Facial Expression Modeling and Synthesis for Patient Simulator Systems: Past, Present, and Future

MARYAM POUREBADI and LAUREL D. RIEK, Computer Science and Engineering, UC San Diego

Clinical educators have used robotic and virtual patient simulator systems (RPS) for dozens of years, to help clinical learners (CL) gain key skills to help avoid future patient harm. These systems can simulate human physiological traits; however, they have static faces and lack the realistic depiction of facial cues, which limits CL engagement and immersion. In this article, we provide a detailed review of existing systems in use, as well as describe the possibilities for new technologies from the human–robot interaction and intelligent virtual agents communities to push forward the state of the art. We also discuss our own work in this area, including new approaches for facial recognition and synthesis on RPS systems, including the ability to realistically display patient facial cues such as pain and stroke. Finally, we discuss future research directions for the field.

CCS Concepts: • **Computer systems organization → Robotics**; • **Applied computing → Health informatics**; **Interactive learning environments**; • **Computing methodologies → Simulation by animation**; **Appearance and texture representations;**

Additional Key Words and Phrases: Human robot interaction, healthcare robotics, Clinical simulators, health-care training and education, facial expression synthesis, facial modeling, neurological impairment, social robotics

ACM Reference format:
Maryam Pourebadi and Laurel D. Riek. 2022. Facial Expression Modeling and Synthesis for Patient Simulator Systems: Past, Present, and Future. *ACM Trans. Comput. Healthcare* 3, 2, Article 23 (February 2022), 32 pages.
https://doi.org/10.1145/3483598

1 INTRODUCTION

For more than five decades, researchers in the field of **Human–Robot Interaction (HRI)** have been building and studying how robots can collaborate with humans, support them with their work, and assist them in their daily lives [92, 101, 160]. For example, autonomous mobile robots work side by side with skilled human workers in factories and retail sectors [165]. Social robots inform and guide passengers in large and busy airports [187]. In both clinical and home settings, robots have been used to assist healthcare workers, clean rooms, ferry supplies, and support people with disabilities and older adults in rehabilitation and task assistance [160].

There is emerging interest in using robotics technology to address key challenges in healthcare, particularly those related to the quality, safety, and cost of care delivery. However, there are several key contextual challenges to realizing this vision. One big concern is the rapidly increasing costs of healthcare. For example, in the United States, healthcare is expensive across a range of services including administrative costs, pharmaceutical spending, individual services, and the use of high-income trained healthcare workers [12]. Another challenge is the

This material is based upon work supported by the National Science Foundation under Grant No. IIS-1253935.
Authors' address: M. Pourebadi and L. D. Riek, Computer Science and Engineering UC San Diego 9500 Gilman Drive, MC 0404 La Jolla, CA 92093; emails: {pourebadi, lriek}@eng.ucsd.edu.

2637-8051/2022/02-ART23 $15.00
https://doi.org/10.1145/3483598

Fig. 1. A typical patient simulation center setup. Clinical learners treat a non-expressive robotic patient simulator. Its physiology is controlled by a clinical educator.

dynamic nature of clinical environments with occupational hazards that put healthcare workers at risk of injury and disability [93, 182, 183]. Additionally, the global shortfall in professional healthcare workers with sufficient clinical education and skills is challenging [203].

Providing healthcare systems with robots may help address these gaps. For example, robots can support the independence of people with disabilities by enabling transitions to home-based care. Robots can also help clinicians and caregivers with care tasks including physical, cognitive, and manipulation tasks [20, 88, 95, 160, 198], as well as healthcare worker education (see Figure 1).

Robots can potentially enable healthcare workers to spend more time with patients and less time engaging in "non-value added" physical tasks, and reduce the errors caused by the overburden of these tasks [88, 182]. These physical tasks include transportation, inventory, and spending time searching and waiting [160]. For example, Tug robots [14] are medical transportation robots that autonomously move through hospitals, delivering supplies, meals, and medication to patients.

Moreover, robots can assist in clinical learning. For example, humanoid patient simulators can mimic human function (physiology) or anatomy (biology). Some of these simulators are engineered systems that model information integration and flow to help clinical learners study human physiology. Others present models of human patient biology and cognition to provide clinicians with a platform to practice different skills including task execution, testing and validation, diagnosis and prognosis, training, and social and cognitive interaction.

Robotic patient simulators (RPSs), **virtual patient simulators (VPSs)**, and **augmented reality patient simulators (APSs)** are three main technologies used to represent realistic, expressive patients within the context of clinical education. **Clinical educators (CE)** can use them to convey realistic scenarios, and **clinical learners (CL)** can practice different procedural and communication skills without harming real patients.

Although there are many benefits associated with using RPS, VPS, and APS systems, their designs suffer from a lack of **facial expressions (FEs)**, which are both a key social function and clinical cue conveyed by real patients. While enabling RPS and VPS systems with an expressive face can address this challenge, still it creates a bigger challenge with designing expressive systems: Facial expressions are very person-dependent and can vary from person to person [212]. It is challenging to analyze, model, and synthesize FEs of a small subgroup of patients

on simulators' faces and develop generalized expressive simulator systems that are capable of representing a diverse group of patients (including but not limited to different ages, genders, and ethnicities who are affected by different diseases and conditions) [212].

Another challenge is that incorrectly (or not) exhibiting symptoms on a simulator's face may reinforce incorrect skills in CLs and could lead to future patient harm. Furthermore, developers may face physical limitations preventing them from advancing the state of the art. For example, VPSs are limited by flat two-dimensional (2D) display mediums, making them unable to represent a physical 3D human-shape volume that clinicians can palpate to perform clinical assessments. Other challenges include the simulator's usability, controllability, high costs, and physical limitations, as well as the need of recruiting experts with various skills.

Tackling these technical challenges to advance the state-of-the art needs work on several fronts. These include the creation of capable and usable RPS and VPS systems, new techniques for recognizing and synthesizing facial expressions on simulators, novel computational methods for developing humanlike face model for them, and new means for evaluating these systems. Ultimately, addressing these gaps can provide healthcare education with realistic, expressive simulators capable of mimicking patientlike expressions. This has the potential to positively affect CLs' retention, and eventually, revolutionize healthcare education.

In this review, we discuss research at the intersection of robotics, computer vision, and clinical education to enable socially interactive robots and virtual agents to simulate human-patient-like expressions and interact with real humans. In Section 2, we provide an overview of the root causes of preventable patient harm and contextualize clinical education as a means for addressing it. We outline common learning modalities, including VPS and RPS systems, and outline key opportunities to improve them. Sections 2.4–5 discuss the importance of incorporating humanlike FEs in RPS and VPS systems and algorithmic approaches for doing so. In Section 6, we discuss our recent research on creating expressive VPS and RPS systems, with diverse appearances and features, which show promise as an important clinical education tool. Finally, Sections 7 and 9 explore open problems in the field and discuss new directions for future work.

2 BACKGROUND

2.1 Patient Safety and Healthcare Education

The World Health Organization defines patient safety as "the absence of preventable harm to a patient during the process of healthcare and reduction of risk of unnecessary harm associated with healthcare to an acceptable minimum" [34]. Taking an action (errors of omission) or inaction (errors of commission) by healthcare workers, system failures, or a combination of these two factors may cause or lead to preventable patient harm [98].

Preventable patient harm represents the root cause of many adverse events experienced in healthcare departments including intensive care units and is a leading cause of mortality and morbidity in the world. Conservative estimates suggest preventable patient harm causes over 400,000 preventable deaths per year in the U.S. hospitals alone [106], and 4–8 million experience serious harm and injury. It is estimated that between 27% and 33% of patients experience an adverse event as a result of their care [9, 69, 175, 190].

While better-designed healthcare systems, services, and processes, as well as new technologies, can help reduce the incidence of patient harm, in the short term one of the best approaches is high-quality clinical education. Recent work shows that healthcare education and training is the most effective mechanism to reduce the incidence of patient harm and improve patient safety [166].

One way advance the state of the art of healthcare education is through the development intelligent learning modalities, such as simulation systems. Simulators provide CLs the chance to safely study the causes and effects of errors, while avoiding harm to real patients. Using simulators also improves CLs' comprehension, confidence, efficiency, and enthusiasm for learning [107]. When compared with non-digital learning methods, using patient simulators can more effectively improve CLs' skills and at least as effectively improve knowledge [112].

Table 1. Simulators: The Structure, Functionality, and Controlability

Type	Medium / Platform	Physiological variables	Visual appearance	Control	Scheduling time
SHPs	Real: 3D real-human body	Can present some of the variables.	Can display dynamic FEs, gestures, and some of the abnormal visual findings.	Controlled by a human.	High
APSs	Hybrid: Visual appearance projected to a 3D physical surface.	Can easily present all the variables.	Can be programmed to richly display dynamic FEs, gestures, and all abnormal visual findings.	Ranges from fully automated to teleoperated to pre-recorded mode.	Low
VPSs	Virtual: 2D monitor or TV or Tablet	Can only present the visual physiological variables due to 2D display limitations.	Can be programmed to richly display dynamic FEs, gestures, and all abnormal visual findings.	Ranges from fully automated to teleoperated to pre-recorded mode.	Low
RPSs	Mchanical: 3D humanlike physical robot	Can exhibit 5000+ physiology changes on it. Verbal responses controlled using a live operator.	Mostly have a static face. They can be programmed to display some of dynamic FEs, gestures, and abnormal visual findings.	Ranges from fully automated to teleoperated to pre-recorded mode.	Low

Clinical educators may also benefit from using simulation systems to run a variety of desired clinical simulation scenarios on realistic patients based on a learner's need, instead of patients' availability. Examples of these scenarios include nursing simulation scenarios [10], physician scenarios [32], and surgical simulation scenarios [33]. Studies also indicate that using simulation improves the performance of learner evaluation and educational needs diagnosis by CEs [42]. This work, and others, are encouraging and suggest that augmenting existing healthcare simulation systems with emerging AI-based technologies offers promising opportunities to substantially reduce preventable patient harm, as well as risks to clinicians.

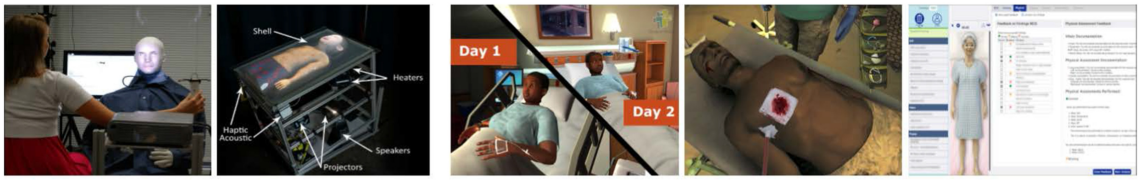
2.2 Patient Simulator Types, Benefits, and Challenges

There are four types of simulated patients used in simulation-based clinical learning: standardized human patients, augmented reality patient simulators, virtual patient simulators, and robotic patient simulators. Table 1 illustrates the structure, functionality, and controllability for each type of patient simulator. This is further discussed below.

Standardized human patients (SHPs) are live actors who assume the roles of patients. They convey a series of symptoms and/or a scenario defined by CEs [47]. SHPs are beneficial, as they provide CLs with a real-human case study to practice their history-taking and clinical assessment skills. As a result, SHPs enable the learning process to sometimes deviate from a predefined senario, as this type of simulator can adapt to unexpected changes on the fly.

However, SHPs cannot accurately exhibit many symptoms of real patients, such as facial paralysis or physiological changes. Furthermore, recruiting SHPs can be difficult and expensive, especially ones at younger ages because of child labor laws and scheduling difficulties [41, 47, 91, 188].

APSs, also known as physical-virtual simulators, use **augmented reality (AR)** techniques to combine physical human-shaped surfaces with dynamic visual imagery projected on its surface [72]. APSs combine the benefits of two worlds: Its physicality can convey a realistic, embodied similarity to people, while its virtual component can display dynamic appearances and FEs without being limited by hardware infrastructure.



(a) (b) (c) (d) (e)

Fig. 2. Left: APSs with rendered faces based on projector placement. (a) An APS system with FPI [163], and (b) An APS system with rear-projected imagery [71]. Right: Examples of VPSs software with rendered faces. (c) Shadow Health [35], (d) CliniSpace [8], and (e) i-Human [30].

However, it is still challenging to display an accurate representation of naturalistic symptoms even in an AR environment. APSs also present some challenges depending on the AR modalities and techniques used. Recent work [63] suggests to avoid the use of commercially available head-mounted displays for augmented reality surgical interventions, because perceptual issues can affect user performance. In **front-projected imagery (FPI)** [163], the shadow of users can hover over the projection [201] and cause the CEs fail to display desired scenarios. Rear-projected imagery [71] can solve both multi-user and projection occlusion problems; however, it requires a sufficient physical space behind the augmented platform to place the projectors [72] (see Figure 2, left).

VPS are interactive digital simulations of real patients in clinical settings displayed on a screen (see Figure 2, left). For example, the Shadow Health VPS keeps CLs engaged with digital patients and lets them practice communication skills, assessing virtual patients, and documenting their findings [35]. CliniSpace offers both a standalone healthcare education system and a fully immersive game [8]. i-Human VPS agents are capable of presenting human physiology and pathophysiology, as well as 3D anatomy of the human body [30]. Gabby is a VPS system that provides support to African-American women to decrease their preconception health risks and eliminate racial and ethnic disparities in maternal and child health [52, 193].

VPSs benefit from virtually portraying physiological variables (e.g., heart rate) without being limited by hardware infrastructure. The virtual display also provides the opportunity to richly and quickly display changes in the appearance, symptoms, behavior, or body language. Furthermore, Kononowicz et al. [112] found that VPS systems can help improve knowledge and skill-building (e.g., clinical reasoning, procedural, and teamwork skills) when compared with non-digital educational methods, including didactic-learning modalities (e.g., lectures, reading exercises, group discussion in the classroom), and non-digital models such as SHPs. Another advantage to VPSs is that they make clinical education more accessible to CLs in low-resource settings, which Kononowicz et al. [112] discuss as being effective in a range of countries worldwide.

Physical RPSs are lifelike physical robots that can simulate realistic patient physiologies and pathologies (see Figure 3) [151]. The use of physical simulators originated with *Resusci Anne*, a static mannequin created to teach cardiopulmonary resuscitation in 1960. It was used to train more than half a billion people in life-saving skills [18, 143]. Later in the 1960s, in an effort to train anesthesiologists, researchers developed a physical RPS called SimOne, able to show palpable pulses, heart sounds, and movement. Its software provided several pre-programmed events, such as different changes in heart rate or blood pressure [64]. Since then, many companies have built more advanced RPS systems to support a range of clinical scenarios, including Gaumard Scientific and Laerdal.

Recent RPSs benefit from the ability to interactively convey thousands of physiological signals. Their high-fidelity physical bodies are comparable to the bodies of real patients, affording CLs a practice platform for physical examinations and procedures.

