration and boredom (rows 5–6). These observations are consistent with our assumption that facial expressions and gestures provide important cues for inferring students' learning outcomes.

With our outcome prediction available in real time, teachers or intelligent tutors would be able to provide interventions and adjust learning schedules in time, to better help and assist students.

E. Learning Outcome Based on Affect and Engagement

In academic settings, emotion and engagement can be tightly correlated with learning outcomes and gains [32], [59], [60], [61]. For example, positive emotions enhance performance on tasks of problem solving [62], [63]. Emotions such as frustration, boredom, and anxiety negatively influence learning outcomes of students [43]. To explore the correlation of emotion, engagement, and learning outcomes,

TABLE X
STUDENT ENGAGEMENT DATASET SAMPLES PER CLASS. COLUMN (a) PRESENTS THE SAMPLES DISTRIBUTION FOR EACH CLASS IN THE REAL-WORLD RAW DATA. COLUMN (b) PRESENTS THE SAME DISTRIBUTION AFTER DOWN-SAMPLING AND BALANCING THE ORIGINAL DATASET. BOTH VERSIONS WILL BE MADE PUBLICLY AVAILABLE FOR NON-COMMERCIAL RESEARCH PURPOSES

Class	Original dataset (a)	Balanced dataset (b)
Paper	4,655	638
Screen	13,483	826
Wander	583	509
Total	18,721	1,973

we collected additional labels of student engagement on our MathSpringSP+ videos [47]. Specifically, we extract frames from videos in MathSpringSP+ dataset and annotate each frame with engagement labels (i.e., 'looking at their screen,' 'looking at their paper,' or 'wandering'). The task is then to classify each frame into one of three engagement categories.

1) Student Engagement Dataset: We selected 400 videos of 19 students in MathSpringSP+ dataset who consented to have their data publicly available for research, and sampled videos frames at one-frame-per-second (FPS). As a result, a total of 18,721 frames have been collected for engagement annotations. We used Amazon Mechanical Turk (MTurk) to label each frame with one of the following three categories: 'looking at their screen,' 'looking at their paper,' or 'wandering.' Each frame was assigned to three different crowdworkers, and we combined three crowdworker selections into a single label by majority vote.

The resulting dataset contains 18,721 annotated frames. However, the class distribution is quite unbalanced: the 'screen' class counts 22 times more samples than the 'wander' class and three times more samples than the 'paper' class (Table X (a)). After analyzing the distributions of the different samples for each class, we notice that the 'paper' and 'screen' classes contain a large number of similar frames. We therefore create a second smaller version of the original dataset by removing the similar samples for each class and balance the dataset. After selecting and removing the similar frames, we obtain a more equally distributed dataset, Table X (b), consisting of around 2,000 frame samples. Finally, we split the balanced dataset into a training and a testing set. For our test set, 20% of the samples were selected; the remaining 80% were used for training. In order to test and train the model on samples coming from different students, we chose the test samples from only three of the original 19 students.

2) Baseline Models: Given the collected, annotated, and balanced student engagement dataset, we then explore different baseline models to predict student engagement. We mainly compare two types of baselines: models based on deep convolutional networks, and models relying on head pose estimation.

Deep Convolutional Networks: We explored different deep convolutional neural networks for the task of classifying the frames. The architectures we used as the backbone model are: MobileNet [64], VGG16 [65] and Xception [66]. The backbone models were pre-trained on ImageNet [67]. On top of the pre-trained model, we added the following custom layers: one 2D global average pooling layer, one fully-connected layer with 128 neurons and ReLU activation, and a final output layer with three neurons and softmax activation. To avoid overfitting, we used multiple data augmentation techniques at the input layer (Gaussian noise, color channel changes, and cropping) and neurons drop-out at the head layers. We compared the performance of different models using the global and per-class accuracy scores. After training with frozen weights for the backbone, we fine-tuned the last layers of the backbone to achieve better accuracy (the number of layers fine-tuned depends on the model complexity).

