Contents lists available at ScienceDirect

Computers & Education: X Reality

journal homepage: www.journals.elsevier.com/computers-and-education-x-reality

Automated emotion recognition of students in virtual reality classrooms

Michael Shomoye, Richard Zhao [*]

Department of Computer Science, University of Calgary, Calgary, Alberta, T2N 1N4, Canada

ARTICLE INFO

Keywords:
Emotion recognition
Facial expression
Virtual reality
Virtual learning environment

ABSTRACT

In contemporary educational settings, understanding and assessing student engagement through non-verbal cues, especially facial expressions, is pivotal. Such cues have long informed educators about students' cognitive and emotional states, assisting them in tailoring their teaching methods. However, the rise of online learning platforms and advanced technologies such as virtual reality (VR) challenge the conventional modes of gauging student engagement, especially when certain facial features become obscured or are entirely absent. This research explores the potential of Convolutional Neural Networks (CNNs), specifically a custom-trained model adapted from the ResNet50 architecture, in recognizing and distinguishing subtle facial expressions in real-time, such as neutrality, boredom, happiness, and confusion. The novelty of our approach is twofold: First, we optimize the power of CNNs to analyze facial expressions in digital learning platforms. Second, we innovate for the context of VR by focusing on the lower half of the face to tackle occlusion challenges posed by wearing VR headsets. Through comprehensive experimentation, we compare our model's performance with the default ResNet50 model and evaluate it against full-face and VR-occluded face datasets. Ultimately, our endeavor aims to provide educators with a sophisticated tool for real-time evaluation of student engagement in technologically advanced learning environments, subsequently enriching the teaching and learning experience.

1. Introduction

Over the years, educational methodologies have evolved from traditional chalkboard-based instruction to technologically advanced learning environments. Amidst these transformations, comprehending student engagement continues to be a paramount concern for educators. Past research indicated a correlation between student engagement and academic performance. (Chukwuemeka et al., 2022).

In classrooms where both students and teachers are physically present, teachers have had various ways to assess the students' engagement. One of the traditional methods has been through observing non-verbal cues, especially facial expressions. Teachers have long relied on these cues as indicators of student's cognitive and emotional states (Ekman & Friesen, 1971). Students' faces give teachers a lot of information which helps them adjust their teaching methods depending on how students react to the material being taught. However, with the rise of online education, this direct observation is often limited or non-existent, especially when students choose to keep their cameras off or face connectivity issues (Sitzmann et al., 2006).

Two expressions often encountered by teachers, "boredom" and "neutrality," can be challenging to distinguish but are crucial for effective teaching (D'Mello et al., 2012; Kunter et al., 2008). While the former may indicate disengagement, the latter could suggest calm attentiveness. Our work focuses on the limitation posed by missing facial cues in virtual education settings. We employ Convolutional Neural Networks (CNNs), leveraging the ResNet50 architecture (He et al., 2016), to accurately analyze facial expressions such as neutrality and boredom without starting from scratch.

In the era of technological advancements, virtual reality (VR) has the potential to reshape many aspects of our everyday lives. Both educational technologists and pedagogical theorists are currently exploring how VR can create captivating and effective learning environments. Over time, VR has transformed from an entertainment and gaming concept into a tool with wide ranging implications across industries such as healthcare, manufacturing, and notably education (Zhao et al., 2019). With VR, learners can now engage with 3D environments where abstract concepts can be brought to life. Imagine a biology student walking through the system of the body, or a history student freely wandering around an ancient civilization while interacting with its culture. This feeling of being present in these worlds could potentially revolutionize how we engage in education by making it an experiential and successful process. While researchers have studied the impacts of VR in fields such

[*] Corresponding author.
E-mail addresses: michael.shomoye@ucalgary.ca (M. Shomoye), richard.zhao1@ucalgary.ca (R. Zhao).

https://doi.org/10.1016/j.cexr.2024.100082
Received 25 October 2023; Received in revised form 20 May 2024; Accepted 31 August 2024
Available online 6 September 2024

as engineering, gaming and medicine, its effectiveness in educational settings is still ongoing research.

As we envision the future, we expand our research to explore the integration of VR in educational settings, such as a virtual classroom where all students join through wearing a VR headset. VR integration in education is tantalizing and growing in popularity because it provides immersive teaching and learning experiences. However, it also presents challenges such as barriers, potential health effects and difficulties with analyzing expressions due to obstructions caused by VR headsets (Radianti et al., 2020). Therefore, this research does not only focus on highlighting the benefits, but also addresses the obstacle of occlusion when analyzing users' facial expressions (Fig. 1). To address this issue, our CNN model has been cleverly designed to focus on the visible half of the face — the lower half. This adaptation ensures that our custom-trained model remains practical and relevant for analyzing facial expressions in these advanced immersive environments.

Our research covers the model's architecture, development, experiments, and results. We aim to offer educators a tool for real-time facial expression analysis during lectures, providing reports without storing student images, ensuring privacy (De la Cruz, 2022).

2. Related literature

Keeping track of student's reactions, emotions and engagements on online learning platforms can be quite challenging. It deprives teachers of the real time feedback mechanism that allows them to adapt their teaching strategies on the spot, potentially creating a disconnect between instruction and student's needs. There have been efforts to address these limitations by using artificial intelligence and machine learning algorithms to analyze how engaged students are in classrooms (D'Mello & Graesser, 2015). These technologies are still in development, as they cannot fully replicate the depth of human understanding that comes from direct observation.

While it is well documented that online learning environments have limitations, especially when it comes to supporting nuanced verbal cues, it is important to recognize that virtual classrooms cause more than just negative effects on education. For example, Ally (2004) mentioned the flexibility and accessibility provided by virtual platforms can often outweigh the advantages of classrooms, particularly in situations where geographic or physical limitations pose challenges to conventional learning. Yildirim et al. (2020) delved into the perspectives of pre-service teachers regarding the incorporation of VR as a resource. Both pre-service teachers and school-based teacher educators acknowledge the value of VR in teacher training curricula (Cooper et al., 2019).

Recently, VR has been used as a tool in games - serious games, for the purpose of improving the educational experience for students. Zhao et al. (2020) focused on developing a VR game specifically designed for manufacturing education. They aim to explore the potential of VR technology in improving manufacturing education by creating an interactive learning experience. Furthermore, they evaluate its effectiveness based on user feedback and assessment of learning outcomes. The research shed light on how the VR game affects students learning experience and knowledge retention, in manufacturing education.