2.3 Open Problems in Simulation-based Education

Despite the many benefits of using patient simulators, there are several challenges with existing systems that may impede how effective they are at supporting CL education, particularly with regard to skill transfer

(a) (b) (c) (d) (e) (f)

Fig. 3. Examples of RPSs with physical faces. (a) Laerdal's Little Resusci Anne [64], (b) Code Blue III by Gaumard Scientific [26], (c) Laerdal's SimNewB [36], (d) Laerdal's Mama Natalie [31], (e) Simroid by Morita Corp [3], and (f) Gaumard's Pediatric HAL [1].

(how well skills map from simulated patients to real patients). Table 3 provides a summary of the existing types of patient simulators, as well as their benefits and challenges.

One main challenge with existing RPS and VPS systems is low usability and controllability, which can cause delay and distraction. These simulators are very complicated and difficult for CEs to control, particularly when running complex simulations in a dynamic learning environment. Running clinical scenarios on these simulators has several time-consuming tasks and requires scheduling. As a result, CEs often cannot run the necessary simulations to support effective learning strategies. Furthermore, clinicians tend to have fairly low technology literacy, so a poorly designed system along with poor socio-technical integration can adversely affect skill learning performance [160]. Finally, using robots in healthcare settings can potentially add disruption and delay to the simulation process, which will change the clinical workflow in unforeseen directions [160, 181].

The other main challenge is that most current commercial VPS and RPS systems suffer from a major design flaw: They completely lack FEs and thus the ability to convey key diagnostic features of different disorders and social cues, which can eventually cause problems with learner immersion and skill transfer. This is critical for scenarios that require dynamic changes in appearance (e.g., abnormal visual findings such as drooping, which cannot be easily portrayed on a mannequin). Therefore, this lack of expression limits the extent to which a CL will become engaged with and immersed in a simulation, which may adversely affect their learning performance [130]. Consequently, CLs may be learning to incorrectly read patient social cues and signals, and may need to be retrained. Due to the importance of FEs as a key social function and clinical cue in patients, it is essential to study the synthesis of expressions (both symmetric and asymmetric) in simulators.

While RPS, APS, and VPS systems with expressive faces can address the previous challenge, they introduce several technical challenges and opportunities with designing expressive systems. First, because facial expressions and their intensities are very person dependent and can vary greatly from person to person [212], it can be challenging to develop one generalized system to recognize, model, and display facial expressions of a wide range of different individuals and cultures. Furthermore, some of the simulators, such as VPS systems, are limited by a flat 2D display medium, making them unable to convey a physical 3D human-shape that clinicians can palpate to perform clinical assessments. Inaccurately exhibiting symptoms on a simulator's face may reinforce incorrect skills in CLs and eventually lead to incorrect diagnoses in their future career [65].

Other challenges with creating expressive simulators include the need to recruit experts with various skills for development, high development costs, and systematic physical limitations.

Therefore, to design robots and avatars with humanlike expressive faces capable of accurately exhibiting patientlike symptoms, it is beneficial to examine the effect of expressive mechanical or rendered faces. To do this, roboticists and engineers need to closely co-design systems with developers and designers with a range of expertise and also include a diverse set of stakeholders, including CLs, CEs, and patients [151, 160, 161].

Adopting an interface to a robotic or avatar face similar to a human-patient's face to mimic real FEs and symptoms requires knowledge on building and controlling physically embodied robots and/or animating virtual systems. It also requires having knowledge of the nature of human facial expressions and the existing methods

of analyzing (recognizing, detecting, and tracking) human facial features. Moreover, it requires knowledge of the existing methods on developing models of humanlike facial expressions and techniques to incorporate and synthesize patientlike FEs onto the simulator's face.

2.4 The Face as a Communication Modality for Robots and Virtual Systems

The human face is a key expressive modality for communicating with others and understanding their intentions and expressions. Facial expressions are a form of visual communication that help to enhance other modalities of communication, such as spoken or gestural language, and enable people to spontaneously communicate important information [60, 133]. In clinical settings, healthcare workers use other non-verbal cues to infer patient physiological states, such as pallor, blinking, eye gaze, blushing, and sweating.

RPS and VPSs with expressive faces also can benefit from this humanlike ability to create better connections and interactions with users and be more favorably perceived [67]. This is why many roboticists develop physical or virtual embodiments capable of displaying facial expressions. Sometimes these expressions are conveyed physically (e.g., with mechanically moving parts), and sometimes they are conveyed virtually (e.g., using 2D displays) (see Figure 4).

While building accurate physical and virtual platforms for robots can enhance interaction, poorly designed faces can adversely affect the interaction and create distractions [67]. In the 1970s, Mori introduced the uncanny valley concept that explains people's negative reaction to certain lifelike robots [135]. The idea is that as robots become more humanlike, they become more attractive until they reach a certain point, after which, people perceive the robots as being creepy and/or immoral. This effect has since been validated across multiple experimental studies [108, 191].

It is important to consider the variability of facial expressions while designing robotic platforms capable of generating humanlike expressions. For many years, facial expressions were considered a universal language to express internal emotional states across all cultures [104]. However, recent cross-cultural studies suggest that culture is a well-documented source of variance in facial expressions. Studies by Jack et al. [104, 105] and Elfenbein et al. [81] suggest that humans across different cultures communicate emotions using different sets of facial expressions, and therefore, the notion of "universal" facial expressions proposed by Ekman [78] is now refuted in the light of demonstrated cultural nuances.

Another important consideration is the source videos / models used to create expressions on VPS or RPS systems. Many of these systems are trained on datasets of actors, presenting exaggerated facial expressions with little variance or cultural nuances, and tend to propagate the now unfavored Ekman "universal" framing of facial expressions with **action units–(AU)** based models. This can lead to bias and errors in both facial expression analysis and synthesis systems (see Section 3.6).

These studies raise an awareness that the impact of including different facial expressions, features, and functionalities in designing virtual and physical faces requires meticulous attention while designing realistic appearance and performance for human faces. Furthermore, the results suggest to carefully study the effects of using different facial analysis methods before, during, and after the realistic face design process.

3 AUTOMATIC FACIAL EXPRESSION ANALYSIS

To build robots and virtual avatars that can replicate realistic, understandable, humanlike expressions, it is necessary to be able to recognize how people express FEs. This section discusses common methods for manually and automatically detecting, locating, and analysing humanlike expressions in the presence of noise and clutter. First, we list a few key concepts.

Facial Landmarks (FL), also known as facial feature points or facial fiducial points, are visually highlighted points in the facial area, mainly located around facial components and contours, such as the eyes, mouth, nose, and chin.

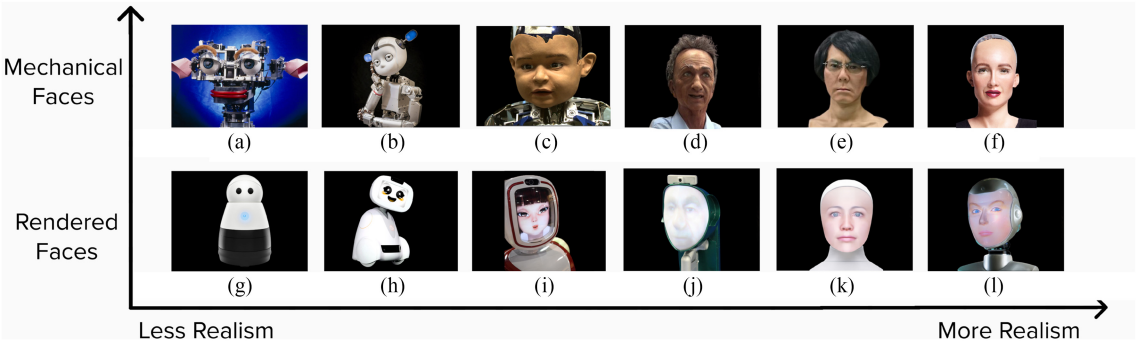


Fig. 4. Top: Physical robots with mechanical faces: (a) Kismet [4], (b) Simon [11], (c) Diego-San [7], (d) Charles [2], (e) Geminoid HI-5 [6], (f) Sophia [38]. Bottom: Virtual and hybrid robots with rendered faces: (g) Kuri [21], (h) BUDDY [19], (i) FURo-D [17], (j) Mask-Bot 2i [147], (k) Furhat [67], and (l) Socibot [37].

Facial AUs are individual components representing the movements of one or several specific facial muscles in each facial component surrounded with specific FLs [16]. Researchers introduced 46 main facial AUs [185], and others have added 8 head movement AUs and 4 eye movement AUs [16]. Examples include AU6-Cheek Riser, AU12-Lip Corner Puller, 5-Upper Lid Raiser, or AU-26 Jaw Drop. To express each specific facial expression, people need to move a specific subset of AUs in different facial components of their face. For example, researchers have identified AU6 and AU10 are associated with the expression of pain, and AU 10 with the expression of disgust [131].

Facial Action Coding System (FACS) is a system for manually describing facial actions according to their appearance, first published in 1978 and later updated in 2002 [78]. The main focus of FACS systems is to recognize facial expression *configuration*, which refers to the combination of AUs. This means that the system associates facial expression changes into a set of facial AUs (of 46 uniquely defined AUs) that produce them. This system also characterizes the variation of AU *intensity*, which represents the degree of difference between the current state of facial expression and neutral face [144]. FACS provides a 5-point intensity scale (A–E) for representing the AU intensity (A is weakest intensity and E is strongest intensity).

Manual FACS are based on annotations done by trained FACS coders who manually recognize both configuration and intensity of AUs in video recordings of an individual according to AUs described by FACS [78]. However, manual FACS rating requires extensive training and is subjective and time consuming. Thus, it is impractical for real-time applications [96].

Nowadays, many researchers work on automating FACS systems to analyze AUs [90]. Using automatic FACS instead of a manual approach can be beneficial, because training experts and manually scoring videos is time consuming. Furthermore, studies suggest using automatic FACS can enhance reliability, accuracy, and temporal resolution of facial measurements [125]. In developing these systems, in addition to *configuration* and *intensity* variation, researchers also analyze facial expression *dynamics* (i.e., the timing and the duration of different AUs). Dynamics can be important for human facial movement interpretation [90]. For example, facial expression dynamics can be beneficial for learning complex physiological behavioral states such as different types of pain [200].

The rest of this section briefly describes the main stages involved in automatic **facial expression analysis (FEA)**, as suggested in a recent survey by Martinez et al. [125], which include face detection and tracking, facial point detection and tracking, facial feature selection and extraction, AU classification based on extracted features, and new approaches on jointly estimating landmark detection and AU Intensity. Finally, we include a list of facial expression analysis software used by the community.

3.1 Face Detection and Tracking

To engage in facial expression analysis, systems need to be able to engage in "face localization," which Deng et al. define as including face detection, alignment, parsing, and dense face localization [75]. Deng et al. introduced RetinaFace [74, 75], "a robust, single-stage, multi-level face detector." It performs face localization on different scales of the image plane using joint extra-supervised and self-supervised multi-task learning. Many acknowledge that RetinaFace provides one of the most robust and strongest approaches to face detection. Others have made strides on related problems, for example, Hu et al. [100] explored a new approach of training separate detectors for face images with different scales. Their result reduced error by a factor of two compared to prior state-of-the-art methods.

In general, most current methods for face detection employ deep learning techniques, including Cascade-**Convolutional Neural Network– (CNN)** based models, **region-based Convolutional Neural Network (R-CNN)** and **Faster Regions with Convolutional Neural Network Features– (Faster-R-CNN)** based models, Single Shot Detector models, and Feature Pyramid Network-based models; see Reference [129] for a recent survey.

3.2 Facial Feature Point Detection and Face Alignment

Facial feature point detection (FFPD) (also known as landmark localization) generally refers to a supervised or semi-supervised process of detecting the locations of FLs. FFPD algorithms are sensitive to facial deformations that can be due to either rigid deformations (e.g., scale, rotation, and translation) or non-rigid deformations (e.g., facial expression variation, head poses, illuminations, noise, clutter, or occlusion) [194]. Enabling FFPD methods to align faces in an input image can lower the effect of changes in face scale as well as in-plane rotation.

Cascaded regression-based methods are one type of FFPD method that recognize either local patches or global facial appearance variations and directly learn a regression function to map facial appearance to the FL locations of the target image [205]. These methods do not explicitly build any global shape model, but they may implicitly embed the information regarding the global shape constraints (i.e., estimate the shape directly from the appearance without learning any shape model or appearance model).

Deep learning regression-based methods combine deep learning models, such as CNN, with global shape models to enhance performance. Early work in this field employed Cascaded CNNs [177], which predict landmarks in a cascaded way. Researchers then presented Multi-task CNNs [208] to further benefit from multi-task learning to increase the performance rate. Studies show the cascade regression with **deep learning (DL)** performs better than cascade regression and cascade regression better than direct regression [205].

In terms of facial feature point detection and face alignment, the **Face Alignment Network (FAN)** proposed by Bulat and Tzimiropoulos [58] is considered to be the state of the art. They constructed FAN by combining landmark localization with a residual block. They then trained the network on a 2D facial landmark dataset and evaluated it for large-scale 2D and 3D face alignment experiments. Researchers have proposed different follow-up methods to reduce the complexity of the original approach. For example, MobileNets is a class of efficient models that uses light-weight **deep neural networks (DNN)** to improve the performance [99].

3.3 Facial Feature Selection and Extraction

If the number of facial features becomes relatively large in comparison to the number of observations in a dataset, then some algorithms may not be able to train models effectively. High-dimensional vectors may cause two problems for classifiers: one, data may become sparser in high-dimensional space, and, two, too many extracted features may cause overfitting [102].

Li and Deng [117] provide a recent comprehensive survey on deep facial expression recognition and include discussion of feature learning and feature extraction techniques. A few examples are briefly discussed below. *CNNs* have been widely employed for the purpose of feature extraction, due to their ability to being robust

when encountering facial location changes and variations [87]. For example, researchers in Reference [176] used R-CNN to combine multi-modal texture features for facial expression recognition in the wild. Moreover, researchers [116] proposed a Faster-R-CNN technique to prevent from the explicit feature extraction step by producing region proposals.

Deep autoencoders and their variations have also been used for feature extraction. For example, researchers [114] used the **deep sparse autoencoder network (DSAE)** on a large dataset of images to prune learned features and develop high-level feature detectors using unlabeled data. The proposed DSAE-based detector is robust to different transformations, including translation, scaling, and rotation. As another example, researchers [162] employed contractive Autoencoder network that adds a penalty term to induce locally invariant features, leading to a set of robust features.

3.4 Facial Feature Classification

In the classification step, the classifier predicts expressions by categorizing the facial features into different categories. Similarly to the facial feature extraction stage, classification performance directly affects the performance of the facial expression recognition system.

Early facial feature classification work used techniques such as Naive Bayes [123, 179], multi-layer perceptrons [55, 150], and SVMs [157]; however, these have fallen out of favor given newer deep learning methods. While traditional facial expression analysis approaches usually perform the feature extraction step and the feature classification step independently, deep facial expression analysis approaches are able to perform both steps in an end-to-end training manner by adding a loss layer as the final layer to the DNN to adjust the error and then directly estimating the probability distribution over a set of classes [117].