Head Pose Estimation: The head pose is a 3-dimensional vector (i.e., yaw, pitch and roll) describing the rotation of the head in Euler angles. We utilize a state-of-the-art head pose estimation model to obtain accurate 3D head poses and infer students' engagement states based on values of head poses. This is intuitive as students' head poses will differ greatly when students are either looking at their screen, looking at their paper, or gaze wandering. While the eye gaze direction might provide a more accurate estimation of where the student is looking, it is more difficult to calculate especially when eyes are occluded. When students are looking at their paper or gaze wandering, their eyes could be fully occluded, making it impossible to calculate eye gaze directions accurately. Therefore, we choose to use head poses, which are highly correlated to gaze directions but are more robust and easier to compute, as indicators of students' engagement states. In addition, extracting a larger set of features (e.g., head pose, facial action unit, and eye gaze) might potentially boost the performance of this type of baseline. However, while this requires delicate feature selection and engineering, we would not expect a significant performance improvement. For simplicity, we use head pose only as a baseline for the task. Specifically, to estimate the head poses of students, we use a deep neural network FSA-Net [68] that predicts the head pose based on feature aggregation and regression. Given a facial image, detected and cropped using MTCNN [69], a deep cascaded multi-task face detector, FSA-Net combines feature maps from different layers by spatially grouping and aggregating features to harvest multi-scale information. The learned meaningful intermediate features are then used to perform soft stage-wise regression. Following Ruiz et al. [70], the head pose estimation model was pre-trained on the 300W-LP synthetic dataset [71] which contains 122,450 facial images with labelled head poses. The dataset synthesized faces across large poses (above 45°), ensuring that the trained model is robust to self-occlusion in our student dataset.

Given the predicted 3D head pose (yaw, pitch, and roll) for each image, we focus on two approaches for baseline classifiers. Our first method is a conditional approach with yaw and pitch head angles as the features. By inspecting and analyzing head pose angles for different classes, we design three conditions to distinguish head poses as either 'looking at their screen,' 'looking at their paper,' or 'wandering.' When students look at their paper, visible positive spikes in the pitch angle and a negative spike in the yaw angle could be observed. When students look at their screen, the yaw and pitch angles are neutral at around 0. Therefore, the conditions for the conditional

TABLE XI
RESULTS. GLOBAL ACCURACY SCORE OF PREDICTING STUDENT ENGAGEMENT USING DIFFERENT DEEP LEARNING AND HEAD POSE ESTIMATE APPROACHES. FROM THESE RESULTS WE CAN CONCLUDE THAT THE DEEP LEARNING MODELS ARE MORE SUITABLE FOR TASK OF CLASSIFYING STUDENT ENGAGEMENT COMPARED TO HEAD POSE ESTIMATORS. ALSO, DEPENDING ON THE COMPLEXITY OF THE DEEP LEARNING MODEL, WE REACH DIFFERENT ACCURACY SCORES, WITH THE BEST RESULTS OBTAINED BY THE MODEL WITH LESS COMPLEXITY, MOBILENET

Method	Accuracy (%)
MobileNet (pretrained ImageNet)	94
Xception (pretrained ImageNet)	88
VGG16 (pretrained ImageNet)	85
Head pose Estimator (Logistic Reg.)	60
Head pose Estimator (Conditional)	55

classifier are as follows 1) if the yaw angle is negative and the pitch angle is positive, we classify the set of angles as 'looking at their paper'; 2) if the yaw and pitch poses are both 0.0 ±0.05, we classify the set of angles as 'looking at their screen'; 3) if both conditions are not met, we classify the set of angles as 'wandering'. Our second approach uses the classical Logistic Regression to model the probability of a certain class. Each set of head angles (yaw and pitch of the student's head pose in a frame) corresponds to a data point with each data point being annotated as one of the three labels. We trained a 2-feature Logistic Regression classifier and each class was weighted with respect to the class size for balancing the dataset. Cross-Entropy loss was used as the loss function and Stochastic Average Gradient Descent as the optimizer.

3) Experimental Results: We here discuss the accuracy of predicting student engagement for different baselines (Table XI). The deep learning models show a range of results, 85%–94% accuracy, depending on the model size and number of parameters. It is important to notice that all the deep learning models reach similar performances when trained with frozen backbone weights (between 74–79% test accuracy), but they improve when we further fine-tune the backbone models. A smaller model such as MobileNet allows us to fine-tune more layers without overfitting, compared to deeper or larger models like VGG16 and Xception. This allows the MobileNet model to obtain a feature representation of the input images that is more relevant for this classification task, and by consequence this model reaches a higher final accuracy compared to the others. The results and training strategy may vary when we use different dataset configurations. We can also conclude from the results in Table XI that all convolutional neural networks significantly outperform the head pose estimation strategies. The reason for low accuracy scores of head pose estimation strategies could be the accumulated errors in the pipeline, such as errors in estimating head poses, and errors in designing conditional parameters or training of Logistic Regression. It is also possible that relying on head poses only is not sufficient to predict the engagement label. Further details on the per class accuracy for the best deep learning and head pose estimation models are given in Figure 8.