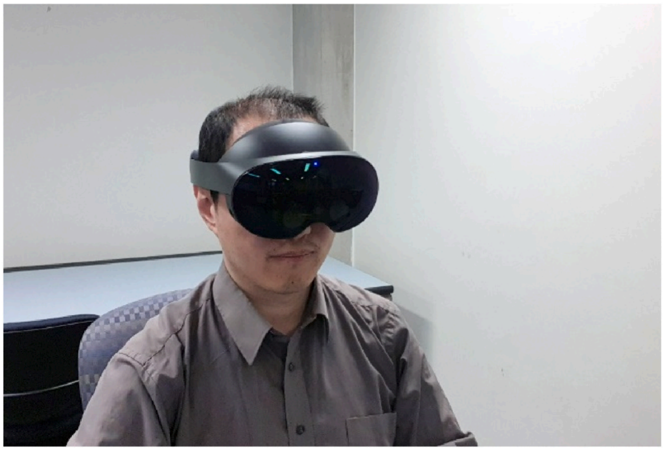

Fig. 1. A person wearing a VR headset (Meta Quest Pro), showing occlusion in the upper half of the face. This is generally the case for any currently available VR headsets in the market.

Traditional face to face classrooms has benefited from the nuanced interaction of signals and expressions between teachers and students. For instance, when a teacher notices confusion or disengagement among students through their expressions, they may take it as a sign to pause, revisit complex concepts, directly ask if further clarification is needed or provide examples. Disregarding these cues could lead to a learning environment where students either mentally disengage or go through heightened levels of anxiety and frustration, which could potentially affect students' grades at the end of a session.

VR-based classrooms also impact student engagement and motivation, resulting in increased interest and active participation in science lessons. Kasapakis et al. (2023) delved into the benefits of incorporating VR technology into education specifically focusing on the impact of realistic nonverbal cues on the overall learning experience. As a result, they recommended that educators and designers consider incorporating these cues into VR based educational experiences to optimize learning outcomes. Similarly, Liu et al. (2020) focused on investigating the impact of a VR based classroom on students' performance in science lessons. To conduct their experiment, they divided students into two groups: one receiving classroom instruction, while the other experiencing instruction within a VR based classroom. The outcome of their work shows that students in a virtual classroom perform better than those who received instructions in physical education class.

Researchers have examined the relationship between facial expressions displayed by users in VR and their emotional states. One study found that emotions expressed by a virtual face were recognized in a comparable way as emotions expressed by natural facial expressions (Dyck et al., 2008). Dubovi (2023) showed that in a virtual learning environment, emotions could be detected through facial expressions. For instance, joy was found to be the predominant emotion during virtual learning, but personality traits and emotion regulation strategies also influenced the expressions of negative emotions such as anger and sadness. Furthermore, a data-driven approach using interactive VR games and collected multimodal measures (self-reports, physiological and facial signals) from participants was used to better understand emotions' underlying processes (Somarathna & Mohammadi, 2024). The results showed the role of different components in emotion differentiation, with the model including all components demonstrating the most significant contribution. These findings have implications for using VR environments in emotion research and highlight the role of physiological signals in emotion recognition within such environments. Marín-Morales et al. (2020) presented a systematic review of the emotion recognition research undertaken with physiological and behavioral measures as elicitation devices and showed that emotions, such as arousal, anxiety, stress, and fear, could be measured in VR. A system designed to track and mirror facial expressions in VR users found that it could accurately reconstruct high-fidelity facial expressions and provide insights into the user's emotional states (Lou et al., 2019).

Half-face facial recognition in the context of VR is an ongoing challenging task due to the occlusion caused by wearing a VR headset. Mills and Cleary (2022) employed deep learning techniques and RGBD imagery of participants wearing VR headsets with obscured faces to analyze emotions. They compared this group with a control group composed of obscured faces. They developed a custom model for emotions and occluded faces recognition. Additionally, another research (Houshmand & Khan, 2020) focuses on recognizing expressions despite occlusion caused by VR headsets. In their work, they used a transfer learning method. They started by using pre-trained networks, such as

VGG and ResNet, which were then fine-tuned using facial expression recognition datasets, such as FER+ and RAF DB. They had a better result with their model compared to the default benchmark datasets.

The effectiveness of ResNet50 in facial recognition was explored in a study by Patel et al. (2019), highlighting its robustness in various challenging environments. The researchers utilized ResNet50 to train on a dataset consisting of thousands of images captured under lighting conditions, angles and expressions. By comparing it with other CNN architectures, they found that ResNet50 demonstrated performance in terms of both speed and accuracy, making it an excellent choice for their facial recognition tasks. Similarly, Nyarko et al. (2022) carried out a comprehensive analysis by comparing ResNet50, AlexNet and Inception-V3 architectures on masked face recognition. Their work highlighted ResNet50's superiority over other deep learning architectures in terms of both accuracy and computational efficiency. However, while the potential of ResNet50 in facial recognition is evident, challenges persist. Hwang et al. (2023) study shed light on potential vulnerabilities in ResNet50-based facial recognition systems, emphasizing the need for adversarial training to bolster security.

3. Research methodology

3.1. Data collection: the cornerstone of machine learning

In machine learning, especially for applications as nuanced and complex as facial recognition, data serves as the foundational layer upon which the entire architecture is built. Recognizing the intricate nature of facial expressions and the challenges they present in educational settings, we collected data from two different sources: a public dataset called AffectNet and a dataset that was captured by our research team to mimic a VR-based classroom setting.

3.2. AffectNet dataset

AffectNet is a large-scale, publicly available dataset specifically designed for facial expression recognition in the wild, meaning the images are not confined to controlled environments but reflect real-world conditions. It contains over one million facial images collected from the Internet, annotated for eleven different facial expressions: happiness, sadness, surprise, neutral, etc. (Fig. 2) The strengths of using AffectNet as a source include its large sample size and diversity, encompassing different ages, genders, and ethnicities, thereby making the model more generalizable. However, it is important to note that while AffectNet provides a broad scope, it may not be specifically tailored for educational settings, especially VR-based educational settings. For example, the boredom expression which is important for our study is missing in the AffectNet dataset. We have therefore collected a second dataset.

3.3. Custom dataset and its collection

To address the limitations of the AffectNet dataset and to make our model more attuned to the specificities of a VR-based educational environment, a second dataset was captured by our research team.