For this purpose, many researchers have adapted CNN techniques for expression detection and classification [57, 119, 211]. The results of work done by Zeng et al. [119] shows that CNN classifiers trained faster and performed well. Another study indicates CNN classifiers also provide better accuracy compared to other neural network-based classifiers [157]. One main challenge to some of CNN classifiers is that they are sensitive to occlusion [119].

In addition to using deep neural networks for end-to-end training, other researchers [40, 76, 142, 170] have used DNNs for feature extraction and then added independent classifiers to the system for expression classification.

3.5 Jointly Estimating Landmark detection and Action Unit Intensity

Early FEA work often included a computationally intensive and laborious process (e.g., face and facial landmark detection, hand-crafted feature extraction, and limited classification methods). Nowadays, researchers benefit from having access to comprehensive, large-scale facial datasets, as well as advanced computing resources to develop more efficient facial analysis methods [68, 84, 85, 110, 118, 120].

One line of research is the work done on jointly estimating landmark and action unit intensity. For example, Wu et al. [204] proposed a constrained joint cascade regression framework to simultaneously perform landmark detection and AU intensity measurement. This method learns a constraint to model the correlation between AUs and face shapes. Next, they use the learned constraint as well as the proposed framework to estimate the landmark location and recognize AUs. The results of the study suggests the connection between these two parameters can improve the performance for both tasks.

Furthermore, many researchers consider the work done by Ntinou et al. [141] as the state-of-the-art method for jointly estimating landmark localization and AU intensity. In this work, researchers employed heatmap regression to model the the existence of an AU at specific location. For this purpose, they used a transfer learning technique between the face alignment network and the AU network.

It is worth mentioning that the newer directions for estimating AU intensity seek learning models with little or no supervision, including work done by Sanchez et al. [168], Wang and Peng [196], Wang et al. [195], and Zhang et al. [210].

One of the applications for AU intensity estimation is to further analyze and synthesize facial expressions representing specific feelings, such as pain. Many researchers have already conducted studies that indicate there is a relationship between a combination of AUs and pain, including work done by Kaltwang et al. [109] and Werner et al. [199]. Furthermore, it is worth mentioning that a fully functional automatic pain estimation system requires enough representative data, and for that purpose, there are some pain datasets publicly available (cf. Reference [122]).

3.6 Facial Expression Analysis Software

Dynamic FEA systems integrate automatic FACS to assess human expressions. Several commercial and open source FEA software packages are available, including iMotions, AFFDEX, FaceReader, IntraFace, and OpenFace 2.0.

iMotions developed a commercial tool for FEA that offers assessing FEs in combination with EEG, GSR, EMG, ECG, and eye tracking [24]. This tool lets users record videos with a mobile phone camera or laptop webcam and then detects changes in FLs. The researcher can set the tool to apply either the AFFDEX algorithm by Affectiva Inc. [80] or the **Computer Expression Recognition Toolbox (CERT)** algorithm used by FaceReader tool [121] to classify expressions. Different classifier algorithms such as CERT and AFFDEX employ various facial datasets, FLs, and statistical models to train the ML system to perform the classification task [24].

Affectiva's AFFDEX **software developer kit (SDK)** [128] is a commercially available real-time facial expression coding toolkit that is able to simultaneously recognize the expressions of several people and is available across different platforms (IOS, Windows, Android). The AFFDEX algorithm uses Viola-Jones [192] for detecting a face and identifying 34 landmarks, **Histogram of Oriented Gradient (HOG)** to extract facial textures, SVM classifiers to classify facial action and, finally, code seven facial expressions based on combinations of facial according to FACS [24]. AffdexMe is the name of the IOS-based AFFDEX SDK that enables developers to emotion-enable their own apps and digital experiences. The tests we performed on the trial version of this SDK show that the app can efficiently analyze and respond to seven basic emotions in real-time.

FaceReader [23] is a commercially available automated expression analysis system developed by Noldus. It enables developers to integrate expression recognition software with eye tracking data and physiology data. This tool provides an assessment of seven expressions, head orientation, gaze direction, AUs, heart rate, valence and arousal, and person characteristics.

FaceReader's algorithm uses the Viola-Jones algorithm [192] to find a face, then makes a 3D face model using facial points and face texture. It then analyzes the face using DL methods, and classifies the expressions using an ANN. Studies show that FaceReader is more robust than AFFDEX [173].

IntraFace is a software package developed by De La Torres et al. [73] for automated facial feature tracking, head pose estimation, facial attribute recognition, and facial expression analysis. This package also includes an unsupervised technique for synchrony detection that supports the function of discovering correlated facial behavior between two people.

IntraFace uses the SDM method to extract and track facial feature landmarks, and normalize the image with respect to scale and rotation [73]. They then extract HOG features at each landmark and perform a linear SVM for classifying facial attributes. Finally, they use the Selective Transfer Machine learning approach to classify facial expressions and AUs.

OpenFace 2.0 is an open source and cross-platform tool for facial behavior analysis released by the Multimodal Communication and Machine Learning Laboratory at Carnegie Mellon University in 2018 [13]. OpenFace 2.0 is capable of performing facial landmark detection, head pose estimation, facial action unit recognition, and eye-gaze estimation in real time [46].

OpenFace 2.0 uses a newly developed Convolutional Experts Constrained Local Model [206] and optimized FFPD algorithm for facial landmark detection and tracking that enables real-time performance [46]. Using this approach also enables OpenFace 2.0 to cope with challenges such as non-frontal or occluded faces and low

illumination conditions. The algorithm of this tool is able to operate on recorded video files, image sequences, individual images, and real-time video data from a webcam without any specialist hardware. GANimation [154, 155] is an anatomically aware facial synthesis method that automatically generates anatomical facial expression movements from a single image. This method provides the opportunity to control the magnitude of activation of each AU and combine several of them.

Latent-pose-reenactment [61] uses latent pose descriptors for neural head reenactment. This system can use videos of a random person and map their expressions to generate realistic reenactments of random talking heads.

4 FACIAL ACTION MODELING FOR SYNTHESIS

In many robotics and AI applications, in addition to recognizing FEs in people, we also need the ability to synthesize them on robotic and virtual characters. We discuss this further in Section 5; however, it is first important to discuss facial modeling.

Facial action modeling (FAM) builds a bridge between facial analysis (recognizing and tracking facial movements) and facial expression synthesis (translating modeled FEs onto an embodied face and animating its facial components) [167]. Thus, technology developers need to incorporate two key ideas in the design of face models: (1) patterns that model the human face (e.g, shape, appearance), both in its neutral state and the way facial movements (i.e., AUs) change to display different expressions, and (2) patterns of the temporal aspects of facial deformation (e.g., acceleration, peak, and amplitude).

The complexity of facial modeling can vary based on the **degrees of freedom (DoF)** of the embodiment (e.g., a mechanical robot or virtual face). It is less complex to build face models for more machinelike robots with very simple faces, such as Jibo [28], which only has one eye with varying properties and details. The complexity of designing a face model increases as the face becomes more realistic and detailed. For both robots with hyper realistic faces (e.g., Charles [161] and Geminoid HI-2 [138]), or a humanlike computer-generated virtual face (e.g., Furhat [25]), developers need to design highly accurate models to engage in synthesis.

There are two groups of information processing strategies for face modeling: theory-driven modeling and data-driven modeling [103].

4.1 Theory-driven Modeling Methods

Ekman and Friesen's FACS theory [78] describes the facial movements through observing the effect of each facial muscle on facial appearance and decomposes the visible movements of the face in the form of 46 AUs. Formerly, many researchers adopted FACS theory for facial modeling and embedded FEs derived from this theory constrained into their social robots [56, 97]. In this approach, programmers selected a small set of (static) FEs (e.g., tightening and slightly raising the corner of the lip unilaterally to express contempt) [79]. They then asked actors to contract k different combinations of muscle AUs to display the selected FEs to generate k different face images and score the face with FACS to verify muscle AUs depicted in each image. Finally, they asked observers to select which image better mimics each specific FEs and therefore identify which combinations of muscle AUs are signals for each specific FE.

However, there are several challenges with the theory-driven modeling methods. For one, these models are based on FEs that precisely met criteria selected and specified by researchers [79]. Moreover, since these models are based on static FEs, they lack dynamical data including the temporal order of FE movements (e.g., acceleration, peak, amplitude) [103], resulting in less realistic facial models and ultimately less human-lik simulators. Furthermore, even in studies on cross-cultural FE analysis where subjects pose cultural-specific expressions, still most subjects are identified as Westerners [137], leading to less diverse face models. Finally, people may have asymmetric facial expressions, such as people who have facial paralysis or deformities are rarely included, thus also limiting the diversity of facial models [134]. As a result, expressive robotic and avatar faces developed using theory-driven modeling methods lack the ability to generate a wide range of FEs. Therefore, these embodiments are not able to adequately communicate and interact with users.

4.2 Data-driven Modeling Methods

To address the gaps associated with theory-driven methods, researchers have proposed data-driven modeling methods (or, example-based deformation models) to computationally model (dynamic) FEs based on real data. Data-driven modeling methods usually consist of three main steps: data collection, facial expression and intensity data labeling, and facial expression model creation [103].

4.2.1 Data Collection. Data are generally collected in one of two ways: via recordings of human participants and through the use of artificial data creation.

One way of collecting data is to capture videos of facial expressions of human subjects (e.g., via an actor or layperson performing facial movements, or use of existing datasets). In this method, a researcher can use any statistical analysis method or facial expression analysis software package (see Section 3.6) to derive a parametric representation of facial deformations and identify the AUs correlated to each frame of a video. For example, Wang et al. [197] created a new FE dataset of over 200,000 images with 119 persons, 4 poses, and 54 expressions, which is about enough to evaluate the effects of unbalanced poses, expressions on the performance of the FE tasks.

Another way of collecting data is by generating artificial data through artificial data creation methods. In this method, developers usually use facial movement generators to randomly generate an enormous range of artificial dynamic facial expression videos. For example, Jack et al. use a facial movement generator, which randomly selects a subset of AUs, assigns a random movement to each AU by setting random values for each temporal parameter, combines randomly activated AUs, and finally projects them to a robotic face to create random facial animation videos [66].

4.2.2 Facial Expression and Intensity Data Labeling. Researchers have used different techniques for labeling FE data correlated to each frame of videos and their intensities, including manual labeling by both lay participants and domain experts and unsupervised data labeling via use of machine learning.

For instance, Jack et al. [66] recruited participants to watch videos of facial expressions. If the projected video formed a pattern that correlated with the perceivers' prior knowledge of one of six expressions, then they manually assigned a label to identify the expression and its intensity rating accordingly. Other researchers working on labeling FEs use domain experts (e.g., clinicians) to manually label data [134]. Other researchers develop facial expression datasets that use different semi-supervised or unsupervised techniques to label the data [197].

4.2.3 Facial Expression Model Creation. The next step is the learning phase, where the system uses the shape and texture variations of several sample images in datasets to build a face model and generate its appearance parameters. The parameters of the face model are reversible, meaning that they represent the shape and the texture of all images in the dataset, and therefore, are able to regenerate realistic images similar to each of the learned sample images. Thus, researchers can reverse-engineer specific dynamic FE patterns. This helps to derive the unique patterns of correlated AUs that are activated over time, which are correlated with human perception of each expression. For example, Chen et al. [66] developed their models by calculating a 41-dimensional binary vector per emotion detailing all AUs, and also seven values detailing the temporal parameters of each AU.

Using these three steps, developers can learn and build mathematical models of the dynamic FEs within a video stream that make it possible to reconstruct these FEs on a robot or avatar's face and animate them later [66].

5 FACIAL EXPRESSION SYNTHESIS AND ANIMATION

Facial expression synthesis and animation (FSA) refers to techniques used to animate dynamic expressions on the faces of avatars or robots using previously developed face models. FSA techniques provide the facial movement vocabulary that maps the developed model of AU movements and densities into the mesh topology of the social robot or avatar heads [148]. Using this technique makes the simulated face able to display AU

Table 2. An Overview of Technical Approaches for the Purpose of Facial Expression Synthesis and Animation [29, 83, 115, 167]

Categories	Process	Benefits	Drawbacks
Skeletal-based	Rigs a skeletal model to automatically associate each bone and joint into various parts of the embodiment's face and animates it using skeletal motion data.	• Manipulating the virtual face is Less labor cost, as animators only need to manipulate a set of vertices (bones) instead of each individual vertex.	• Generating the accurate mapping between bones and facial parts is labor consuming and time consuming. • This can lead to unrealistic artificial-looking animations and inaccurate synthesis and unrealistic FEs.
Blend-shapes	Creates a number of main mesh topologies of the face, and uses an automatic interpolation function to blend these topologies to create a smooth transition between them.	• It has low computational time • It is easy to implement	• It needs a great number of main topologies of different expressions. • This method only provides synthetic FEs in between the existing examples • Mesh design and animation creation is labor and time consuming. • Inconvenient for real-time application.
Parametric-based	Uses a system of parameters to create both the face model and different deformation model based on the visual or physical effect of muscle actions.	• It creates more realistic animations. • It can create various deformations. • It makes it possible to create interactive animations by incorporating text, audio, or video data.	• It requires using lots of high-quality motion capture equipment. • It is not usually feasible to perform it in real time.

movements corresponding to developed facial expression models. Concerning FSA, many articles have reviewed state-of-the-art methods and techniques, including References [83, 148, 167].

5.1 FSA Technical Approaches

Existing surveys in facial expression synthesis and animation include Reference [115], Reference [167], and Reference [83]. The surveys suggest there are three primary categories of techniques for synthesis purposes: skeletal-based, shape blend-based, and performance-driven approaches. Table 2 provides a summary of common approaches, which are further discussed below.

Skeletal-based approach (also known as the key-framing approach) works by rigging a skeletal model using an interactive tool to mimic the contraction of facial muscles and generate synthetic facial movements [29, 83]. For this purpose, animators use the 3D rigging tool first to construct a rig of bones and joints based on an estimation of the locations of facial muscles. They manually define the combinations of muscles representing each and every facial expression, and associate each bone into different parts of the avatar's visual presentation accordingly. Using this mapping, animators can automatically animate the virtual face using skeletal motion data.

Animating a virtual model using this approach is less labor-intensive, as animators only need to manipulate a set of vertices (bones) instead of each individual vertex. However, the downside of the skeletal-based approach is that generating the accurate mapping between the bones with facial parts is labor- and time-consuming. Furthermore, because it is difficult to accurately model facial movements based on bone movements, using this method can generate unrealistic artificial-looking animations and lead to inaccurate synthesis and unrealistic FEs on a virtual robot [167].

Blend-shape approach works by creating a number of main mesh topologies of the expressions and poses examples collected from the face of a real subject (one for each main expression), and then using an automatic

interpolation function to linearly blending these topologies to create a smooth transition between them [167]. To achieve smooth animations, animation developers need to generate hundreds of blended topologies.