4) Discussion and Future Work: We have shown that our model can successfully predict student engagement. We suggest that the presented student engagement dataset and models enable future exploration of how affect, engagement, and

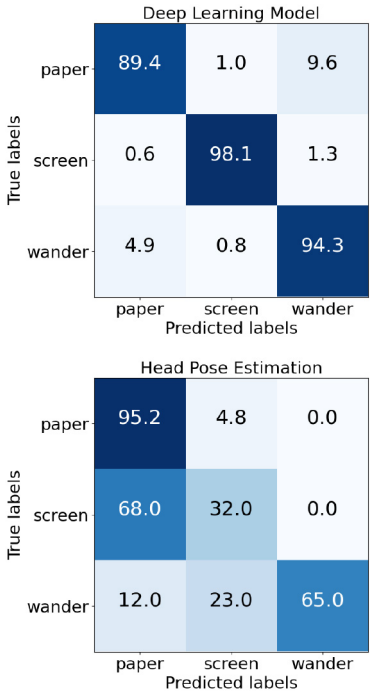

Fig. 8. Confusion Matrix Comparison. The head pose estimation model (bottom) obtains a lower per-class accuracy score, compared to the deep learning model solution MobileNet (top), which not only reaches an overall higher accuracy but also consistently classifies different classes.

learning outcomes correlate. While previous work often investigates how emotion and engagement impact learning outcomes, we are more interested in using learning outcomes as indicators of emotion and engagement. As we mentioned in the introduction, when solving problems, a smile does not necessarily mean a student is happy, but could mean the student is embarrassed for not knowing the answer to a question. Relying only on facial expressions and gestures may not be sufficient to infer a student's actual affect and engagement state. Learning outcomes could serve as an effective indicator of a student's real emotion and engagement. For example, positive outcomes (SOF, SHINT, ATT) could be indicative of positive learning activities, with paying attention and positive emotions, while negative outcomes (SKIP, GUESS, NOTR, GIVEUP) are indicative of negative learning activities, with inattentive states and negative emotions. With our publicly available MathSpringSP+ videos and annotations of student engagement and problem outcome, we facilitate future investigations of the correlations of affect, engagement and learning outcomes and how to utilize learning outcomes to help predict affect and engagement of students.

Additionally, while we consider exercise outcome indicative of a student's actual engagement or disengagement with a math problem solving activity, the main difficulty is that this classification label of the student-problem interaction cannot be predicted by our approach until the exercise is completed and the student clicks on "next problem" to request a new exercise. We acknowledge that it might be more beneficial for a student's learning experience if we were able to detect the student's disengagement (such as being able to understand that the student is looking away, and their attention has been lost) before the math problem solving activity has finished, so

that MathSpring could intervene, and potentially change the effort excerted on the exercise at hand. With our work, we can at least aim to change the student's interaction with the new exercise, a goal we will address in future work.

VI. Conclusion

We introduce a large labeled dataset of student interactions with an intelligent online math tutor that consists of 68 sessions, where 54 individual students solved 2,749 math problems. Using this dataset we design a transfer learning model ATL-BP that improves problem outcome predictions for students interacting with the ITS and answering math problems. By modeling the temporal structure of the videos with ATL-BP, we achieve a substantial increase in classification F-score and accuracy compared to previous state-of-the-art in this task. Additionally, using a novel affect representation along with user personalization, we achieve a further increase in performance. More generally, these promising results suggest that leveraging affect representations might be valuable in behavior analysis applications more generally. Our final method achieves a 50% relative increase in mean F-score as well as an absolute 11 percentage point increase in accuracy compared to previous work. We collected a dataset of children student interactions and present results on this dataset. We show that fine-tuning of the Affect network with age-appropriate images and video further improves performance in this scenario. These results pave the way for future improvements in solutions for this task. Finally, we present additional annotations of student engagement ('Looking at screen,' 'Looking at paper,' and 'Wandering'), which enable future explorations of correlations of learning outcomes, emotion, and engagement. Future tutor systems may use our proposed outcome and engagement prediction model in order to deliver real-time interventions to improve the learning of students.