3.3.1. User study design and accessibility

We designed user study materials as video presentations with clear instructions. The video utilized text-to-speech instructions generated through lovo.ai, directing participants on the required facial expressions. To cater to those with hearing impairments, subtitles were added using the "VideoProc Vlogger" software. Each expression was timed for 5 s, introduced by a countdown. To facilitate natural and authentic facial expressions, specific background visuals and music tracks were chosen for each expression type. For example, a cheerful background coupled with the lively tune was used to elicit the happiness expression. Conversely, a plain, dull background along with sleep-inducing music was selected to evoke expressions of boredom, which ensured genuine expressions and an engaging user experience.

Participants for our study were students from the Computer Science department at our institution, contacted through a dedicated mailing list. The recruitment email introduced the study's objectives, an e-poster, participation incentives, and expectations. Our inclusive approach invited a diverse demographic, and we recruited 30 participants, offering a broad dataset for our facial expression model. We asked participants to self-report demographical information. Among our participants, the average age is 26.9. 16 participants identified as male, 12 identified as female, 1 identified as non-binary, and 1 was undisclosed. Ethnically, 2 identified as East Asian, 6 identified as Black, 8 identified as South Asian, 12 identified as West Asian (Middle Eastern), and 2 identified as Latino.

3.3.2. User study setup

We configured a lab space to meet the requirements of the study, focusing on both technical and comfort aspects. A computer played the instructional video, while a Nikon camera captured the facial expressions. The setup focused on recording detailed expressions vital for training our machine learning models. We ensured participants knew they could opt out anytime, adhering to ethical standards. Each participant was recorded individually, focusing on four facial expression classes: Neutral, Boredom, Confusion, and Happiness.

3.3.3. Data processing: from videos to useable data

We gathered data in video format, capturing all four facial expressions per participant, resulting in 30 comprehensive videos. These were subsequently divided according to expression classes, totaling 120 videos. They were recorded at 23.98 fps but were down-sampled to 15 fps to reduce computational strain. After converting these videos into

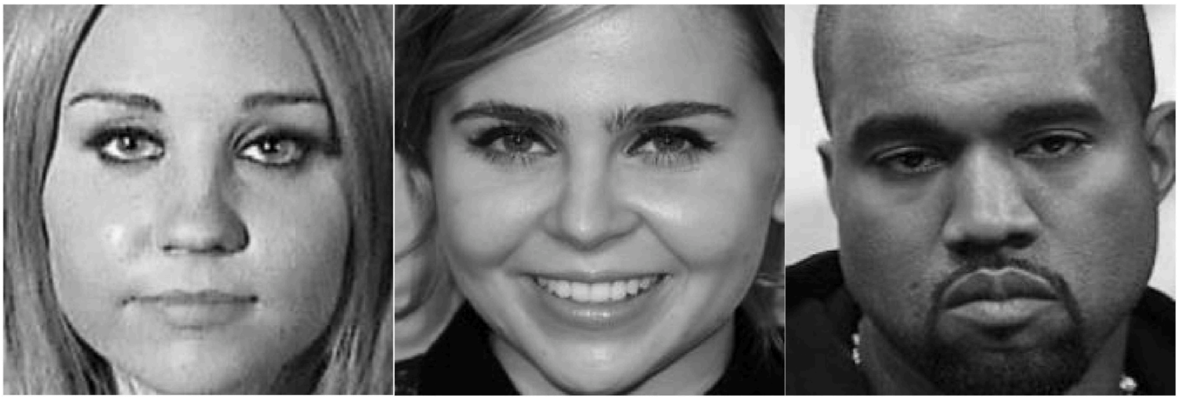

Fig. 2. Sample images from the AffectNet dataset, showing different emotions: neutral, happiness, and boredom, respectively.

image frames, we reviewed and discarded unsuitable frames where the images were blurry, resulting in 9000 useable frames.

We utilized the “dlib” Python library, a facial landmark detector identifying 68 key face coordinates (x, y) (Fig. 3). This pre-trained model is highly accurate in pinpointing features such as eyes, nose, and mouth.

3.4. *Processing for full face recognition and facial occlusion*

To target only facial regions, we cropped bounding boxes around each face, specifically, we set the x-axis to 20 and the y-axis to 25 extracting the full face akin to passport-sized photos. This dataset serves as the foundation for our full-face recognition model, emphasizing facial expressions by removing unrelated data.

Considering the relevance of VR in education, our research focuses on expressions when the upper face is obscured by a VR headset. Thus, we processed data to highlight the lower face, the area visible even when a VR headset is worn. Using different cropping parameters “x-axis at 3 and the y-axis at 15”, we emphasized this lower region for our occlusion-specific model. Conclusively, all cropped images, both full and occluded half-face, were resized to 244x244 pixels, ensuring they fit the parameters required by our training model.

3.5. *Custom-trained ResNet50*

In our quest to build a specialized deep learning model capable of distinguishing between nuanced facial expressions, particularly in the context of online education and VR classrooms, we opted for a custom-trained ResNet50 architecture. ResNet50, a convolutional neural network with 50 layers, has been proven to perform exceptionally well in image classification tasks. However, to better suit the specificities of our study, certain modifications were imperative.

Efficiency and reproducibility are two pillars of any scientific research. Towards that, several key parameters were set before training the model. After experimenting with different batch sizes, we found that a batch size of 16 yielded the best results in terms of accuracy and computational efficiency. All our models were trained for 100 epochs. This number was found to be sufficient for the model to converge without overfitting. Setting the random seed for both NumPy and TensorFlow to 42 ensures that the experiments are reproducible.

3.5.1. *Model architecture: ResNet50 base*

For the feature extraction part of our model, we leveraged the pre-trained ResNet50 architecture. The input shape is set to accommodate 64x64 images with 3 color channels (RGB). We excluded the fully connected layers at the end of the pre-trained model to adapt the model for our custom classification layer. The model used pre-trained weights from the ImageNet database for quicker and more effective training.

3.5.2. *Layer freezing and unfreezing strategy*

In transfer learning, it is often beneficial to freeze the weights of the pre-trained model during the initial phase of training. This ensures that the useful features captured by the pre-trained model are not distorted. We froze all layers except for the last one. This approach was taken to avoid overfitting, which becomes more likely as more layers are unfrozen and fine-tuned.

The ResNet50 architecture, inherently comprising multiple residual blocks, was further fine-tuned. Specifically, we unfroze the layers in blocks 4 and 5 for training, while keeping the preceding layers frozen. This localized training enabled the model to learn higher-level features pertinent to our specific use-case, without altering the generic feature-maps learned initially. This method focused training on the deeper layers responsible for higher-level feature extraction, offering the dual benefits of time-efficiency and specialized learning. Fig. 4 shows the entirety of our custom-trained ResNet50 architecture.