This approach is commonly used to animate virtual faces, as it benefits from low computational time and is easy to implement. However, the performance of this approach greatly depends on the existing examples of different expressions [29]. Furthermore, this method only provides synthetic FEs in between the existing examples [83]. Furthermore, manually designing the main mesh topologies and manipulating each vertex to create animations is labor intensive and time consuming, making it an inconvenient modeling technique for creating real-time, long animations [167].

Parameter-based approach (also known as Motion Capture or the Performance-driven approach) uses a system of sensors and cameras to record motions and FE movements of a subject [167]. It then learns the face and deformation parameters from the captured data (including visual or physical effects of muscle actions) and finally transfers synthetic FEs onto the virtual robot's face.

In comparison with the other two methods, the performance-driven approach has the potential to be more realistic [29]. The use of parameter-based models makes it possible to create a wide range of deformations. These techniques also support creating interactive animations by incorporating text, audio, or video data in the model developing process [167]. However, to get the best and most accurate simulation using this method, it is necessary to use lots of high-quality motion capture equipment. Although this method is greatly used by major film making companies, it is not a convenient approach for technology developers and animators [83].

5.2 Advanced FSA Methods

Recently, researchers have performed more research-oriented studies of facial expression generation, that reflect ongoing attempts to address several of the challenges with respect to the expressivity of a facial expression synthesis system. More specifically, recent studies have focused on automatically synthesizing facial expressions from a few or single images using the newest advances in **Generative and Adversarial Networks (GAN)**.

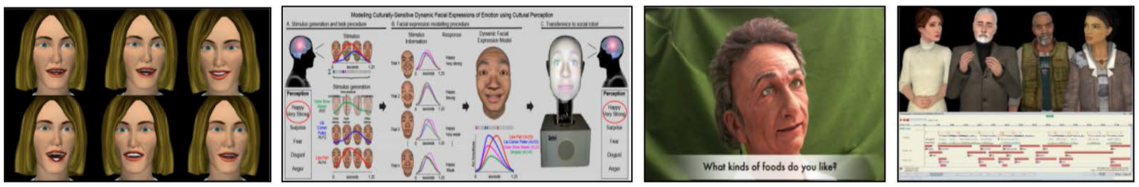
For example, Pumarola et al. [154, 155] introduced GANimation to automatically generate facial expressions in a continuous domain, without using any facial landmarks. They conditioned the network on a one-dimensional vector that represents the existence and the magnitude of each AU. This provides the opportunity to control the magnitude of activation of each AU and combine several of them. Additionally, they trained the network in a fully unsupervised manner, only requiring images annotated with their activated AUs, leading to an approach that is robust to changing backgrounds and lighting conditions.

In addition, other recent work addresses face reenactment and synthesis in a landmark-driven way. For instance, Burkov et al. [61] recently proposed a "neural head reenactment system" that uses a latent pose representation, based solely on image reconstruction losses. This system can use videos of a random person and maps their expressions to generate realistic reenactments of random talking heads.

Another recent work in this field is by Zakharov et al. [207], who developed a system that can generate plausible video sequences of speech expressions and mimicry of a particular person. They use a deep network that combines adversarial fine-tuning into a meta-learning framework to train lifelike digital speaking heads based on only a few photos of a person (e.g., a few-shot approach). This model can generate photorealistic animations of both random people and portrait paintings.

Gecer et al. [89] proposed a novel multi-branch GAN architecture that synthesizes photo-realistic expressions. It adopts a multimodal approach by including multiple 3D features (e.g., shape, texture, normals, etc.). They then trained the network to generate all modalities in a local and global correspondence, and condition the GAN by expression labels to create 3D faces with various expressions.

OpenPose, proposed by Cao et al. [62], is a open source, real-time system that detects the 2D pose (including the face) of multiple people in a single image. It employs a non-parametric representation to learn which body or facial parts is related to which person in the image. The system achieves high accuracy and real-time performance, regardless of the number of people.



(a) (b) (c) (d)

Fig. 5. (a) Figure of the Greta virtual agent [27, 145]. (b) Figure of synthesizing dynamic facial expressions onto the Furhat robot [67]. (c) Figure of the Charles robot mimicking a human [161]. (d) Figure of the Faceposer software interface [22].

5.3 FSA Exemplar

Researchers have mapped the synthesized motions to the face of different embodiments using FSA software packages (see Figure 5). For example, Faceposer SDK [22] for the Steam Source engine [39] is a virtual platform that uses synthesis framework to transfer facial expressions and skeletal animations to a virtual character's control points for animation. After generating facial movements and transformation parameters from a source video using one of the methods described in Section 4, Faceposer's synthesis framework converts the parameters into 21 control points (Flex sliders). The system saves the values of the Flex sliders in a .VCD scene file consisting of a header section with date, simulator, and timescale information; a variable definition section; and a value change section. Finally, after importing the .VCD file to the Faceposer SDK as the input, the SDK transfers the FEs on an avatar's face accordingly and animates the avatar.

Moreover, Pelachaud [145, 146] introduced Greta, which is a conversing socio-emotional virtual agent. This agent's software provides users with a real-time platform to control socio-emotional virtual characters and develop natural interaction with humans. Greta animation engine receives body animation parameters and facial animation parameters as inputs and synthesizes the expressions on a virtual character using Ogre3D or Unity3D [27].

Furthermore, Chen et al. [66] introduced a social physical-virtual agent displayed on a Furhat robot [25], which is capable of re-displaying facial expression using state-of-the-art 3D animation techniques. The introduced agent's algorithm provides full control over face designs, and includes realistic lip movements, as well as high-level control over the eyes and other facial movements [25]. It also provides the user with the opportunity to change the projected face's ethnicity, gender, language, and even its species. To measure the humanlike-ness of their synthesis approach, they performed an experiment to compare two FE synthesis methods (one generated through their reverse-engineering and synthesizing method, and one manually pre-programmed on their social robot). Their results suggest that users perceived their reverse-engineered expressions as more humanlike than the existing expressions of the robot [66].

Charles is a humanoid, hyper-realistic robot head from Hanson Robotics [158, 161]. Charles is able to display lifelike human expressions as it has wrinkles on the skin and 22 DoF in the face and neck. The robot has micro-controllers to control the motors that move the brow, eyes, midface, lips, mouth, jaw, head, and neck. Its control system generates motions using a direct AU-to-motor mapping system to synthesize expressions.

6 OUR WORK TO DATE

For the past decade, our team has developed new expressive, interactive RPS and VPS systems. This work has included new computational pipelines and control systems for FEA, **facial action modeling (FAM)**, and FSM, new robotic and virtual patient simulator embodiments, and new methods for modeling and synthesizing a range of conditions including dystonia, pain, **Bell's Palsy (BP)**, and stroke [130, 132–134, 149, 151, 152, 159, 161]. We briefly discuss this work below.

Fig. 6. Robotic patient simulators are tele-operated, life-size mannequins that can exhibit thousands of physiological signals, and can breathe, bleed, and respond to medications. However, they are largely inexpressive, leading to poor training outcomes for CLs, and possibly poor clinical outcomes for patients. Our work addresses this gap by introducing patient simulator systems with a much wider range of expressivity, including the ability to express pain, neurological impairment (e.g., stroke, Bell's Palsy), and other clinically relevant expressions, via simulators with diverse genders, races, and ages.

6.1 Creating New Embodiments for Robots and Avatars

Our work on creating new embodiments for robots and avatars thus far has focused on designing physical and virtual faces and leveraging robots' and avatars' expressivity, diversification, and control modalities to improve human health and safety [151] (See Figure 6).

The contributions of this work are as follows. We created different virtual avatars with diverse ethnic backgrounds and genders using the aforementioned Source SDK tool. We also supported the redesign of our team's bespoke robotic head and a low-cost expressive face[159] by increasing its DoF to 21 and performing iterative experimentation to increase their realism and efficacy. Additionally, we have developed appearances for these expressive faces to include more diverse backgrounds and to represent different age groups, genders, and races.

The robots are capable of conveying dynamic humanlike expressions meeting or exceeding the current state of the art. This work may one day help clinicians improve their clinical communication and cultural competency skills with real patients with different ethnic and cultural backgrounds.

6.2 End-to-End Analysis-Modeling-Synthesis Framework Development

In addition to building robotic and virtual embodiments, we also developed an end-to-end **Analysis-Modeling-Synthesis (AMS)** framework that included: automatic FEA, FAM, and dynamic FSA systems (see Figure 7).

The contributions of this work are as follows. First, we extended an FEA system previously developed by our team [133] to improve automatic FACS ratings of facial AUs. The extended FEA system benefits from preprocessing techniques such as noise reduction and facial alignment techniques to diminish the effects of facial deformations, including translation, rotation, and distance to the camera. Next, a CLM-based tracker [70] is used in the FEA system, as it is robust to illumination and occlusion. This tracker robustly locates the FL locations on an input frame based on the global statistical shape models and the independent local appearance information around each landmark.

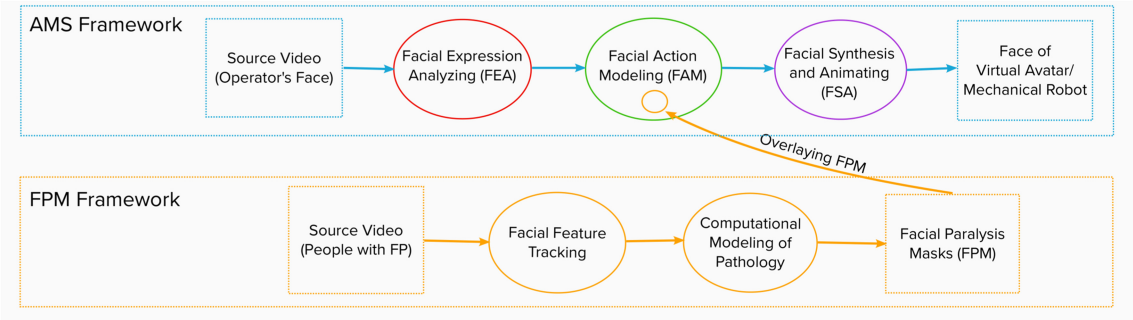


Fig. 7. In our work, we have developed of an end-to-end AMS framework to recognize, model, and synthesize facial expressions of real humans to the face of a physical or virtual robotic head. The AMS framework included automatic FEA, FAM, and dynamic FSA systems. Furthermore, we developed a novel FPM framework to build accurate computational models of people with Bell's Palsy that are constructible in real time.

Second, we proposed a novel data-driven FAM system developed in three steps: First, we collected real dynamic facial expression data, second, we labeled the FE data correlated to each frame of videos and their intensities using manual outsourcing technique, and third, we generated reversible appearance parameters by calculating a 46-dimensional binary vector detailing all AUs. This FAM system makes it possible to computationally model dynamic FEs tracked by the FEA system based on real human facial expression data, which ultimately can make it easier for developers to generate a diverse set of realistic face models derived from real patients.

Third, we extended an FSA system previously developed by our team [133] for synthesizing realistic, patient-like FEs on both our bespoke RPS head and virtual avatar faces. The method is based on data-driven synthesis, which maps motion from video of an operator/CE onto the face of an embodiment (e.g., virtual avatar or robot). This platform-independent software makes it possible for SMs to easily and robustly synthesize and animate realistic expressions on the faces of a range of embodiments, and makes it easy for CEs to perform simulation.

This end-to-end AMS framework models and synthesizes patient-data-driven facial expressions, and can easily and robustly map these expressions onto both simulated and robotic faces. By leveraging this work, other roboticists and engineers will be able to discover platform-independent methods to control the FEs of both robots and virtual agents. This can also help improve how clinicians interact with patients, and increase their cultural competence when interacting with patients from diverse backgrounds.

6.3 Modeling and Synthesizing Clinically Relevant Expressions

For the past decade, our team has developed new methods for modeling and synthesizing a range of clinically relevant conditions, including dystonia, pain, BP, and stroke [130, 132–134, 149, 151, 152, 159, 161] (see Figure 8). We briefly summarize several of these projects below (dystonia, pain, stroke) and then discuss in a bit more detail results of recent Bell's Palsy modeling and synthesis project.

6.3.1 Dystonia. Dystonia is a movement disorder characterized by involuntary motions, often in the head and neck. People with dystonia often struggle during interaction due to the biases of others. Thus, we were curious to explore if a robot conveying dystonia could serve as a facilitator to help improve human–human communication. Our team interviewed four people with head and facial movement disorders and synthesized their movements on a physical robot. We then experimentally explored using these robots as social facilitators to improve communication between people with and without disabilities. Our results suggest that a robot may be useful for this purpose [161]. We also observed a significant relationship between people who hold negative attitudes toward robots and negative attitudes toward people with disabilities.

6.3.2 Pain. Another clinically relevant facial expression we have explored is pain. We have modeled and synthesized both acute and chronic pain, on both virtual avatars and physical robots [130, 132]. In one study, we explored people's perceptions of pain, both on a humanoid robot and comparable virtual avatar, using autonomous facial expression synthesis techniques.

We conducted an experiment with clinicians and laypersons to explore differences in pain perception across the two groups, and also to study the effects of embodiment (robot or avatar) on pain perception. The results of this study indicated that clinicians have lower overall accuracy in detecting synthesized pain in comparison to lay participants. It also suggested that all participants are overall less accurate detecting pain from a humanoid robot in comparison to a comparable virtual avatar [130].

6.3.3 Stroke. Additionally, we built expressive patient simulator systems that can recognize and synthesize asymmetrical facial expressions similar to patients with stroke. Stroke, a substantial contributor to the global disease burden, affects 15 million people each year and is the second leading cause of death worldwide [77, 171]. Of those affected by stroke, five million die and another five million are permanently disabled [171]. One of the contributors to this disease burden is diagnostic failures: stroke is the fourth most common misdiagnosis reported by clinicians [180].

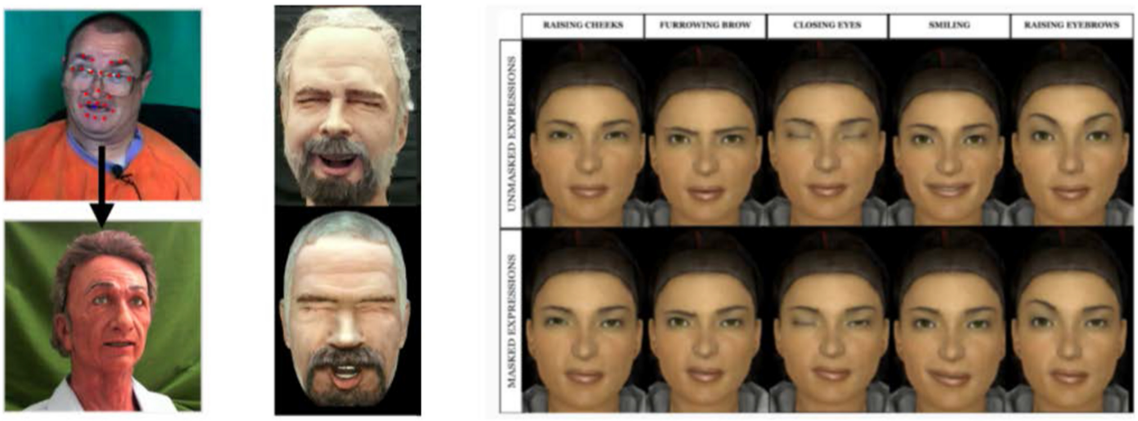
Research shows CLs often fail to master the neurological examination on simulated patients. This may result in inadequately performing the exam on real patients [15]. Even if a CL performs the exam well, they may have little confidence in the accuracy of their findings. Given the subjective nature of interpretation of these findings, low-confidence in the neurological exam, irrespective of how well it is performed, may lead to an uncertain interpretation of the results. This uncertainty can lead to missed opportunities for acute interventions, prompt treatments, and prevention of serious harm [44, 136].