Acknowledgment

The authors acknowledge funding for this research by the National Science Foundation, Grant 1551572, Grant 1551590, Grant 1551589, and Grant 1551594. The authors also thank the participants of their experimental studies.

References

[1] "MathSpring math tutor." https://www.mathspring.org. Accessed: Oct. 15, 2018. [Online]. Available: http://tutor.mathspring.org/ms/welcome.html

[2] I. Arroyo, B. P. Woolf, W. Burelson, K. Muldner, D. Rai, and M. Tai, "A multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect," *Int. J. Artif. Intell. Educ.*, vol. 24, no. 4, pp. 387–426, 2014.

[3] B. P. Woolf et al., "The effect of motivational learning companions on low achieving students and students with disabilities," in *Proc. Int. Conf. Intell. Tutoring Syst.*, 2010, pp. 327–337.

[4] I. Arroyo, D. G. Cooper, W. Burleson, B. P. Woolf, K. Muldner, and R. Christopherson, "Emotion sensors go to school," in *Proc. Conf. Artif. Intell. Educ. Build. Learn. Syst. Care Knowl. Represent. Affect. Model.*, 2009, pp. 17–24.

[5] S. D'Mello et al., "A time for emoting: When affect-sensitivity is and isn't effective at promoting deep learning," in *Proc. Int. Conf. Intell. Tutoring Syst.*, 2010, pp. 245–254.

[6] G. Gordon et al., "Affective personalization of a social robot tutor for children's second language skills," in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 3951–3957.

[7] A. Joshi et al., "Affect-driven learning outcomes prediction in intelligent tutoring systems," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, 2019, pp. 1–5.

[8] T. Almaev, B. Martinez, and M. Valstar, "Learning to transfer: Transferring latent task structures and its application to person-specific facial action unit detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3774–3782.

[9] J. Chen, X. Liu, P. Tu, and A. Aragones, "Person-specific expression recognition with transfer learning," in *Proc. 19th IEEE Int. Conf. Image Process.*, 2012, pp. 2621–2624.

[10] J. Chen, X. Liu, P. Tu, and A. Aragones, "Learning person-specific models for facial expression and action unit recognition," *Pattern Recognit. Lett.*, vol. 34, no. 15, pp. 1964–1970, 2013.

[11] E. Sangineto, G. Zen, E. Ricci, and N. Sebe, "We are not all equal: Personalizing models for facial expression analysis with transductive parameter transfer," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 357–366.

[12] G. Zen, L. Porzi, E. Sangineto, E. Ricci, and N. Sebe, "Learning personalized models for facial expression analysis and gesture recognition," *IEEE Trans. Multimedia*, vol. 18, no. 4, pp. 775–788, Apr. 2016.

[13] M. Shahabinejad, Y. Wang, Y. Yu, J. Tang, and J. Li, "Toward personalized emotion recognition: A face recognition based attention method for facial emotion recognition," in *Proc. 16th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, 2021, pp. 1–5.

[14] H. Kaya, F. Gürpınar, and A. A. Salah, "Video-based emotion recognition in the wild using deep transfer learning and score fusion," *Image Vis. Comput.*, vol. 65, pp. 66–75, Sep. 2017.

[15] M. Xu, W. Cheng, Q. Zhao, L. Ma, and F. Xu, "Facial expression recognition based on transfer learning from deep convolutional networks," in *Proc. 11th Int. Conf. Nat. Comput. (ICNC)*, 2015, pp. 702–708.

[16] X. Sun, J. Zeng, and S. Shan, "Emotion-aware contrastive learning for facial action unit detection," in *Proc. 16th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, 2021, pp. 1–8.

[17] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[18] I. Dua, A. U. Nambi, C. V. Jawahar, and V. Padmanabhan, "AutoRate: How attentive is the driver?" in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, 2019, pp. 1–8.