The challenge of overfitting—where the model performs exceedingly well on training data but poorly on unseen data—needed addressing. To mitigate overfitting, dropout layers were strategically placed in our neural network, following both the base ResNet50 model and the subsequent dense layer. In each dropout layer, a percentage of the layer’s input units are randomly set to zero during training, which helps to prevent any single neuron from becoming overly specialized. Specifically, we used a dropout rate of 0.5 following the base model and 0.6 after the dense layer. These rates were determined experimentally to maximize model generalization without significantly compromising training accuracy.

Additionally, L2 regularization was applied to the dense layers in the model. Regularization techniques such as L2 are commonly employed to constrain the optimization process, discouraging the learning algorithm from fitting the training data too closely. L2 regularization adds a penalty term to the loss function based on the magnitude of the layer weights. In our case, we used a regularization factor of 0.3, which was empirically found to be effective in reducing overfit.

4. Experimental results

Evaluation metrics are crucial for assessing the performance of machine learning models, particularly in classification tasks where the prediction of categories or classes is involved. These metrics provide a quantitative basis for comparing models, understanding their strengths and weaknesses, and guiding the improvement of algorithms.

Our experiments can be broadly categorized into two segments. The first segment is dedicated to the training of different neural networks on

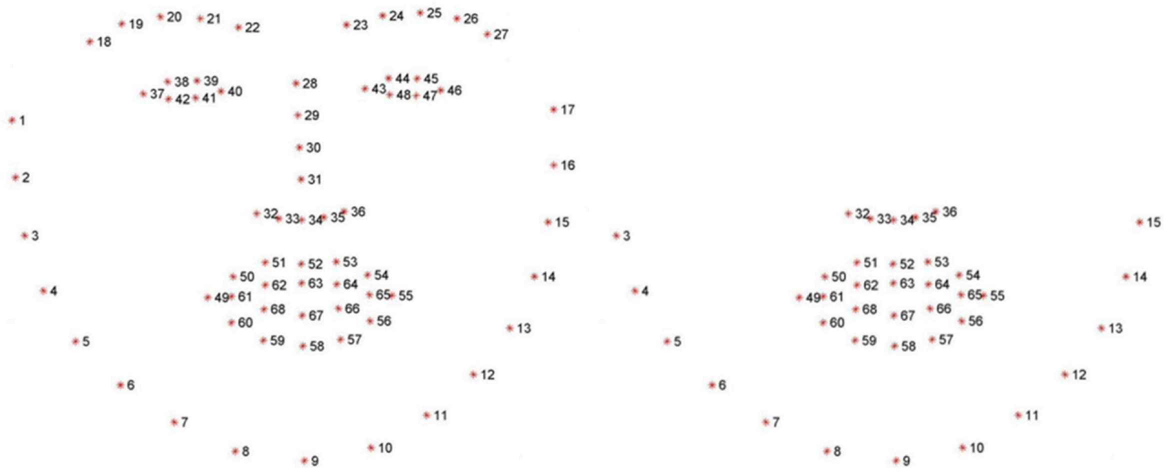

Fig. 3. Facial feature identification from the dlib Python library, showing full face and lower-half face identifications. Image adapted from Sagonas et al. (2013).

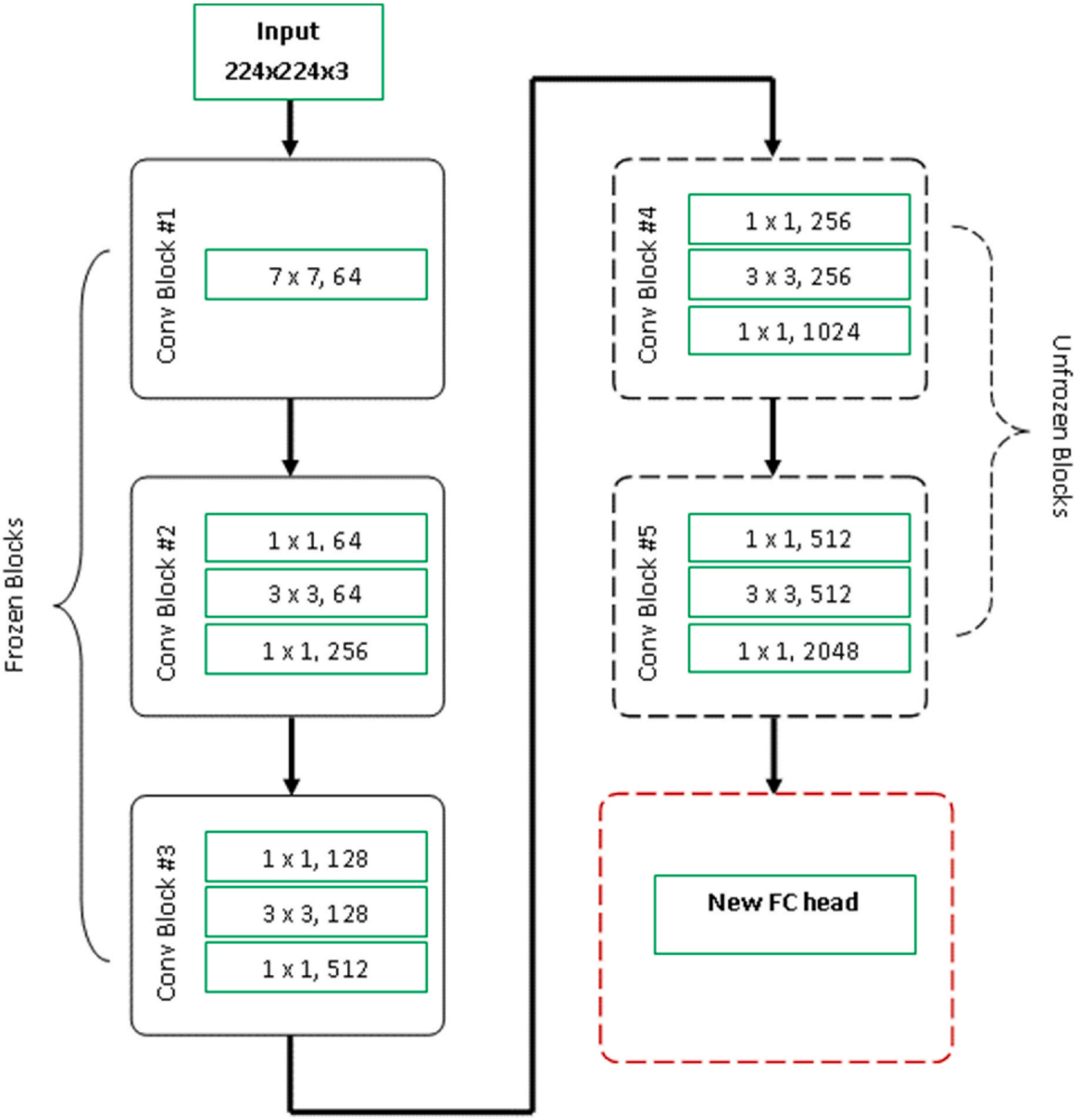

Fig. 4. Block diagram of the proposed custom-trained ResNet50 architecture.

datasets. Following the training process, we saved the learned weights of each model for use in the second segment, which focuses on prediction tasks using the model unseen datasets. The four sets of experiments are summarized in Table 1.