One approach to address the urgent need for a smart training tool for clinicians to practice their stroke diagnosis skills is to make simulators capable of realistically depicting non-verbal, asymmetric facial cues that are important for the rapid diagnosis of neurological emergencies, such as stroke [149, 152]. In our work, we introduced the concept of stroke-mask synthesis on VPS and RPS systems [152]. These systems depict patients with acute stroke and can help train future generations of neurologists in rapid diagnosis of acute neurological injury. Our work can also help researchers in the facial recognition community to explore new methods for asymmetric facial expression analysis and synthesis.

6.3.4 Bell's Palsy. Every year, 22 million people experience stroke, Parkinson's disease, Moebius syndrome, and BP [5, 124, 189], which can cause **facial paralysis (FP)**. Facial paralysis is the inability to move one's facial muscles on the affected side of the face, leading to **asymmetric facial expressions (A-FEs)** [48]. The quality of social interaction that people with A-FEs experience can be poor due to others who have difficulty understanding their emotions [53]. Studies show observers perceive the emotions of a person with FP differently from their actual emotional states [178]. For example, people with severe FP are perceived as less happy than people with mild FP [54].

In clinical contexts, these misperceptions can lead to poor care delivery. Healthcare providers frequently have negatively biased impressions of patients with facial nerve paralysis [186], which may adversely affect the quality of care they receive [161, 164]. If a patient and a healthcare provider do not communicate effectively, then there is a higher chance that their treatment will be unsuccessful [45, 178]. Therefore, new training tools that enable CLs to practice their interaction with FP patients may result in improved care for people with FP and also improve how clinicians calibrate their perception of asymmetric expressions.

However, prior development of facially expressive VPS and RPS systems was based on the assumption that human faces are structurally symmetric and thus have not accounted for expressing A-FEs. Due to the large number of people affected by FP, it is important to also explore synthesizing A-FEs in clinical contexts. To our knowledge, FP patient simulators have not been explored in this way.



(a) (b) (c)

Fig. 8. Three examples of the expressive patient simulator systems our team has built, with clinically relevant-expressions: (a) Dystonia [161], (b) Pain [131], and (c) Bell's Palsy [134].

In our work, we focused on considering people with FP to provide a better training tool for clinicians to practice how to avoid forming biased impressions, improve clinical communication, and, therefore, improve care delivery for people with FP. For this purpose, we introduced the concept of **facial paralysis mask (FPM)** synthesis to incorporate accurate A-FEs on patient simulators based on real patients' facial characteristics, situated within a clinical education context. FPMs are computational models of different pathologies derived from recognized expressions of real people with FP. FPM synthesis is the process of using pre-built FPMs and overlaying them on the facial model of the standard analysis-modeling-synthesis framework described in Section 6.2 to recreate A-FEs on a RPSs and VPSs.

This work explored two research questions. First, *How does one computationally model the facial characteristics of BP, and synthesize them on a patient simulator to help support clinical engagement of those affected?* To address this question, the first step was to collect self-recorded, publically available videos from people with BP conveying four expressions (raising eyebrow, furrowing brow, smiling, and closing the eye).

Next, we presented a novel algorithm for the FPM framework to build accurate computational masks that can model facial characteristics of people with BP and are constructible in real time (see Figure 7). This algorithm tracks faces in each source video and uses the 2D coordinates of the 34 facial features of the unaffected side of the face to calculate the 2D coordinates of the other part of the face, assuming that the person did not have A-FEs. Dividing the actual coordinates of the affected side by the calculated coordinates of the affected side gave us the scaling parameters $\beta_{i,x}$ and $\beta_{i,y}$ for x and y of each of the facial points. A 68-bit array consisting of the scaling parameters of all 68 tracked feature points is the calculated mask for the patient with BP.

Our second research question was *How realistically do these masks convey signs of BP when applied to a virtual patient?* To address this question, we conducted a qualitative, expert-based perceptual experiment to evaluate the realism of the synthesized expressions in comparison to actual patients and get feedback for further refinement. This is a common method for evaluating synthesized FEs [49, 127].

To perform this validation, after collecting videos from a performer without BP, we inputted the videos into the AMS framework (see Section 6.2), and overlaid three pre-built masks of BPs to recreate the AFE (see Figure 7). Next, the generated asymmetric expressions of BP were transferred to the face of a VPS system to create stimuli videos (see Figure 8(c)).

The results of this study suggest that two of the developed BP masks realistically display signs of BP. Furthermore, clinicians' perceptions of the synthesized expressions were comparable to their perceptions of the

expressions of real people with BP. Therefore, the models described in this work have the potential to provide a practical training tool for CLs to better understand the emotions of people with this facial paralysis.

7 FUTURE RESEARCH DIRECTIONS

There are several opportunities to advance the state of the art of expressive RPS and VPS systems within the context of clinical learning, as well as in the broader context of robotics and HRI. These include technical advancements, such as new methods for FEA, FMA, and FSA, as well as socio-technical considerations, such as stakeholder-centered design and ethical questions. We briefly outline these below.

7.1 Advancing Expression Recognition and Synthesis Approaches

As discussed, there are many methods for recognizing and synthesizing facial expressions. However, they have their drawbacks. Many commercially available systems are unable to perform the tasks necessary for FE analysis or synthesis (e.g., FaceReader is not able to provide head pose estimation). Furthermore, systems may may lack state-of-the-art performance, rendering them impractical for clinical applications.

Thus, there are many opportunities to advance the state of the art. For example, some regression-based methods such as CNNs are successful for FL detection and tracking. Furthermore, Gabor features showed promising results for feature extraction, and the CNN and SVM methods improved classification performance. Integrating these approaches into facial expression FEA and FSA systems may improve analysis and/or synthesis of dynamic FEs in individuals with and without facial disorders.

7.2 Combining Domain Knowledge with Facial Model Development

As part of the design process, engaging in stakeholder-centered design with CEs and CLs, as well as conducting observations of live simulations is important. For example, neurologists can help validate if neurological impairment models created by the system are realstic and also ensure the patient simulator's appearance and expressiveness is well aligned with their clinical education goals.

7.3 Real-world, Spontaneous Data Collection

It is important for developers to release systems that are designed and built using enough real-world, spontaneous facial expression data [94]. The number of facial expressions used for training and developing FEA, FAM, and FSA systems should be much higher to lead to more realistic results. In case of having a low number of images for training, it is challenging to choose the best approaches to enlarge the dataset while developing the system. Expressive robot developers also need to make sure the system includes a continuous adoption process that learns each user's expressions over time and adds them to its knowledge base [94]. It is also important to pay close attention to include the variability of the facial data in terms of subjects by including data from subjects well represented in gender and ethnicity, as well as diversity in terms of lighting, head position, and face resolution [94]. Given that patient simulators are designed to mimic humans and are designed for use by humans, we added a discussion on the importance of having designs that are informed by human sensory systems and behavioral outputs. Finally, it is important that datasets are labeled and analyzed in concert with domain experts, but to our knowledge little work has been done in this area. One potential solution can be to create a large training set of photorealistic facial expressions generated using existing face generation platforms labeled by human observers.

There are several existing facial expression datasets and Action Unit datasets that tackle some of the data collection challenges, including DISFA [126], BP4D-spontaneous [209], Aff-Wild 2 [111], and SEWA DB [113]. Furthermore, some of the recent facial expression synthesis methods, such as those mentioned in Section 5, are also intended to address these challenges. However, more work can be done in this field to tackle all the afore-mentioned problems.

Moreover, newer directions also seek learning models with little or no supervision, both for facial landmarks (unsupervised landmark detection) and for Action Units that can help to address these challenges.

In terms of identifying databases of images or videos that reflect real facial expressions, it is important to consider the relationship between internal states and external facial cues. Work done by Benedek et al. [50] indicates people perceive the appearance of the face, especially the eyes of others, to understand both their external goals or actions, and their internal thoughts and feelings. Voluntary facial expressions are sometimes made in the absence of internal states. However, it is difficult to detect internal states in case attention is not presented externally. Therefore, it is critical to identify datasets of real data to better infer the external facial cues and more accurately interpret internal states.

It is worth mentioning that there is the potential of having a pattern of confusions (false alarms and misses) in detected facial expressions. False alarms is the errors of describing a facial expression being present when it was absent. Misses is the errors of describing a facial expression as being absent when it was present. Studies indicate that the pattern of confusion becomes worse when some other challenges occur at the same time, such as illumination or occlusion in an image [153].

7.4 Cultural Considerations

Researchers have also explored the caveats associated with cultural variance in the way observers infer internal experiences from external displays of facial expressions. For example, Engelmann et al. [82] argues that culture influences expression perception in different ways. For one, people from different cultures may perceive the intensity of external facial expressions differently. For example, American participants rated the intensity of same expressions of happiness, sadness, and surprise higher that Japanese participants. Moreover, depending on cultural contexts, there is a difference in the way people infer internal states from external facial cues of expressions. For example, researchers ran an experiment to ask two groups of American and Japanese participants to rate the intensity of internal and external state of a person expressing certain emotions. American participants gave higher rates to external facial cues of emotions, while Japanese participants gave a higher ratings to internal state of emotions. Therefore, it is important to consider these cross-cultural differences in inferring internal states and external expressions.

7.5 Generating Universal Models for Various Pathologies

To generally represent all patients with specific pathologies, one can create a universal model for each that encompasses its predominant features. This can be done by leveraging our previous findings in Section 6.3 to further extend the FPM framework in two directions: (1) Extend the FPM framework to encompass the predominant features of a specific pathology (e.g., stroke) and (2) transfer the framework from being an individual mask generator to a universal model generator. This can be done by using enough source videos of people with the specific pathology, extracting its common features, and creating a general model (see Figure 9). By leveraging this work, CLs will have the potential to more accurately diagnose people with diverse backgrounds, and to be better able to interact with them.

7.6 Sharing Autonomy between Users and Expressive Robots

Considering how to share autonomy between a human and robot is an important aspect to ensuring effective HRI [156]. It can help to reduce an operator's workload, allow both inexperienced and professional operators to control the system [86, 156].

As such, it is important to focus on interaction between the control system and human users in the context of expressive simulator systems. Thus, researchers can design and validate a customizable, shared autonomy system for expressive RPS systems to leverage the advantages of automation while also having users as "active supervisors." For example, in our work, we are designing a shared autonomy system that can support a range

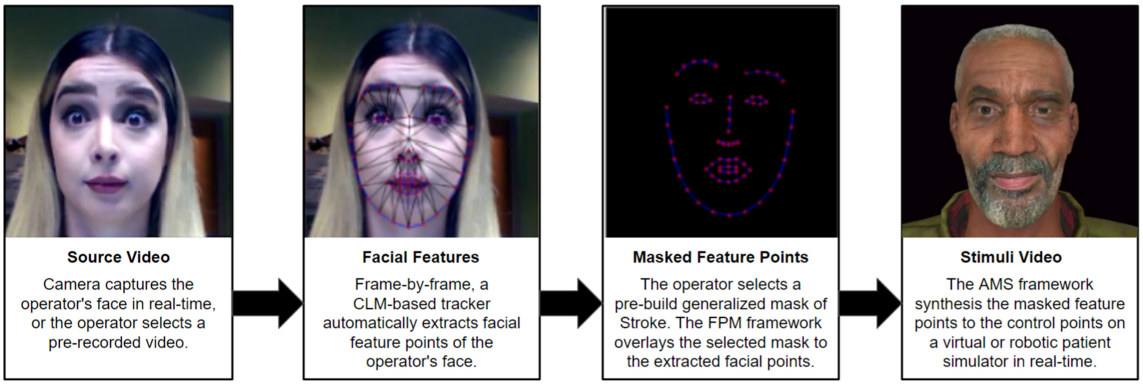


Fig. 9. The context for performing masked synthesis using the Generalized Stroke FPM framework and the AMS framework. This can be performed on either a VPS or RPS system.

of adjustable control modalities, including direct tele-operation (e.g., puppeteering), pre-recorded modes (e.g., hemifacial paralysis during a stroke), and reactive modes (e.g., wincing in pain given certain physiological signals) [151]. It also can help overcome common control challenges, including the operator being overwhelmed, having high workload, and lack of autonomy in robotic simulator systems. This system can help to make robots adjustable to different control paradigms, so that they reliably support CEs' workload in dynamic, safety-critical settings, and improve the operator's ability to focus on their educational goals rather than on robot control.

8 ETHICAL CONSIDERATIONS FOR PATIENT SIMULATOR SYSTEMS

Using FEA and FSA technologies to develop new RPS and VPS systems and integrating them within clinical learning contexts presents a number of ethical and social challenges that require specific attention. It is important researchers and technology developers carefully consider these challenges, and work to design inclusive technologies to avoid unintended consequences. While this is by no means an exhaustive list, a few key challenges are highlighted herein.

8.0.1 Racial and Ethnic Bias in FEA Technologies. There are many concerns regarding racial, ethnic, misogynistic, and ableist biases in FEA technologies, which can perpetuate social and fiscal oppression [51, 139, 140]. For example, many studies show high rate of misidentifying blacks by recognition systems, which can be due using FEA algorithms trained on a racially biased datasets, as well as systemic biases embedded within the systems themselves [59]. Such biased models can then affect FSA, and further perpetuating biases in clinical education [172]. Moreover, there are challenges regarding distancing and dividing effects caused by using FEA systems for controlling patient simulators. For example, an operator of an expressive robot sometimes need to adjust their feelings to express exaggerated facial expressions (e.g., intense smile) or fake facial expressions (e.g., reflecting different feeling than what they genuinely feel at the moment), so the FEA algorithm can detect and/or track the expression. Although some researchers think these adjustments may only cause minor problems or difficulties, others think using these technologies can distance and dehumanise people [43].

8.0.2 Privacy. Another concern is on privacy and the extensive use of data in FEA and FSA systems. Widespread use of these systems in healthcare settings can lead to the collection of large amounts of patients' and clinical workers' actions, locations, personal, physiological, and behavioral information. This can raise many concerns about the ways of protecting the privacy of collected personal data, as well as the ways simulator developers use the data.

8.0.3 Uncanny Valley. Another concern that often arises with highly humanlike RPS and VPS systems is a phenomenon called the Uncanny Valley [135]. This is a theory that suggests that as robots become more humanlike they are more attractive, until they reach a certain point, where people's affinity for these humanlike robots descends into a feeling of strangeness and unease [108, 135]. This is reflected in both their appearance and their behavior [169]. While CLs require highly humanlike RPS and VPS systems to learn proper clinical skills, ones that miss the mark can cause learner distress, and adversely affect their learning, Thus, RPS/VPS designers should carefully consider learners' perceptions as part of their design process.