[19] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Sep. 2015, pp. 1–12. [Online]. Available: https://dx.doi.org/10.5244/C.29.41

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[21] N. Ruiz et al., "Leveraging affect transfer learning for behavior prediction in an intelligent tutoring system," in *Proc. 16th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, 2021, pp. 1–8.

[22] S. K. D'Mello, A. Olney, C. Williams, and P. Hays, "Gaze tutor: A gaze-reactive intelligent tutoring system," *Int. J. Human-Comput. Stud.*, vol. 7, no. 5, pp. 377–398, 2012. [Online]. Available: https://doi.org/10.1016/j.ijhcs.2012.01.004

[23] M. Mayo and A. Mitrovic, *Optimising ITS Behaviour With Bayesian Networks and Decision Theory*, Int. Artif. Intell. Educ. Soc., 2001.

[24] A. Mitrović and J. Holland, "Effect of non-mandatory use of an intelligent tutoring system on students' learning," in *Proc. Int. Conf. Artif. Intell. Educ.*, 2020, pp. 386–397.

[25] A. L. Mondragon, R. Nkambou, and P. Poirier, "Evaluating the effectiveness of an affective tutoring agent in specialized education," in *Proc. Eur. Conf. Technol. Enhanced Learn.*, 2016, pp. 446–452.

[26] S. Feng, A. J. Magana, and D. Kao, "A systematic review of literature on the effectiveness of intelligent tutoring systems in STEM," in *Proc. IEEE Front. Educ. Conf. (FIE)*, 2021, pp. 1–9.

[27] K. VanLehn, "The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems," *Educ. Psychol.*, vol. 46, no. 4, pp. 197–221, 2011.

[28] J. A. Kulik and J. Fletcher, "Effectiveness of intelligent tutoring systems: A meta-analytic review," *Rev. Educ. Res.*, vol. 86, no. 1, pp. 42–78, 2016.

[29] A. T. Corbett and J. R. Anderson, "Student modeling and mastery learning in a computer-based programming tutor," in *Proc. Int. Conf. Intell. Tutoring Syst.*, 1992, pp. 413–420.

[30] A. C. Graesser, K. VanLehn, C. P. Rosé, P. W. Jordan, and D. Harter, "Intelligent tutoring systems with conversational dialogue," *AI Mag.*, vol. 22, no. 4, p. 39, 2001.

[31] S. Amershi, C. Conati, and H. Maclaren, "Using feature selection and unsupervised clustering to identify affective expressions in educational games," in *Proc. Workshop Motivational Affect. Issues (ITS) 8th Int. Conf. (ITS)*, 2016, pp. 1–8.