4.1. *Experiment 1: comparative analysis of ResNet50, VGG19, and MobileNet*

The goal of this experiment is to determine the effective neural network architecture for transfer learning in our facial recognition research. We chose common image classification architectures for comparisons based on recent works (Jahnavi et al., 2023; Mascarenhas & Agarwal, 2021; Ikechukwu et al., 2021; Mohapatra et al., 2021). Although there are performance statistics for ResNet50 (He et al., 2016), VGG19, which is an updated architecture of VGG (Simonyan & Zisserman, 2015), and MobileNet (Howard et al., 2017), these existing metrics are general and do not accurately represent how these models perform in the specific scenario of emotion detection in VR-occluded faces. Hence, we conducted this comparative analysis. Tables 2 and 3 show the results.

ResNet50: We chose ResNet50 because of its depth and the presence of residual connections, which have been proven to aid in solving the vanishing/exploding gradient problem in deep networks. Additionally, ResNet50 is widely adopted in the research community.

VGG19: We selected VGG19 based on its reputation for performance in image classification tasks and its straightforward architecture, which makes it easier to understand and modify.

MobileNet: We included it in our experiment because it serves as the architecture for Google Teachable Machine (Carney et al., 2020). This architecture is specifically designed to be lightweight and efficient, making it highly suitable for real time applications or situations where computational resources are limited.

All three networks were trained using our dataset consisting of four classes: “Neutral”, “Boredom”, “Confusion”, and “Happiness.” We conducted the trainings over 100 epochs to ensure that the model reached a level of convergence. We also extended our experiment to account for the occluded "half-face" scenario, simulating the conditions under which a portion of the face is obscured when a student is wearing a VR headset.

The results indicated a clear preference for the ResNet50 architecture

Table 1
Summary of our four experiments.

Experiments	Model Comparison	Dataset Used	Facial Expression Classes	Dataset for Comparison
Experiment 1	Default ResNet vs. VGG19 vs. MobileNet	Our specialized dataset (training and prediction)	Neutral, Boredom, Confusion & Happiness	Full Face and Half Face
Experiment 2	Default ResNet50 vs Custom-ResNet50	AffectNet (training) + Our specialized dataset (prediction)	Neutral, Boredom & Happiness	Full Face and Half Face
Experiment 3	Default ResNet50 vs Custom-ResNet50	Our specialized dataset (training and prediction)	Neutral, Boredom, Confusion & Happiness	Full Face and Half Face
Experiment 4	Custom-ResNet50	Our specialized dataset (training and prediction)	Boredom vs Happy Boredom vs Neutral Boredom vs Confusion.	Full Face and Half Face

Table 2
Experiment 1 training results. (Full face = F, half face = H).

Model	Accuracy	Precision	Recall	F1 Score
ResNet50	F: 0.7986 H: 0.8059	F: 0.6646 H: 0.6807	F: 0.3821 H: 0.4211	F: 0.4879 H: 0.5160
VGG19	F: 0.7573 H: 0.7638	F: 0.5301 H: 0.5530	F: 0.2582 H: 0.2871	F: 0.3428 H: 0.3730
MobileNet	F: 0.7796 H: 0.7921	F: 0.6066 H: 0.6415	F: 0.3364 H: 0.3821	F: 0.4287 H: 0.4760

Table 3
Experiment 1 prediction results.

Model	Precision	Recall	F1 Score
ResNet50	F: 0.5370 H: 0.4845	F: 0.5332 H: 0.4579	F: 0.5294 H: 0.4619
VGG19	F: 0.3030 H: 0.3836	F: 0.2972 H: 0.3752	F: 0.2870 H: 0.3753
MobileNet	F: 0.2719 H: 0.3545	F: 0.2455 H: 0.3032	F: 0.2130 H: 0.2385

in terms of both accuracy and generalization when applied to our dataset. This was consistent across both the full-face and occluded half-face conditions. This justifies our choice in using ResNet50 as the base architecture for our customization.

4.2. Experiment 2: effectiveness when testing on a different dataset

In the second experiment, we aimed to evaluate the generalizability and adaptability of our custom ResNet50 model by training it on an online dataset and test its effectiveness with our own dataset, which is a different dataset that more closely resembles a VR classroom setting. This would be similar to real-life use cases of teachers utilizing such a system in virtual classrooms where the model was pre-trained on a different dataset. We contrasted this with the performance of the default ResNet50 model. This was conducted to test the robustness of our custom-trained model in working with new, unseen data. Tables 4 and 5 show the results.

Table 4
Experiment 2 training results. (Full face = F, half face = H).

Model	Accuracy	Precision	Recall	F1 Score
Default ResNet50	F: 0.7543 H: 0.7159	F: 0.6678 H: 0.5971	F: 0.5231 H: 0.4537	F: 0.5823 H: 0.5117
Custom-trained ResNet50	F: 0.8041 H: 0.8045	F: 0.7574 H: 0.7604	F: 0.6063 H: 0.6037	F: 0.6739 H: 0.6716

Table 5
Experiment 2 prediction results.