8.0.4 Risks and Benefits of Diverse FSA. Just like humans, humanlike patient simulators that resemble a certain gender, race, or culture in their design can face judgement and aggression based on the biases towards such social identities. Designing humanlike robots with diverse appearance and behavior has numerous benefits. For example, building a humanlike robot resembling a patient who has had a stroke for healthcare education application provides the clinical learners with a great opportunity to practice their communication and procedural skills on these robots, preparing them for treating real human patients with stroke in their future careers [149, 152].

However, diversifying the appearance and behavior for simulators also introduces risks. For example, roboticists may implicitly or explicitly reinforce gender biases by assigning a specific gender to the robot during the design process, and CLs/CEs might as well during simulation sessions [172].

People also more readily dehumanize robots racialized in the likeness of marginalized social identities than those racialized White [174]. As such, people with racist behavioral biases represented similar racist biases while interacting with humanlike RPS or VPS systems of a similar race.

9 DISCUSSION

The technologies and methods discussed in this review can cultivate a bridge between robotics and healthcare research, and improve existing clinical training practices, by enabling VPS and RPS systems to become more diverse, interactive, and immersive for CLs and CEs. This will enable CLs to further engage during training sessions, will help them to significantly improve their communication and procedural skills, and ultimately save more lives. Building on these approaches will lead to systems with a much wider range of expressivity, such as the ability to express clinically relevant facial expressions. Through studies with stakeholders, including patients, clinicians, and clinical learners, technologists can improve the expressiveness of simulator robots, and improve the interactions between humans and robots for expressive patient simulators and beyond. Ultimately, this work may help clinicians deliver better clinical care, by both improving their diagnostic skills and by providing new educational opportunities for reducing racial disparities, by teaching them to be less biased when interacting with real patients [161]. Furthermore, disseminating the results of this work (and software) to the research community will help both the broader robotics and healthcare communities employ these novel systems in their own application domains.

APPENDIX

Table 3. Simulators: Types, Benefits, and Drawbacks

Type	Prior Work	Process	Benefits	Drawbacks
SHP	[41, 47, 91, 188]	Living humans acting as patients through a clinical scenario that is defined by CEs.	• Provides live, real-human case study • Tangible body and face • Capable of expressing facial cues	• Hard to manually control some of the physiological parameters • Difficult and expensive to recruit, especially ones at younger ages
APSs	[72, 184, 202]	A physical human-shaped surface with dynamic visual imagery of body parts projected on the surface.	• Their physicality can convey haptic similarity to people • Their virtual component can display dynamic appearances and FEs without being limited by hardware infrastructure	• Difficult to display an accurate representation of naturalistic symptoms • Limited field of view, projection occlusion, hard to have multiple users
VPSs	[8, 30, 35]	Interactive digital simulations of real patients, displayed on a screen.	• Can virtually portraying physiology variables without being limited by hardware infrastructure • Are able to richly and quickly display changes in the appearance, symptoms, behavior, or body language of simulator • Improve knowledge and skill-building compared with non-digital methods • make clinical education more accessible to CLs in low-resource settings • Can reduce preventable patient harm • Have a positive influence on CLs' comprehension, confidence, efficiency, and enthusiasm for learning • Enable CEs to run desired clinical simulation scenarios on realistic patients.	• **VPSs ONLY:** Limited by the flat 2D display that makes them unable to represent a physical 3D human-shape volume with touchable physiology variables **COMMON BETWEEN VPSs & RPSs:** • Complicated for CEs to use and control • Running clinical scenarios on them is time-consuming and need scheduling • Low technology literacy levels of CLs along with poorly designed system can adversely affect learning performance • Can add disruption and change the clinical workflow in unforeseen directions • Completely lack FEs, and thus lack the ability to convey key diagnostic features of disorders and social cues • Lack of FEs may adversely affect CLs' learning performance, and they will ultimately need to be retrained. • Even in systems with expressive faces, inaccurately exhibiting symptoms on a simulator's face may reinforce wrong behaviors in CLs and result in failure to recognize a disease in Cs' future careers. • Difficult to display accurate representations of naturalistic symptoms on them.
RPSs	[18, 26, 36, 64]	Lifelike physical robots that can simulate realistic patient physiologies and pathologies.	• Versatility: Can exhibit 5000+ variables • Have physical bodies comparable to the bodies of humans, where CLs physically examine and conduct procedures on. • Can reduce preventable patient harm • Have a positive influence on CLs' comprehension, confidence, efficiency, and enthusiasm for learning • Enable CEs to run desired clinical simulation scenarios on realistic patients.	• **RPSs ONLY:** Limited by hardware infrastructure

REFERENCES

[1] Meet Pediatric HAL S2225. 2020. Retrieved July 31, 2020 from https://www.gaumard.com/s2225.
[2] Meet the Robot That Can Mimic Human Emotion. Retrieved July 31, 2020 from https://www.cambridge-news.co.uk/news/cambridge-news/cambridge-university-robot-human-emotion-14431300.
[3] SIMROID Patient Simulation System for Dental Education. Retrieved July 31, 2020 from https://www.morita.com/group/en/products/educational-and-training-systems/training-simulation-system/simroid/.
[4] 1998. Kismet. Retrieved June 20, 2020 from https://robots.ieee.org/robots/kismet/.
[5] 2002. The World Health Report 2002: Reducing Risks, Promoting Healthy Life. Retrieved September 9, 2020 from https://www.who.int/whr/2002/en/whr02_en.pdf?ua=1/.
[6] 2010. Hiroshi Ishiguro: The Man Who Made a Copy of Himself. Retrieved July 31, 2020 from https://spectrum.ieee.org/robotics/humanoids/hiroshi-ishiguro-the-man-who-made-a-copy-of-himself.
[7] 2013. Diego-San Research Robot. Retrieved from https://www.hansonrobotics.com/diego-san/.
[8] 2014. CliniSpace Offers Healthcare Training Applications & Engine Platform. Retrieved June 20, 2020 from https://www.healthysimulation.com/5499/clinispace-offers-healthcare-training-applications-engine-platform/.
[9] 2016. Adverse Events in Rehabilitation Hospitals: National Incidence among Medicare Beneficiaries.
[10] 2017. Advancing Care Excellence for Seniors (ACE.S). Retrieved June 20, 2020 from http://www.nln.org/professional-development-programs/teaching-resources/ace-s.
[11] 2017. How Our Robots Will Charm Us (and Why We Want Them to). Retrieved June 20, 2020 from https://sonarplusd.com/en/programs/barcelona-2017/areas/talks/how-our-robots-will-charm-us-and-why-we-want-them-to.
[12] 2018. Here's the Real Reason Health Care Costs So Much More in the US. Retrieved July 31, 2020 from https://www.cnbc.com/2018/03/22/the-real-reason-medical-care-costs-so-much-more-in-the-us.html.
[13] 2018. OpenFace. Retrieved July 31, 2020 from http://multicomp.cs.cmu.edu/resources/openface/.
[14] 2018. Tug, One Platform, Multi-Purpose. Retrieved June 20, 2020 from https://aethon.com/products/.
[15] 2018. UCSD's Practical Guide to Clinical Medicine. Retrieved December 30, 2019 from https://meded.ucsd.edu/clinicalmed/neuro2.html.
[16] 2019. Facial Action Coding System (FACS)—A Visual Guidebook. Retrieved June 20, 2020 from https://imotions.com/blog/facial-action-coding-system/.
[17] 2019. Future Robot. Retrieved June 20, 2020 from http://www.futurerobot.com.
[18] 2019. How Far Has CPR Feedback Come? Retrieved June 20, 2020 from https://www.laerdal.com/us/information/resusci-anne-then-and-now/.
[19] 2020. BUDDY the First Emotional Companion Robot. Retrieved July 31, 2020 from https://buddytherobot.com/en/buddy-the-emotional-robot/.
[20] 2020. Da Vinci by Intuitive. Retrieved June 20, 2020 from https://www.intuitive.com/en-us/products-and-services/da-vinci.
[21] 2020. Explore Kuri. Retrieved July 31, 2020 from https://www.heykuri.com/explore-kuri/.
[22] 2020. Faceposer. Retrieved June 20, 2020 from https://developer.valvesoftware.com/wiki/Faceposer.
[23] 2020. FaceReader. Retrieved June 20, 2020 from https://www.noldus.com/facereader.
[24] 2020. Facial Expression Analysis. Retrieved June 20, 2020 from https://imotions.com/biosensor/fea-facial-expression-analysis/.
[25] 2020. Furhat Robot. Retrieved June 20, 2020 from https://furhatrobotics.com/furhat-robot/.
[26] 2020. Gaumard Simulators. Retrieved June 20, 2020 from http://www.gaumard.com/aboutsims/.
[27] 2020. Greta. Retrieved June 20, 2020 from https://github.com/isir/greta.
[28] 2020. Hey, I'm Jibo. Retrieved July 31, 2020 from https://jibo.com/.
[29] 2020. How to Create 3D Face Animation the Smart and Slick Way. Retrieved August 9, 2020 from https://gallery.mailchimp.com/e6002b57be315a1fb16689676/files/989f9065-8a00-462b-a3c1-fab00be0e276/Free_ebook.pdf?ct=t()&mc_cid=635a967b21&mc_eid=1744f12fd6.
[30] 2020. i-Human. Retrieved June 20, 2020 from http://www.i-human.com.
[31] 2020. MamaNatalie—Birthing Simulator. Retrieved June 20, 2020 from https://www.laerdal.com/us/mamaNatalie.
[32] 2020. MedEdPortal—Physician Resident Scenarios. Retrieved June 20, 2020 from https://www.mededportal.org.
[33] 2020. Minnesota Simulation Alliance. Retrieved June 20, 2020 from http://www.mnsimlib.org/.
[34] 2020. Patient safety. Retrieved July 31, 2020 from https://www.who.int/patientsafety/en/.
[35] 2020. Shadow Health. Retrieved June 20, 2020 from https://www.shadowhealth.com/.
[36] 2020. SimNewB. Retrieved June 20, 2020 fromhttp://www.laerdal.com/us/doc/88/SimNewB.
[37] 2020. SOCIBOT. Retrieved July 31, 2020 from https://robotsoflondon.co.uk/socibot.
[38] 2020. Sophia. Retrieved July 31, 2020 from https://www.hansonrobotics.com/sophia/.
[39] 2020. Source SDK 2013. Retrieved June 20, 2020 from https://developer.valvesoftware.com/wiki/Source_SDK_2013.

[40] Dinesh Acharya, Zhiwu Huang, Danda Pani Paudel, and Luc Van Gool. 2018. Covariance pooling for facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 367–374.
[41] G. Adamo. 2003. Simulated and standardized patients in osces: Achievements and challenges. *Med. Teach.* 25, 3 (2003), 262–270.
[42] A. H. Al-Elq. 2010. Simulation-based medical teaching and learning. *J. Fam. Commun. Med.* 17, 1 (2010), 35.
[43] M. Andrejevic and N. Selwyn. 2019. Facial recognition technology in schools: critical questions and concerns. *Learn Media Technol.* 45, 2 (2020), 115–128.
[44] A. E. Arch, D. C. Weisman, S. Coca, K. V. Nystrom, C. R. Wira, and J. K. Schindler. 2016. Missed ischemic stroke diagnosis in the emergency department by emergency medicine and neurology services. *Stroke* 47, 3 (2016), 668–673.
[45] R. C. Arkin and M. J. Pettinati. 2014. Moral emotions, robots, and their role in managing stigma in early stage parkinson's disease caregiving. In *Workshop on New Frontiers of Service Robotics for the Elderly in IEEE International Symposiumon Robot and Human Interactive Communication*.
[46] T. Baltrušaitis, Amir Zadeh, Y. C. Lim, and L. P. Morency. 2018. OpenFace 2.0: Facial behavior analysis toolkit. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*.
[47] H. S. Barrows. 1993. An overview of the uses of standardized patients for teaching and evaluating clinical skills. *Acad. Med.* 68 (1993), 443–443.
[48] R. F. Baugh, G. J. Basura, K. E. Ishii, S. R. Schwartz, C. M. Drumheller, and R. Burkholder. 2013. Clinical practice guideline Bell's palsy. *Otolaryngol. Head Neck Surg.* 149, 3_suppl, S1–S27.
[49] C. Becker-Asano and H. Ishiguro. 2011. Evaluating facial displays of emotion for the android robot Geminoid F. In *Proceedings of the IEEE Workshop on Affective Computational Intelligence*.
[50] Mathias Benedek, David Daxberger, Sonja Annerer-Walcher, and Jonathan Smallwood. 2018. Are you with me? Probing the human capacity to recognize external/internal attention in others' faces. *Vis. Cogn.* 26, 7 (2018), 511–517.
[51] Ruha Benjamin. 2019. Race after technology: Abolitionist tools for the new jim code. *Soc. Forces* 98, 4 (2019), 1–3.
[52] T. Bickmore, A. Rubin, and S. Simon. 2020. Substance use screening using virtual agents: Towards automated screening, brief intervention, and referral to treatment (SBIRT). In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (IVA'20)*.
[53] K. R. Bogart and L. Tickle-Degnen. 2015. Looking beyond the face: A training to improve perceivers' impressions of people with facial paralysis. *Patient Educ. Counsel.*
[54] K. R. Bogart, L. Tickle-Degnen, and N. Ambady. 2014. Communicating without the face: Holistic perception of emotions of people with facial paralysis. *Basic Appl. Soc. Psychol.* 36, 4 (2014), 309–320.
[55] H. Boughrara, M. Chtourou, B. C. Amar, and L. Chen. 2016. Facial expression recognition based on a mlp neural network using constructive training algorithm. *Multimedia Tools Appl.* 75, 2 (2016), 709–731.
[56] C. L. Breazeal. 2001. *Designing sociable robots*. MIT press.
[57] R. Breuer and R. Kimmel. 2017. A deep learning perspective on the origin of facial expressions. arXiv:1705.01842.
[58] Adrian Bulat and Georgios Tzimiropoulos. 2017. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*. 1021–1030.
[59] J. Buolamwini and T. Gebru. 2018. Gender shades: intersectional accuracy disparities in commercial gender classification. In *Proceedings of Machine Learning Research at Conference on Fairness, Accountability, and Transparency*.
[60] J. K. Burgoon, N. Magnenat-Thalmann, M. Pantic, and A. Vinciarelli. 2017. *Social Signal Processing*. Cambridge University Press.
[61] Egor Burkov, Igor Pasechnik, Artur Grigorev, and Victor Lempitsky. 2020. Neural head reenactment with latent pose descriptors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13786–13795.
[62] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2019. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.* 43, 1 (2019), 172–186.
[63] M. Carbone, R. Piazza, and S. Condino. 2020. Commercially available head-mounted displays are unsuitable for augmented reality surgical guidance: A call for focused research for surgical applications. *Surg. Innov.* 27, 3 (2020), 254–255.
[64] D. F. Carter. 1969. Man-made Man: Anesthesiological medical human simulator. *J. Assoc. Adv. Med. Instrum.* 3, 2, 80–86.
[65] H. R. Champion and A. G. Gallagher. 2003. Surgical simulation–a 'good idea whose time has come'. *Br. J. Surg.* 3, 2 (2003), 80–86.
[66] C. Chen, O. G. B. Garrod, J. Zhan, J. Beskow, P. G. Schyns, and R. E. Jack. 2018. Reverse engineering psychologically valid facial expressions of emotion into social robots. In *IEEE International Conference on Automatic Face and Gesture Recognition*.
[67] C. Chen, K. B. Hensel, Y. Duan, R. A. Ince, O. G. B. Garrod, J. Beskow, R. E. Jack, and P. G. Schyns. 2019. Equipping social robots with culturally-sensitive facial expressions of emotion using data-driven methods. In *IEEE International Conference on Automatic Face Gesture Recognition*.
[68] Wen-Sheng Chu, Fernando De la Torre, and Jeffrey F. Cohn. 2019. Learning facial action units with spatiotemporal cues and multi-label sampling. *Image Vis. Comput.* 81 (2019), 1–14.
[69] D. C. Classen, R. Resar, F. Griffin, F. Federico, T. Frankel, N. Kimmel, J. C. Whittington, A. Frankel, A. Seger, and B. C. James. 2011. Global trigger tool" shows that adverse events in hospitals may be ten times greater than previously measured. *Health Affairs*. 30, 4 (2011), 581–589.