[32] R. S. J. D. Baker, S. K. D'Mello, M. M. T. Rodrigo, and A. C. Graesser, "Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments," *Int. J. Human-Comput. Stud.*, vol. 68, no. 4, pp. 223–241, 2010.
[33] S. Craig, A. Graesser, J. Sullins, and B. Gholson, "Affect and learning: An exploratory look into the role of affect in learning with AutoTutor," *J. Educ. Media*, vol. 29, no. 3, pp. 241–250, 2004.
[34] S. D'Mello, R. W. Picard, and A. Graesser, "Toward an affect-sensitive AutoTutor," *IEEE Intell. Syst.*, vol. 22, no. 4, pp. 53–61, Jul./Aug. 2007.
[35] N. Jaques, C. Conati, J. M. Harley, and R. Azevedo, "Predicting affect from gaze data during interaction with an intelligent tutoring system," in *Proc. Int. Conf. Intell. Tutoring Syst.*, 2014, pp. 29–38.
[36] A. Kapoor, W. Burleson, and R. W. Picard, "Automatic prediction of frustration," *Int. J. Human-Comput. Stud.*, vol. 65, no. 8, pp. 724–736, 2007.
[37] S. Lallé, C. Conati, and R. Azevedo, "Prediction of student achievement goals and emotion valence during interaction with pedagogical agents," in *Proc. 17th Int. Conf. Auton. Agents MultiAgent Syst.*, 2018, pp. 1222–1231.
[38] R. Nkambou, "A framework for affective intelligent tutoring systems," in *Proc. 7th Int. Conf. Inf. Technol. Based High. Educ. Training*, 2006, pp. 2–8.
[39] B. D. Nye et al., "Engaging with the scenario: Affect and facial patterns from a scenario-based intelligent tutoring system," in *Proc. Int. Conf. Artif. Intell. Educ.*, 2018, pp. 352–366.
[40] R. W. Picard, "Affective computing: From laughter to IEEE," *IEEE Trans. Affect. Comput.*, vol. 1, no. 1, pp. 11–17, Jan. 2010.
[41] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 10, pp. 1175–1191, Oct. 2001.
[42] B. Woolf, W. Burleson, I. Arroyo, T. Dragon, D. Cooper, and R. Picard, "Affect-aware tutors: Recognising and responding to student affect," *Int. J. Learn. Technol.*, vol. 4, nos. 3–4, pp. 129–164, 2009.
[43] A. C. Strain and S. K. D'Mello, "Emotion regulation during learning," in *Proc. Int. Conf. Artif. Intell. Educ.*, 2011, pp. 566–568.
[44] S. D. Craig, A. C. Graesser, and R. S. Perez, "Advances from the office of naval research STEM grand challenge: Expanding the boundaries of intelligent tutoring systems," *Int. J. STEM Educ.*, vol. 5, no. 1, p. 11, 2018.
[45] S. Karumbaiah, R. Lizarralde, D. Allessio, B. P. Woolf, I. Arroyo, and N. Wixon, "Addressing student behavior and affect with empathy and growth mindset," in *Proc. 10th Int. Conf. Educ. Data Min. (EDM)*, 2017, pp. 96–103.
[46] Y. Kim, "Empathetic virtual peers enhanced learner interest and self-efficacy," in *Proc. Workshop Motivation Affect Educ. Softw. Conjunction 12th Int. Conf. Artif. Intell. Educ.*, 2005, pp. 9–16.
[47] K. Delgado et al., "Student engagement dataset," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3628–3636.
[48] H. Yu et al., "Measuring and integrating facial expressions and head pose as indicators of engagement and affect in tutoring systems," in *Proc. Int. Conf. Human-Comput. Interact.*, 2021, pp. 219–233.
[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[50] A. Mollahosseini, B. Hassani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, Jan.–Mar. 2019. [Online]. Available: http://arxiv.org/abs/1708.03985
[51] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Xi'an, China, May 2018, pp. 59–66. [Online]. Available: https://doi.org/10.1109/FG.2018.00019
[52] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, Dec. 2009.
[53] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6000–6010.
[54] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[55] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.
[56] S. Jastrzebski et al., "Finding flatter minima with SGD," in *Proc. ICLR*, 2018, pp. 1–4.
[57] R. A. Khan, A. Crenn, A. Meyer, and S. Bouakaz, "A novel database of children's spontaneous facial expressions (LIRIS-CSE)," *Image Vis. Comput.*, vols. 83–84, pp. 61–69, Mar./Apr. 2019.
[58] V. LoBue and C. Thrasher, "The child affective facial expression (CAFE) set: Validity and reliability from untrained adults," *Front. Psychol.*, vol. 5, p. 1532, Jan. 2015.
[59] S. D'Mello, B. Lehman, R. Pekrun, and A. Graesser, "Confusion can be beneficial for learning," *Learn. Instruct.*, vol. 29, pp. 153–170, Feb. 2014.
[60] R. M. Carini, G. D. Kuh, and S. P. Klein, "Student engagement and student learning: Testing the linkages," *Res. High. Educ.*, vol. 47, no. 1, pp. 1–32, 2006.
[61] J. Parsons and L. Taylor, "Improving student engagement," *Current Issues Educ.*, vol. 14, no. 1, pp. 1–32, 2011.
[62] J. D. Herrington et al., "Emotion-modulated performance and activity in left dorsolateral prefrontal cortex," *Emotion*, vol. 5, no. 2, pp. 200–207, 2005.
[63] A. M. Isen, K. A. Daubman, and G. P. Nowicki, "Positive affect facilitates creative problem solving," *J. Pers. Soc. Psychol.*, vol. 52, no. 6, pp. 1122–1131, 1987.
[64] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.
[65] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015, *arXiv:1409.1556*.
[66] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," 2017, *arXiv:1610.02357*.
[67] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[68] T.-Y. Yang, Y.-T. Chen, Y.-Y. Lin, and Y.-Y. Chuang, "FSA-net: Learning fine-grained structure aggregation for head pose estimation from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1087–1096.
[69] J. Xiang and G. Zhu, "Joint face detection and facial expression recognition with MTCNN," in *Proc. 4th Int. Conf. Inf. Sci. Control Eng. (ICISCE)*, 2017, pp. 424–427.
[70] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 2074–2083.
[71] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3D solution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 146–155.