Model	Data Class	Precision	Recall	F1 Score
Default ResNet50	Boredom	F: 0.43 H: 0.44	F: 0.46 H: 0.58	F: 0.44 H: 0.50
	Neutral	F: 0.39 H: 0.47	F: 0.51 H: 0.59	F: 0.44 H: 0.52
	Happiness	F: 0.69 H: 0.88	F: 0.43 H: 0.36	F: 0.53 H: 0.51
	Average	**F: 0.5033** **H: 0.5948**	**F: 0.4666** **H: 0.511**	**F: 0.47** **H: 0.5113**
Custom-trained ResNet50	Boredom	F: 0.61 H: 0.49	F: 0.46 H: 0.61	F: 0.52 H: 0.54
	Neutral	F: 0.66 H: 0.46	F: 0.72 H: 0.45	F: 0.69 H: 0.46
	Happiness	F: 0.76 H: 0.76	F: 0.82 H: 0.60	F: 0.79 H: 0.67
	Average	**F: 0.6766** **H: 0.57**	**F: 0.6666** **H: 0.5533**	**F: 0.6666** **H: 0.5566**

We trained both the default ResNet50 and our custom-trained version on the AffectNet dataset. Then we used these trained models to make predictions on our own dataset. The performance metrics show that our custom-trained ResNet50 model performed better than the default ResNet50 model in making accurate predictions on our specialized dataset. This highlights the versatility and generalizability of our custom-trained model when faced with new data.

4.3. Experiment 3: default ResNet50 vs custom-trained ResNet50

Our third experiment aimed to evaluate the performance of our custom model using our specialized dataset. This was similar to Experiment 2, but in this case, we partitioned our specialized dataset into a training set, a validation set, and prediction set, focusing on both full-face and occluded half-face images. The models were then trained, validated, and predicted using these sets. Tables 6 and 7 show the results. Our custom-trained model again outperformed the default ResNet50 model, irrespective of the full face or half face context.

4.4. Experiment 4: binary classification experiments involving boredom

The fourth experiment aimed to test the discriminatory power of our custom-trained model, specifically in differentiating “Boredom” from other expressions such as “Neutral,” “Happiness,” and “Confusion.” This is of particular interest as the expression for “Boredom” often closely resembles “Neutral,” making it a challenging task. We configured the models to perform binary classifications. We set up three separate scenarios: Boredom vs Neutral, Boredom vs Happiness, and Boredom vs Confusion. The experiment was conducted using both full-face and occluded half-face datasets. Tables 8 and 9 show the results. The result of this experiment provided intriguing insights. Not only did our custom model manage to effectively differentiate “Boredom” from other

Table 6
Experiment 3 training results. (Full face = F, half face = H).

Model	Accuracy	Precision	Recall	F1 Score
Default ResNet50	F: 0.7986 H: 0.8059	F: 0.6646 H: 0.6807	F: 0.3821 H: 0.4211	F: 0.4879 H: 0.5160
Custom-trained ResNet50	F: 0.8496 H: 0.9111	F: 0.7717 H: 0.8946	F: 0.5661 H: 0.7304	F: 0.6502 H: 0.8033

Table 7
Experiment 3 prediction results.

Model	Data Class	Precision	Recall	F1 Score
Default ResNet50	Boredom	F: 0.45 H: 0.50	F: 0.61 H: 0.51	F: 0.52 H: 0.51
	Confusion	F: 0.47 H: 0.33	F: 0.37 H: 0.50	F: 0.41 H: 0.40
	Neutral	F: 0.56 H: 0.48	F: 0.46 H: 0.34	F: 0.51 H: 0.40
	Happiness	F: 0.66 H: 0.62	F: 0.69 H: 0.49	F: 0.68 H: 0.55
	Average	**F: 0.5370** **H: 0.4845**	**F: 0.5332** **H: 0.4579**	**F: 0.5294** **H: 0.4619**
Custom-trained ResNet50	Boredom	F: 0.71 H: 0.58	F: 0.51 H: 0.61	F: 0.59 H: 0.60
	Confusion	F: 0.74 H: 0.45	F: 0.70 H: 0.57	F: 0.72 H: 0.50
	Neutral	F: 0.63 H: 0.75	F: 0.82 H: 0.64	F: 0.71 H: 0.69
	Happiness	F: 0.80 H: 0.77	F: 0.85 H: 0.65	F: 0.82 H: 0.70
	Average	**F: 0.7224** **H: 0.6379**	**F: 0.7177** **H: 0.6145**	**F: 0.7125** **H: 0.6217**

Table 8
Experiment 4 training results. (Full face = F, half face = H).

Data Classes	Accuracy	Precision	Recall	F1 Score
Boredom vs Neutral	F: 0.9036 H: 0.8654	F: 0.9007 H: 0.8654	F: 0.9071 H: 0.8729	F: 0.9012 H: 0.8646
Boredom vs Happiness	F: 0.9614 H: 0.9621	F: 0.9641 H: 0.9723	F: 0.9586 H: 0.9514	F: 0.9577 H: 0.9603
Boredom vs Confusion	F: 0.8407 H: 0.8600	F: 0.8383 H: 0.8510	F: 0.8443 H: 0.8729	F: 0.8368 H: 0.8598

Table 9
Experiment 4 prediction results.

Data Class	Precision	Recall	F1 Score
Boredom vs Neutral	F: 0.6596 H: 0.6442	F: 0.653 H: 0.641	F: 0.6493 H: 0.6389
Boredom vs Happiness	F: 0.8432 H: 0.8102	F: 0.8380 H: 0.8085	F: 0.8373 H: 0.8082
Boredom vs Confusion	F: 0.6471 H: 0.6373	F: 0.647 H: 0.63	F: 0.6468 H: 0.6249

expressions, but it also achieved a reasonable level of accuracy across both full-face and occluded half-face conditions.

4.5. *Performance of the models in prediction time*

A critical aspect of evaluating the efficacy of different Convolutional Neural Network models was not only their accuracy in emotion recognition but also the performance of their prediction processes. The effectiveness of the ResNet50, VGG19, and MobileNet models in this research was evaluated based on their total prediction time when given our entire custom dataset. Table 10 presents a comparative overview of each model's prediction time, reflecting their capacity to process and analyze facial expressions. Each model was given our custom-captured dataset of 9000 useable frames. This measure directly impacts the feasibility of integrating the models into VR environments where immediate feedback and interaction are paramount. While MobileNet had the least prediction time (as it was designed to be simple and fast), the performance of ResNet50 still provided real-time capabilities of per-frame prediction necessary for such a system in a virtual classroom setting.

Table 10
Total prediction time results.

Models	ResNet50	VGG19	MobileNet
Total Prediction Time	2321 s	2296 s	652 s

4.6. *Discussions*

The importance of any research lies in its ability to systematically address and provide insights into the questions that form its foundation. In this section, we delve into an analysis of the experimental results and provide a link between the framework of our research and its practical implications, ensuring a coherent narrative flow, from identifying problems to exploring potential solutions.