[70] D. Cristinacce and T. Cootes. 2008. Automatic feature localisation with constrained local models. *J. Pattern Recogn. Soc.* 41, 10, 3054–3067.
[71] S. Daher, J. Hochreiter, N. Norouzi, L. Gonzalez, G. Bruder, and G. Welch. 2018. Physical-virtual agents for healthcare simulation. In *Proceedings of the International Conference on Intelligent Virtual Agents*.
[72] S. Daher, J. Hochreiter, R. Schubert, L. Gonzalez, J. Cendan, M. Anderson, D. A Diaz, and G. F. Welch. 2020. The physical-virtual patient simulator a physical human form with virtual appearance and behavior. *J. Soc. Simul. Healthcare*. 15, 2, 115–121.
[73] F. De la Torre, W. Chu, X. Xiong, F. Vicente, X. Ding, and J. F. Cohn. 2015. IntraFace. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*.
[74] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. 2020. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5203–5212.
[75] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. 2019. Retinaface: Single-stage dense face localisation in the wild. arXiv:1905.00641. Retrieved from https://arxiv.org/abs/1905.00641.
[76] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2014. DeCAF: A deep convolutional activation feature for generic visual recognition. In *Proceedings of the 31st International Conference on Machine Learning*. PMLR, 647–655.
[77] E. S. Donkor. 2018. Stroke in the 21 st Century: A snapshot of the burden, epidemiology, and quality of life. *NCBI Stroke Res. Treatm.*
[78] P. Ekman and W. Friesen. 1978. *Facial Action Coding System: Investigator's Guide*. Consulting Psychologists Press.
[79] P. Ekman, E. R. Sorenson, and W. V. Friesen. 1969. Pan-cultural elements in facial displays of emotion. *Science*. 164, 3875, 86–88.
[80] R. El Kaliouby and P. Robinson. 2005. Real-time inference of complex mental states from facial expressions and head gestures. In *Real-time Vision for Human-Computer Interaction*. Springer, New York, NY.
[81] Hillary Anger Elfenbein, Martin Beaupré, Manon Lévesque, and Ursula Hess. 2007. Toward a dialect theory: Cultural differences in the expression and recognition of posed facial expressions. *Emotion* 7, 1 (2007), 131.
[82] Jan B. Engelmann and Marianna Pogosyan. 2013. Emotion perception across cultures: the role of cognitive mechanisms. *Front. Psychol.* 4 (2013), 118.
[83] N. Ersotelos and F. Dong. 2008. Building highly realistic facial modeling and animation: A survey. *Vis. Comput.* 24, 1, 13–30
[84] Itir Onal Ertugrul, Jeffrey F. Cohn, László A. Jeni, Zheng Zhang, Lijun Yin, and Qiang Ji. 2020. Crossing domains for au coding: Perspectives, approaches, and measures. *IEEE Trans. Biometr. Behav. Ident. Sci.* 2, 2 (2020), 158–171.
[85] Itir Önal Ertugrul, Laszlo A. Jeni, and Jeffrey F. Cohn. 2019. PAttNet: Patch-attentive deep network for action unit detection. In *Proceedings of the British Machine Vision Conference (BMVC'19)*. 114.
[86] S. Ethier, W. J. Wilson, and C. Hulls. 2002. Telerobotic part assembly with shared visual servo control. In *Proceedings of the IEEE International Conference on Robotics and Automation*.
[87] Beat Fasel. 2002. Robust face analysis using convolutional neural networks. In *Object Recognition Supported by User Interaction for Service Robots*, Vol. 2. IEEE, 40–43.
[88] A. E. Frank, A. Kubota, and L. D. Riek. 2019. Wearable activity recognition for robust human-robot teaming in safety-critical environments via hybrid neural networks. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'19)*.
[89] Baris Gecer, Alexandros Lattas, Stylianos Ploumpis, Jiankang Deng, Athanasios Papaioannou, Stylianos Moschoglou, and Stefanos Zafeiriou. 2020. Synthesizing coupled 3d face modalities by trunk-branch generative adversarial networks. In *Proceedings of the European Conference on Computer Vision*. Springer, 415–433.
[90] M. Ghayoumi. 2017. A quick review of deep learning in facial expression. *J. Commun. Comput.* 14, 1(2017), 34–8.
[91] K. H. Glantz. 1996. Conducting research with children: Legal and ethical issues. *J. Am. Acad. Child Adolesc. Psychiatr.* 35, 10, 1283–1291.
[92] M. A. Goodrich and A. C. Schultz. 2007. Human-Robot Interaction: A Survey. Now Publishers Inc.
[93] T. Gorman, J. Dropkin, J. Kamen, S. Nimbalkar, N. Zuckerman, T. Lowe, J. Szeinuk, D. Milek, G. Piligian, and A. Freund. 2013. Controlling health hazards to hospital workers. *New Solutions*. 23, 1_suppl (2014), 1–169.
[94] S. J. Goyal, A. K. Upadhyay, R. S. Jadon, and R. Goyal. 2018. Real-Life facial expression recognition systems: A review. *Smart Comput. Inf.* 77, 311–331.
[95] J. D. Greer, T. K. Morimoto, A. M. Okamura, and E. W. Hawkes. 2019. A soft, steerable continuum robot that grows via tip extension. *Soft Robot.* 6, 1 (2019), 95–108.
[96] J. Hamm, C. G. Kohler, R. C. Gur, and R. Vermaa. 2011. Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders. *J. Neurosci. Methods*. 200, 2 (2011), 237–256.
[97] T. Hashimoto, S. Hitramatsu, T. Tsuji, and H. Kobayashi. 2006. Development of the face robot SAYA for rich facial expressions. In *Proceedings of the SICE-ICASE International Joint Conference*.
[98] P. Hellyer. 2019. Preventable patient harm is expensivE. *Br. Dent. J.* 227, 4 (2019), 275–275.
[99] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861. Retrieved from https://arxiv.org/abs/1704.04861.

[100] Peiyun Hu and Deva Ramanan. 2017. Finding tiny faces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 951–959.
[101] W. Huang. 2015. When HCI Meets HRI: the intersection and distinction. *ACM Special Interest Group Comput.-Hum. Interac.*
[102] Y. Huang, F. Chen, S. Lv, and X. Wang. 2019. Facial expression recognition: A survey. *Symmetry*. 11, 10 (2019), 1189.
[103] R. E. Jack, O. G. Garrod, and P. G. Schyns. 2014. Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Curr. Biol.* 24, 2 (2014) 187–192.
[104] Rachael E. Jack. 2013. Culture and facial expressions of emotion. *Vis. Cogn.* 21, 9-10 (2013), 1248–1286.
[105] Rachael E. Jack, Wei Sun, Ioannis Delis, Oliver G. B. Garrod, and Philippe G. Schyns. 2016. Four not six: Revealing culturally common facial expressions of emotion. *J. Exp. Psychol.: Gen.* 145, 6 (2016), 708.
[106] J. T. James. 2013. A new evidence-based estimate of patient harms associated with hospital care. *J. Patient Safe.* 9, 3 (2013), 122–128.
[107] P. R. Jeffries. 2007. Simulation in nursing education: From conceptualization to education. In *National League for Nursing*.
[108] A. Kalegina, G. Schroeder, A. Allchin, K. Berlin, and M. Cakmak. 2018. Characterizing the design space of rendered robot faces. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*.
[109] Sebastian Kaltwang, Ognjen Rudovic, and Maja Pantic. 2012. Continuous pain intensity estimation from facial expressions. In *Proceedings of the International Symposium on Visual Computing*. Springer, 368–377.
[110] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. 2019. Face Behavior a la carte: Expressions, affect and action units in a single network. arXiv:1910.11111. Retrieved from https://arxiv.org/abs/1910.11111.
[111] Dimitrios Kollias and Stefanos Zafeiriou. 2018. Aff-wild2: Extending the aff-wild database for affect recognition. arXiv:1811.07770. Retrieved from https://arxiv.org/abs/1811.07770.
[112] A. A. Kononowicz, K. A. Woodham, S. Edelbring, N. Stathakarou, D. Davies, N. Saxena, K. T. Car, J. Carlstedt-Duke, J. Car, and N. Zary. 2019. Virtual patient simulations in health professions education: systematic review and meta-analysis by the digital health education collaboration. *J. Med. Internet Res.* 21, 7, e14676.
[113] Jean Kossaifi, Robert Walecki, Yannis Panagakis, Jie Shen, Maximilian Schmitt, Fabien Ringeval, Jing Han, Vedhas Pandit, Antoine Toisoul, Bjoern W Schuller, et al. 2019. Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE Trans. Pattern Anal. Mach. Intell.* 43, 3, 1022–1040.
[114] Quoc V. Le. 2013. Building high-level features using large scale unsupervised learning. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 8595–8598.
[115] M. J. Leo and D. Manimegalai. 2011. 3D modeling of human faces—A survey. In *Proceedings of the International Conference on Trends in Information Sciences Computing*.
[116] Jiaxing Li, Dexiang Zhang, Jingjing Zhang, Jun Zhang, Teng Li, Yi Xia, Qing Yan, and Lina Xun. 2017. Facial expression recognition with faster R-CNN. *Proc. Comput. Sci.* 107 (2017), 135–140.
[117] Shan Li and Weihong Deng. 2020. Deep facial expression recognition: A survey. *IEEE Trans. Affect. Comput.*
[118] Wei Li, Farnaz Abtahi, and Zhigang Zhu. 2017. Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1841–1850.
[119] Y. Li, J. Zeng, S. Shan, and X. Chen. 2019. Occlusion aware facial expression recognition using CNN with attention mechanism. *IEEE Trans. Image Process.* 28, 5 (2019), 2439–2450.
[120] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. 2019. Self-supervised representation learning from videos for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10924–10933.
[121] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett. 2011. The computer expression recognition toolbox (CERT). In *Proceedings of the International Conference on Automatic Face & Gesture Recognition and Workshops*.
[122] Patrick Lucey, Jeffrey F. Cohn, Kenneth M. Prkachin, Patricia E. Solomon, and Iain Matthews. 2011. Painful data: The UNBC-McMaster shoulder pain expression archive database. In *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (FG'11)*. IEEE, 57–64.
[123] Q. Mao, Q. Rao, Y. Yu, and M. Dong. 2017. Hierarchical Bayesian theme models for multipose facial expression recognition. *IEEE Trans. Multimedia*. 19, 4 (2017), 861–873.
[124] A. G. Marson and R. Salinas. 2000. Clinical evidence: Bell's palsy. *West. J. Med*. 173, 4, 266.
[125] Brais Martinez, Michel F. Valstar, Bihan Jiang, and Maja Pantic. 2017. Automatic analysis of facial actions: A survey. *IEEE Trans. Affect. Comput.* 10, 3 (2017), 325–347.
[126] S. Mohammad Mavadati, Mohammad H. Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F. Cohn. 2013. Disfa: A spontaneous facial action intensity database. *IEEE Trans. Affect. Comput.* 4, 2 (2013), 151–160.
[127] D. Mazzei, N. Lazzeri, D. Hanson, and D. De Rossi. 2012. HEFES: An hybrid engine for facial expressions synthesis to control human-like androids and avatars. In *Proceedings of the IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics*.
[128] D. McDuff, A. Mahmoud, M. Mavadati, M. Amr, J. Turcot, and R.E. Kaliouby. 2016. AFFDEX SDK: A cross-platform realtime multi-face expression recognition toolkit. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*.
[129] Shervin Minaee, Ping Luo, Zhe Lin, and Kevin Bowyer. 2021. Going deeper into face detection: A survey. arXiv:2103.14983. Retrieved from https://arxiv.org/abs/2103.14983.