Nataniel Ruiz (Graduate Student Member, IEEE) received the B.Sc. degree from Ecole Polytechnique, Paris, and the M.Sc. degree in computer science from the Georgia Institute of Technology. He is currently pursuing the Ph.D. degree with Boston University focusing on generative models, simulation, and facial analysis.

Hao Yu (Graduate Student Member, IEEE) received the B.Sc. degree in computer science from Zhejiang University, China. He is currently pursuing the Ph.D. degree with the Image & Video Computing Group, Boston University. His research interests include computer vision, human-computer interaction, and applications of machine learning, especially in facial image processing and analysis.

Danielle A. Allessio is a Postdoctoral Research Associate with the College of Information and Computer Sciences, University of Massachusetts, Amherst. Her research interests include intelligent tutoring systems, artificial intelligence, mathematics education, English language learner students, and animated pedagogical agent design.

Mona Jalal (Student Member, IEEE) received the double major master's degree in computer science and electrical engineering from the University of Wisconsin-Madison. She is currently a Graduate Research Fellow of Computer Science with a focus on Computer Vision, Boston University. Her research interests include computer vision, machine learning, and deep learning.

Ajjen Joshi received the B.A. degree from Connecticut College, the M.Sc. degree in computer science from Boston University, and the Ph.D. degree from the Department of Computer Science, Image and Video Computing Research Group, Boston University. He is currently a Senior Research Scientist with Affectiva (acquired by Smart Eye in June 2021). His research interests include computer vision, machine learning, and human-computer interaction.

Tom Murray is a Research Scientist and a author working in the fields of advanced educational technology, online social deliberative skills, and adult developmental theory.

John J. Magee (Member, IEEE) received the B.A. degree in computer science and mathematics from Boston College and the M.A. and Ph.D. degrees in computer science from Boston University. He is an Associate Professor and the Chair of the Department of Mathematics and Computer Science with Clark University. His primary research interests include computer vision, human-computer interaction, accessible computing, and assistive technology.

Kevin Manuel Delgado received the B.Sc. degree in computer science from Boston University in 2021, where he is currently a Data Scientist. His research interests include computer vision and machine learning.

Vitaly Ablavsky (Senior Member, IEEE) received the Ph.D. degree in computer science from Boston University. He is currently a Principal Research Scientist with the Applied Physics Laboratory, University of Washington. His interests include computer vision, machine learning, and autonomous systems.

Stan Sclaroff (Fellow, IEEE) received the Ph.D. from the Massachusetts Institute of Technology in 1995. He is currently a Professor of Computer Science with Boston University and the Dean of Arts and Sciences. His research interests include computer vision, pattern recognition, and machine learning. He is a Fellow of the IAPR.

Ivon Arroyo received the M.Sc. degree in computer science and the Ed.D. degree in education from the University of Massachusetts Amherst, Amherst, where she is an Associate Professor of Computer Science and Education. She has carried out research with the forefront of education, computer science, and psychology, authoring over 100 research articles in the three disciplines.

Beverly P. Woolf received the Ph.D. degree in computer science and the Ed.D. degree in education from the University of Massachusetts Amherst. She has more than 20 years of experience in educational computer science research, production of intelligent tutoring systems, and development of multimedia systems.

Sarah Adel Bargal received the Ph.D. degree from the Department of Computer Science, Boston University in 2019. She is an Assistant Professor and a Provost's Distinguished Faculty Fellow with the Department of Computer Science, Georgetown University. Her research interests are in machine learning, computer vision, and explainable artificial intelligence, with a current focus on making artificial intelligence systems explainable and accountable to humans and society.

Margrit Betke (Senior Member, IEEE) received the Ph.D. degree in computer science and electrical engineering from the Massachusetts Institute of Technology in 1995. She is a Professor of Computer Science with the Computer Science Department, Boston University, where she co-leads the Artificial Intelligence Research Initiative and the Image and Video Computing Research Group. She conducts research in computer vision, human-computer interfaces, medical image analysis, and application of machine learning.