This research provides a detailed comparative analysis of ResNet50 with VGG19 and MobileNet, focusing on their performance in facial recognition in VR classroom settings. From our experiments, ResNet50 exhibited superior performance in terms of accuracy and generalization when applied to both full-face and VR-occluded face datasets. This outcome highlights that while VGG19 and MobileNet have their own merits, ResNet50's architecture, particularly its depth and residual connections, makes it more suitable for the complexities involved in VR-based emotion recognition.

Our work also explored the effectiveness of a customized training approach to CNN adapted from the ResNet50 architecture in recognizing and differentiating subtle facial expressions, such as neutrality, boredom, happiness, and confusion, in VR classroom settings. This effectiveness is evident in Sections 4.2 and 4.3, where our model outperformed the default ResNet50 model in accuracy and adaptability in both full-face and occluded scenarios. In Section 4.4, we also demonstrated how we were able to test our customized training approach model alongside boredom and other expressions such as neutrality, happiness, and confusion, yielding decent results. Our approach to modifying the ResNet50 model, specifically targeting the lower half of the face to address occlusion challenges, appears to be a key factor in its success.

The limitations of current emotion recognition technologies in VR environments in this research revolve around occlusion challenges and data diversity, including the absence of boredom facial expressions, which were missing in previous studies. Boredom is one of the essential facial expressions in educational settings. Our custom-trained model addresses these challenges by focusing on the lower half of the face, which remains visible even with a VR headset occlusion. Occlusion caused by VR headsets significantly affects the accuracy of facial expression recognition. This research incorporated techniques that improved recognition performance for the visible portions of the face by focusing on the lower half of the face and employed data augmentation strategies to simulate what a VR classroom setting would look like.

5. Conclusion

Understanding student engagement through non-verbal cues such as facial expressions is crucial in contemporary educational settings. With the advent of online learning and VR technologies, conventional methods of gauging these cues are being tested, as facial features can be obscured or completely absent when students participate remotely wearing VR headsets. This research presents a novel approach leveraging a custom-trained CNN model adapted from the ResNet50 architecture for recognizing subtle facial expressions when students are learning using VR technology. Our custom-trained model uniquely addresses two key challenges: the capability to analyze facial expressions in digital platforms and the ability to focus on the lower half of the face, overcoming the occlusion issues presented by VR headsets. Our experiments examined the identification of emotions such as boredom, happiness, and confusion, all central to an educational setting. We aim to empower educators with the means to evaluate student engagement in real-time in modern, technologically advanced learning environments, thus enriching both teaching and learning experiences. As the field of education undergoes the transition from traditional to digital methodologies, our model stands as a beacon, exemplifying how technology can seamlessly bridge this gap. As educational technology continues to adapt and evolve, we remain optimistic that such technological advancements will pave the way for richer, more insightful, and inclusive learning experiences for all.

5.1. Applications in future VR classrooms

The advent of VR in educational settings has opened new vistas for immersive and interactive learning experiences. The integration of emotion recognition technologies, particularly the custom ResNet50 model developed in this research, holds significant potential in revolutionizing these virtual learning spaces. Our model can be seamlessly integrated into a future VR classroom, enhancing the teaching-learning dynamic through real-time emotional feedback or post-class facial expression analysis.

In a future VR classroom, each student, equipped with a VR headset, is immersed in a 3D educational environment, alongside their instructor. The students and their instructor can be physically sitting at any location on the planet, but they see each other's virtual avatars in a virtual classroom, sitting together. Integrated within each VR headset is a compact camera system, discreetly positioned to capture the students' facial expressions without causing any discomfort or distraction. As the lesson progresses, our model embedded within the classroom's educational software analyzes the facial expressions of each student in real-time or captures their facial expressions for post-class analysis. By focusing on the lower half of the face, the model overcomes the occlusion challenge posed by the VR headsets. It identifies subtle expressions indicative of emotions such as confusion, boredom, neutrality, and happiness.

When done in real-time, the model's output is relayed to the instructor's VR interface, where a dashboard displays an aggregated emotional overview of the class. This immediate feedback allows the instructor to adjust the pace, content, and delivery of the lesson dynamically. For instance, if a significant number of students show signs of confusion or boredom, the instructor can introduce interactive elements, revisit complex topics, or initiate a Q&A session. When the analysis is done post-class, the system captures students' facial expressions for subsequent analysis, which the instructor can access after class. The resulting report provide a comprehensive view of each student's emotional responses in summary format. This data then enables the instructor to pinpoint moments during the class when a notable increase in boredom or disengagement occurred, and therefore able to adjust the teaching plan in the future.

This integration of emotion recognition technology cultivates an empathetic educational environment. Students, aware that their emotional feedback is valued and acted upon, feel more connected and engaged. Instructors, equipped with insights into their students' emotional states, can foster a more inclusive and responsive learning atmosphere. The envisioned application of the custom model in VR classrooms paves the way for a new era in digital education. This technology, with its potential to understand and respond to students' emotional states, can significantly enhance the effectiveness of VR-based learning.

Our system has several limitations, particularly its emphasis on specific facial expressions. Additionally, our training dataset was collected under controlled lighting conditions, which may not reflect real-world scenarios. Future research should broaden the scope of facial expression analysis to encompass various environmental factors, such as fluctuating lighting conditions and changing distances between the camera and the user, although the latter can be mitigated with head-mounted cameras. Furthermore, the social dimension is crucial for comprehending the emotional dynamics of human beings (Marín-Morales et al., 2020), necessitating an examination of the effects of social interactions in VR settings on emotional states.

Statements on open data and ethics

This study has been reviewed and approved by the Research Ethics Board (REB) at the University of Calgary. All procedures were performed in compliance with relevant laws and institutional guidelines, and the appropriate institutional committee(s) have approved them. Participants provided informed consent for participating.

Participants were protected by having their personal information removed in the study, and this research only published aggregate data. Participation in the study was voluntary and participants knew that they could withdraw from the study at any time during data collection and their data would be immediately deleted. The privacy rights of human subjects within this study are observed through secure storage of all data associated with this study in an appropriate protected repository.

The AffectNet dataset is an open dataset available for non-commercial use.

CRediT authorship contribution statement

Michael Shomoye: Data curation, Methodology, Software, Writing – original draft, Investigation. **Richard Zhao:** Conceptualization, Funding acquisition, Investigation, Methodology, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported and funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) under grant number RGPIN-2022-03178.

References

Ally, M. (2004). Foundations of educational theory for online learning. *Theory and Practice of Online Learning, 2*(1), 15–44.

Carney, M., Webster, B., Alvarado, I., Phillips, K., Howell, N., Griffith, J., … Chen, A. (2020). Teachable machine: Approachable Web-based tool for exploring machine learning classification. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems* (pp. 1–8).