[130] M. Moosaei, S. K. Das, D. O. Popa, and L. D. Riek. 2017. Using facially expressive robots to calibrate clinical pain perception. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*.
[131] M. Moosaei, S. K. Das, D. O. Popa, and L. D. Riek. 2017. Using facially expressive robots to calibrate clinical pain perception. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*.
[132] M. Moosaei, M. J. Gonzales, and L. D. Riek. 2014. Naturalistic pain synthesis for virtual patients. In *Proceedings of the International Conference on Intelligent Virtual Agents*.
[133] M. Moosaei, C. J. Hayes, and L. D. Riek. 2015. Facial expression synthesis on robots: An ROS module. In *Proceedings of the Annual ACM/IEEE International Conference on Human-Robot Interaction*.
[134] M. Moosaei, M. Pourebadi, and L. D. Riek. 2019. Modeling and synthesizing idiopathic facial paralysis. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG'19)*.
[135] M. Mori. 2012. The uncanny valley: The original essay by masahiro mori. *IEEE Robot. Autom. Mag*.
[136] E. Moy, E. Valente, R. Coffey, and A.K. Hines. 2014. Missed diagnosis of stroke in the emergency department: A cross-sectional analysis of a large population-based sample. *Diagnosis* 1, 2, (2014), 155–166.
[137] N. L. Nelson and J. A. Russell. 2013. Universality Revisited. *Emotion Review SAGE Journals* 5, 1 (2013), 8–15.
[138] S. Nishio, H. Ishiguro, and N. Hagita. 2007. Geminoid: Teleoperated android of an existing person. In *Humanoid Robots: New Developments*. 14, 343–352.
[139] Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.
[140] Safiya Umoja Noble. 2020. Tech won't save us: Reimagining digital technologies for the public. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*. 1–1.
[141] Ioanna Ntinou, Enrique Sanchez, Adrian Bulat, Michel Valstar, and Yorgos Tzimiropoulos. 2021. A transfer learning approach to heatmap regression for action unit intensity estimation. *IEEE Trans. Affect. Comput.* (2021), 1–1.
[142] Naima Otberdout, Anis Kacem, Mohamed Daoudi, Lahoucine Ballihi, and Stefano Berretti. 2018. Deep covariance descriptors for facial expression recognition. arXiv:1805.03869. Retrieved from https://arxiv.org/abs/1805.03869.
[143] H. Owen. 2012. Early use of simulation in medical education. *J. Soc. Simul. Healthc.* 7, 2 (2012), 102–116.
[144] M. Pantic and M. S. Bartlett. 2007. Machine analysis of facial expressions. In *Face Recognition*, K. Delac and M. Grgic (Eds.). IntechOpen.
[145] C. Pelachaud. 2015. Greta, an interactive expressive embodied conversational agent. In *Proceedings of the International Conference on Autonomous Agents & Multiagent Systems*.
[146] C. Pelachaud. 2017. Greta: A conversing socio-emotional agent. In *Proceedings of the ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents*.
[147] B. Pierce, T. Kuratate, C. Vogl, and G. Cheng. 2012. DMask-Bot 2i": An active customisable Robotic Head with Interchangeable Face. In *Proceedings of the IEEE-RAS International Conference on Humanoid Robots*.
[148] H. Y. Ping, K. N. Abdullah, P. S. Sulaiman, and A. A. Halin. 2013. Computer facial animation: A review. *Int. J. Comput. Theory Eng*. 5, 4 (2013), 658.
[149] M. Pourebadi, J. N. LaBuzetta, C. Gonzalez, P. Suresh, and L. D. Riek. 2019. Mimicking acute stroke findings with a digital avataR. In *Proceedings of the International Stroke Conference (ISC'19) in AHA/ASA Journal*.
[150] M. Pourebadi and M. Pourebadi. 2016. MLP neural network based approach for facial expression analysis. In *Proceedings of the World Congress in Computer Science, Computer Engineering and Applied Computing (WORLDCOMP'16)*.
[151] M. Pourebadi and L. D. Riek. 2018. Expressive robotic patient simulators for clinical education. In *Proceedings of the R4L Workshop on Robots for Learning—Inclusive Learning at the 13th Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI'18)*.
[152] M. Pourebadi and L. D. Riek. 2020. Stroke modeling and synthesis for robotic and virtual patient simulators. In *Proceedings of the AAAI Artificial Intelligence for Human-Robot Interaction (AAAI AI-HRI'20): Trust & Explainability in Artificial Intelligence for Human-Robot Interaction*.
[153] Varsha Powar and Aditi Jahagirdar. 2012. Reliable face detection in varying illumination and complex background. In *Proceedings of the International Conference on Communication, Information & Computing Technology (ICCICT'12)*. IEEE, 1–4.
[154] Albert Pumarola, Antonio Agudo, Aleix M. Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. 2018. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV'18)*. 818–833.
[155] Albert Pumarola, Antonio Agudo, Aleix M. Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. 2020. Ganimation: One-shot anatomically consistent facial animation. *Int. J. Comput. Vis.* 128, 3 (2020), 698–713.
[156] M. Ramacciotti, M. Milazzo, F. Leoni, S. Roccella, and C. Stefanini. 2016. A novel shared control algorithm for industrial robots. *Int. J. Adv. Robot. Syst*. 13, 6 (2016), 729881416682701.
[157] I. M. Revina and W. R. S. Emmanuel. 2018. A survey on human face expression recognition techniques. *J. King Saud Univ. Comput. Inf. Sci.* 33, 6 (2018), 619–628.
[158] L. D. Riek. 2011. *Expression Synthesis on Robots*. Ph.D. Dissertation. University of Cambridge.
[159] L. D. Riek. 2016. System and method for robotic patient synthesis. US Patent 9,280,147.
[160] L. D. Riek. 2017. Healthcare robotics. *Commun. ACM* 60, 11 (2017), 68–78.

[161] L. D. Riek and P. Robinson. 2011. Using robots to help people habituate to visible disabilities. In *2011 IEEE International Conference on Rehabilitation Robotics*. IEEE, 1–8.
[162] Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. 2011. Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the International Conference on Machine Learning*.
[163] D. Rivera-Gutierrez, G. Welch, P. Lincoln, M. Whitton, J. J. Cendan, D. Chesnutt, H. Fuchs, and B. Lok. 2012. Shader lamps virtual patients: The physical manifestation of virtual patients. In *Medicine Meets Virtual Reality 19*. IOS Press, 372–378.
[164] K. L. Robey, P. M. Minihan, K. M. Long-Bellil, J. E. Hahn, J. G. Reiss, G. E. Eddey, Alliance for Disability in Health Care Education, et al. 2013. Teaching health care students about disability within a cultural competency context. *Disabil. Health J.* 6, 4, 271–279.
[165] S. Robla-Gómez, V.M. Becerra, J.R. Llata, E. González-Sarabia, C. Torre-Ferrero, and J. Pérez-Oria. 2017. Working together: A review on safe human-robot collaboration in industrial environments. *IEEE Access*. 5 (2017), 26754–26773.
[166] T. L. Rodziewicz and J. E. Hipskind. 2020. *Medical Error Prevention*. StatPearls Publishing.
[167] H. Salam and R. Séguier. 2018. A survey on face modeling: Building a bridge between face analysis and synthesis. *Vis. Comput*. 34, 2 (2018), 289–319.
[168] Enrique Sanchez, Adrian Bulat, Anestis Zaganidis, and Georgios Tzimiropoulos. 2020. Semi-supervised facial action unit intensity estimation with contrastive learning. In *Proceedings of the Asian Conference on Computer Vision*.
[169] Ayse Pinar Saygin, Thierry Chaminade, Hiroshi Ishiguro, Jon Driver, and Chris Frith. 2012. The thing that should not be: Predictive coding and the uncanny valley in perceiving human and humanoid robot actions. *Soc. Cogn. Affect. Neurosci.* 7, 4 (2012), 413–422.
[170] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 806–813.
[171] S. R. Shrivastava, P. S. Shrivastava, and J. D. Ramasamy. 2013. Reduction in global burden of stroke in underserved areas. *J. Neurosci. Rur. Pract.*. 4, 4 (2013), 475–476.
[172] Ben Singer. 2013. The human simulation lab—Dissecting sex in the simulator lab: The clinical lacuna of transsexed embodiment. *J. Med. Human.* 34, 2 (2013), 249–254.
[173] S. Stockli, Schulte-Mecklenbeck, S. M. Borer, and A. C. Samson. 2018. Facial expression analysis with AFFDEX and FACET: A validation study. *Behavior Research Methods*. 50, 4 (2018), 1446–1460.
[174] M. Strait, A. Ramos, V. Contreras, and N. Garcia. 2018. Robots racialized in the likeness of marginalized social identities are subject to greater dehumanization than those racialized as white. In *Proceedings of the IEEE-RAS International Conference on Humanoid Robots*.
[175] C. Suarez, M. D. Menendez, J. Alonso, N. Castaño, M. Alonso, and F. Vazquez. 2014. Detection of adverse events in an acute geriatric hospital over a 6-year period using the global trigger tool. *J. Am. Geriatr. Soc.* 62, 5 (2014), 896–900.
[176] Bo Sun, Liandong Li, Guoyan Zhou, Xuewen Wu, Jun He, Lejun Yu, Dongxue Li, and Qinglan Wei. 2015. Combining multimodal features within a fusion network for emotion recognition in the wild. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 497–502.
[177] Y. Sun, X. Wang, and X. Tang. 2013. Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[178] H. Sunvisson, B. Habermann, S. Weiss, and P. Benner. 2009. Augmenting the Cartesian medical discourse with an understanding of the person's lifeworld, lived body, life story, and social identity. *Nurs. Philos.* 10, 4 (2009), 241–252.
[179] L. Surace, M. Patacchiola, E. Battini Sönmez, W. Spataro, and A. Cangelosi. 2017. Emotion recognition in the wild using deep neural networks and Bayesian classifiers. In *Proceedings of the ACM International Conference on Multimodal Interaction*.
[180] A. A. Tarnutzer, S. Lee, K. A. Robinson, Z. Wang, J. A. Edlow, and D. E. Newman-Toker. 2017. ED misdiagnosis of cerebrovascular events in the era of modern neuroimaging. *J. Am. Acad. Neurol*. 88, 15 (2017), 1468–1477.
[181] A. Taylor, H. Lee, A. Kubota, and L.D. Riek. 2019. Simulation-based medical teaching and learning. In *Proceedings of the ACM Conference on Computer Supported Collaborative Work*.
[182] A. Taylor, S. Matsumoto, and L. D. Riek. 2020. Situating robots in the emergency department. In *Proceedings of the AAAI Spring Symposium on Applied AI in Healthcare: Safety, Community, and the Environment*.
[183] Angelique M. Taylor, Sachiko Matsumoto, Wesley Xiao, and Laurel D. Riek. 2021. Social navigation for mobile robots in the emergency department. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'21)*.
[184] M. M. Tedesco, J. J. Pak, E. J. Harris, T. M. Krummel, R. L. Dalman, and J. T. Lee. 2007. Simulation-based endovascular skills assessment: The future of credentialing? *J. Vasc. Surg*. 47, 5 (2017), 1008–1014.
[185] Y. I. Tian, T. Kanade, and J. F. Cohn. 2001. Recognizing action units for facial expression analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*
[186] L. Tickle-Degnen, K.A. Zebrowitz, and H. Ma. 2011. Culture, gender, and health care stigma: Practitioners' response to facial masking experienced by people with Parkinson's disease. *Soc. Sci. Med*. 73, 1 (2011), 95–102.
[187] R. Triebel, K. Arras, R. Alami, L. Beyer, S. Breuers, R. Chatila, M. Chetouani, D. Cremers, V. Evers, M. Fiore, H. Hung, O. A. I. Ramírez, M. Joosse, H. Khambhaita, T. Kucner, B. Leibe, A. J. Lilienthal, T. Linder, M. Lohse, M. Magnusson, B. Okal, L. Palmieri, U. Rafi, M. van Rooij, and L. Zhang. 2016. *SPENCER: A Socially Aware Service Robot for Passenger Guidance and Help in Busy Airports*. In *Field and Service Robotics*. Springer, 607–622.

[188] T. Tsai. 2004. Using children as standardised patients for assessing clinical competence in paediatrics. *Arch. Dis. Childhood*. 89, 12 (2004), 1117–1120.
[189] O. Tysnes and A. Storstein. 2017. Epidemiology of Parkinson's disease. *J. Neur. Transmiss.* 5, 6 (2017), 525–535.
[190] M. Unbeck, K. Schildmeijer, P. Henriksson, U. Jürgensen, O. Muren, L. Nilsson, and K. Pukk Härenstam. 2013. Is detection of adverse events affected by record review methodology? An evaluation of the "Harvard Medical Practice Study" method and the "Global Trigger Tool." *Patient Safety Surg*. 7, 1 (2013), 1–12.
[191] Burcu A. Urgen, Marta Kutas, and Ayse P. Saygin. 2018. Uncanny valley as a window into predictive processing in the social brain. *Neuropsychologia* 114 (2018), 181–185.
[192] P. Viola and M. Jones. 2001. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*
[193] A. W. Walter, C. Julce, N. Sidduri, L. Yinusa-Nyahkoon, J. Howard, M. Reichert, T. Bickmore, and B. W. Jack. 2020. Study protocol for the implementation of the Gabby Preconception Care System—An evidence-based, health information technology intervention for Black and African American women. In *BMC Health Services Research.* 20, 1, 1–14.
[194] N. Wang, X. Gao, D. Tao, and X. Li. 2017. Facial feature point detection: A comprehensive survey. *Neurocomputing*. 275 (2017), 50–65.
[195] Shangfei Wang, Bowen Pan, Shan Wu, and Qiang Ji. 2019. Deep facial action unit recognition and intensity estimation from partially labelled data. *IEEE Trans. Affect. Comput.* 12 (2019), 1018–1030.
[196] Shangfei Wang and Guozhu Peng. 2019. Weakly supervised dual learning for facial action unit recognition. *IEEE Trans. Multimedia* 21, 12 (2019), 3218–3230.
[197] W. Wang, Q. Sun, T. Chen, C. Cao, Z. Zheng, G. Xu, H. Qiu, and Y. Fu. 2019. A fine-grained facial expression database for end-to-end multi-pose facial expression recognition. arXiv:1907.10838.
[198] C. Watson and T. K. Morimoto. 2020. Permanent magnet-based localization for growing robots in medical applications. *IEEE Robot. Autom. Lett.* 5, 2, 2666–2673.
[199] Philipp Werner, Daniel Lopez-Martinez, Steffen Walter, Ayoub Al-Hamadi, Sascha Gruss, and Rosalind Picard. 2019. Automatic recognition methods supporting pain assessment: A survey. *IEEE Trans. Affect. Comput.* 1–1.
[200] A. Williams. 2002. Facial expression of pain: An evolutionary account. *Behav. Brain Sci.* 25, 4 (2002), 439–455.
[201] A. D. Wilson and S. N. Bathiche. 2013. Compact interactive tabletop with projection-vision. US10026177B2 Patent.
[202] M. S. Wilson, A. Middlebrook, C. Sutton, R. Stone, and R. F. McCloy. 1997. MIST VR: A virtual reality trainer for laparoscopic surgery assesses performance. *Ann. Roy. Coll. Surg. Engl.* 79, 6 (1997), 403.
[203] Q. Wu, L. Zhao, and X. Ye. 2016. Shortage of healthcare professionals in China. *Br. Med. J.* 354.
[204] Yue Wu and Qiang Ji. 2016. Constrained joint cascade regression framework for simultaneous facial action unit recognition and facial landmark detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 3400–3408.
[205] Y. Wu and Q. Ji. 2018. Facial landmark detection: A literature survey. *Int. J. Comput. Vis.* 127, 2 (2019), 115–142.
[206] A. Zadeh, T. Baltrušaitis, and L. P. Morency. 2017. Constrained local model for facial landmark detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops.*
[207] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. 2019. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 9459–9468.
[208] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Sign. Process. Lett.* 23, 10 (2016), 1499–1503.
[209] Xing Zhang, Lijun Yin, Jeffrey F. Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M. Girard. 2014. Bp4d-spontaneous: A high-resolution spontaneous 3d dynamic facial expression database. *Image Vis. Comput.* 32, 10 (2014), 692–706.
[210] Yong Zhang, Haiyong Jiang, Baoyuan Wu, Yanbo Fan, and Qiang Ji. 2019. Context-aware feature and label fusion for facial action unit intensity estimation with partially labeled data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 733–742.
[211] X. Zhao, X. Shi, and S. Zhang. 2012. Facial expression recognition via deep learning. In *Proceedings of the IEEE/ACS International Conference on Computer Systems and Applications.*
[212] I. Zubrycki, I. Szafarczyk, and G. Granosik. 2018. Project fantom: Co-Designing a robot for demonstrating an epileptic seizure. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication.*

Received March 2021; revised July 2021; accepted August 2021