Chukwuemeka, U. K., Obayi, A. A., Abiodun, O. S., Chukwunike, A. K., & Anwar, C. (2022). *An enhanced student engagement and academic performance predictive system*. https://doi.org/10.51583/ijltemas.2023.12506

Cooper, G., Park, H., Nasr, Z., Thong, L. P., & Johnson, R. (2019). Using virtual reality in the classroom: Preservice teachers' perceptions of its use as a teaching and learning tool. *Educational Media International, 56*(1), 1–13.

D'Mello, S., & Graesser, A. (2015). Feeling, thinking, and computing with affect-aware learning technologies. In R. A. Calvo, S. D'Mello, J. Gratch, & A. Kappas (Eds.), *The oxford handbook of affective computing* (pp. 419–434). Oxford University Press.

D'Mello, S., Taylor, R. S., Davidson, K., & Graesser, A. C. (2012). Monitoring affective trajectories during complex learning. In *Proceedings of the 5th international conference on educational data mining* (pp. 320–323).

De la Cruz, J. (2022). Online class: Student data privacy. *International Journal of Advance Research in Computer Science and Management Studies, 10*(7).

Dubovi, I. (2023). Facial expressions capturing emotional engagement while learning with desktop VR: The impact of emotional regulation and personality traits. *Interactive Learning Environments*, 1–17.

Dyck, M., Winbeck, M., Leiberg, S., Chen, Y., Gur, R. C., et al. (2008). Correction: Recognition profile of emotions in natural and virtual faces. *PLoS One, 3*(11), Article e3798.

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology, 17*(2), 124–129.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Houshmand, B., & Khan, N. M. (2020). Facial expression recognition under partial occlusion from virtual reality headsets based on transfer learning. In *2020 IEEE sixth international conference on multimedia big data (BigMM)* (pp. 70–75). IEEE.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Hwang, R. H., Lin, J. Y., Hsieh, S. Y., Lin, H. Y., & Lin, C. L. (2023). Adversarial patch attacks on deep-learning-based face recognition systems using generative adversarial networks. *Sensors, 23*(2), 853.

Ikechukwu, A. V., Murali, S., Deepu, R., & Shivamurthy, R. C. (2021). ResNet-50 vs VGG-19 vs training from scratch: A comparative analysis of the segmentation and classification of pneumonia from chest X-ray images. *Global Transitions Proceedings, 2* (2), 375–381.

Jahnavi, K., Battu, V. S. J., Sandeep, N. S., Anitha, R., Deepika, R., & Prakash, K. B. (2023). Detection of COVID-19 using ResNet50, VGG19, MobileNet, and forecasting; using logistic regression, prophet, and SEIRD model. In *2023 7th international conference on computing methodologies and communication (ICCMC)* (pp. 1538–1542). IEEE.

Kasapakis, V., Dzardanova, E., & Agelada, A. (2023). Virtual reality in education: The impact of high-fidelity nonverbal cues on the learning experience. *Computers & Education: X Reality, 2*, Article 100020.

Kunter, M., Tsai, Y. M., Klusmann, U., Brunner, M., Krauss, S., & Baumert, J. (2008). Students' and mathematics teachers' perceptions of teacher enthusiasm and instruction. *Learning and Instruction, 18*(5), 468–482.

Liu, R., Wang, L., Lei, J., Wang, Q., & Ren, Y. (2020). Effects of an immersive virtual reality-based classroom on students' learning performance in science lessons. *British Journal of Educational Technology, 51*(6), 2034–2049.

Lou, J., Wang, Y., Nduka, C., Hamedi, M., Mavridou, I., Wang, F. Y., & Yu, H. (2019). Realistic facial expression reconstruction for VR HMD users. *IEEE Transactions on Multimedia, 22*(3), 730–743.

Marín-Morales, J., Llinares, C., Guixeres, J., & Alcañiz, M. (2020). Emotion recognition in immersive virtual reality: From statistics to affective computing. *Sensors, 20*(18), 5163.

Mascarenhas, S., & Agarwal, M. (2021). A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Classification. *2021 International conference on disruptive technologies for multi-disciplinary research and applications (CENTCON), 1*, 96–99. IEEE.

Mills, I., & Cleary, F. (2022). Facial Emotion recognition analysis using deep learning through RGB-D imagery of VR participants through partially occluded facial types. In *2022 IEEE conference on virtual reality and 3D user interfaces abstracts and workshops (VRW)* (pp. 862–863). IEEE.

Mohapatra, S., Abhishek, N. V. S., Bardhan, D., Ghosh, A. A., & Mohanty, S. (2021). Comparison of MobileNet and ResNet CNN architectures in the CNN-based skin cancer classifier model. *Machine Learning for Healthcare Applications*, 169–186.

Nyarko, B. N. E., Bin, W., Zhou, J., Agordzo, G. K., Odoom, J., & Koukoyi, E. (2022). Comparative analysis of Alexnet, resnet-50, and inception-V3 models on masked face recognition. In *2022 IEEE world AI IoT congress (AIIoT)* (pp. 337–343). IEEE.

Patel, H., Pandya, H., Theckedath, D., & Kini, R. (2019). Facial affect detection using transfer learning: A comparative study. *PsyArXiv*.

Radianti, J., Majchrzak, T. A., Fromm, J., & Wohlgenannt, I. (2020). A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda. *Computers & Education, 147*, Article 103778.

Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2013). 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE international conference on computer vision workshops* (pp. 397–403).

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *3rd international conference on learning representations (ICLR 2015)*. Computational and Biological Learning Society.

Sitzmann, T., Kraiger, K., Stewart, D., & Wisher, R. (2006). The comparative effectiveness of web-based and classroom instruction: A meta-analysis. *Personnel Psychology, 59* (3), 623–664.

Somarathna, R., & Mohammadi, G. (2024). Exploring emotions in multi-componential space using interactive VR games. *arXiv preprint arXiv:2404.03239*.

Yildirim, B., Topalcengiz, E. S., Arikan, G., & Timur, S. (2020). Using virtual reality in the classroom: Reflections of STEM teachers on the use of teaching and learning tools. *Journal of Education in Science Environment and Health, 6*(3), 231–245. https://doi.org/10.21891/JESEH.711779

Zhao, R., Aqlan, F., Elliott, L. J., & Baxter, E. J. (2020). Multiplayer physical and virtual reality games for team-based manufacturing simulation. *ASEE virtual annual conference*, 2020.