Spring 2024

Computational Modeling and Analysis of Facial Expressions and Gaze for Discovery of Candidate Behavioral Biomarkers for Children and Young Adults with Autism Spectrum Disorder

Megan Anita Witherow
Old Dominion University, witherowma@icloud.com

Recommended Citation

Witherow, Megan A.. "Computational Modeling and Analysis of Facial Expressions and Gaze for Discovery of Candidate Behavioral Biomarkers for Children and Young Adults with Autism Spectrum Disorder" (2024). Doctor of Philosophy (PhD), Dissertation, Electrical & Computer Engineering, Old Dominion University, DOI: 10.25777/w99f-ts51
https://digitalcommons.odu.edu/ece_etds/261

COMPUTATIONAL MODELING AND ANALYSIS OF FACIAL EXPRESSIONS AND GAZE FOR DISCOVERY OF CANDIDATE BEHAVIORAL BIOMARKERS FOR CHILDREN AND YOUNG ADULTS WITH AUTISM SPECTRUM DISORDER

by

Megan Anita Witherow
B.S. May 2018, Old Dominion University

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

ELECTRICAL AND COMPUTER ENGINEERING

OLD DOMINION UNIVERSITY
May 2024

Approved by:

Khan M. Iftekharuddin (Director)

Jiang Li (Member)

Norou Diawara (Member)

John W. Harrington (Member)

ABSTRACT

COMPUTATIONAL MODELING AND ANALYSIS OF FACIAL EXPRESSIONS AND GAZE FOR DISCOVERY OF CANDIDATE BEHAVIORAL BIOMARKERS FOR CHILDREN AND YOUNG ADULTS WITH AUTISM SPECTRUM DISORDER

Megan Anita Witherow
Old Dominion University, 2024
Director: Dr. Khan M. Iftekharuddin

Facial expression production and perception in autism spectrum disorder (ASD) suggest the potential presence of behavioral biomarkers that may stratify individuals on the spectrum into prognostic or treatment subgroups. High-speed internet and the ease of technology have enabled remote, scalable, affordable, and timely access to medical care, such as measurements of ASD-related behaviors in familiar environments to complement clinical observation. Machine and deep learning (DL)-based analysis of video tracking (VT) of expression production and eye tracking (ET) of expression perception may aid stratification biomarker discovery for children and young adults with ASD. However, there are open challenges in 1) facial expression analysis (FEA) across age groups to overcome domain shift between child and adult expressions, 2) Facial Action Coding System (FACS)-labeled 3D avatar-based stimuli to improve user engagement for eliciting expressions, and 3) assessment of construct validity and group discriminability criteria to discover candidate biomarkers for ASD.

Consequently, this dissertation proposes three goals. The proposed dissertation goals have been completed in collaboration and consultation with a team of Old Dominion University and Eastern Virginia Medical School investigators. The first proposed aim is a novel deep domain adaptation fusing DL-based texture features with geometric landmark features for generalized child/adult FEA. Novel facial feature selection for DL is performed using a new statistical method based on a mixture of beta distributions. Our model performs competitively over transfer learning and existing domain adaptation methods using multiple benchmark data sets. Second, we propose FACS-labeled customizable avatars for improved user engagement and DL models for multi-label FACS action unit (AU) detection. The DL models incorporate feature fusion, multi-task learning of AUs and expressions, and a novel beta-guided correlation loss to achieve state-of-the-art AU detection performance on our primary benchmark data set. We report the construct validity of proposed stimuli and measurements based on a feasibility study of twenty healthy adults. Finally, we conduct an online pilot study of 11 autistic children and young adults and 11 age-/gender-matched neurotypical individuals. Webcam-based ET and VT are collected while participants recognize and mimic avatar expressions. Extensive statistical analyses, including evaluation of construct validity and group discriminability, identify one candidate ET biomarker and 14 additional ET and VT measurements that may be candidates for more comprehensive future studies with increased sample size for validation and clinical translation.

This dissertation is dedicated to my parents,
Edwin C. Witherow and Deborah C. Witherow, and
In loving memory of Yukiko T. Witherow and Jerry B. Ward.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor, Dr. Khan Iftekharuddin, for the opportunity to work in his Vision Lab ever since I first got started as an undergraduate researcher back in Fall 2015, for his mentorship and encouragement that motivated me to embark on this PhD, and for his guidance, patience, and support throughout the completion of my dissertation research. I would also like to thank the members of my advisory committee, Dr. Jiang Li, Dr. Norou Diawara, and Dr. John Harrington, for their service and their willingness to share their time, expertise, and valuable feedback.

I would like to express my deepest gratitude to my parents, Eddie and Debbie Witherow, who have inspired me and provided me with endless love and support. I would also like to express my heartfelt appreciation to the many other family members and dear friends who have cheered me on throughout my PhD, including my grandfather and grandmother, Marshall and Julie Cherry; my grandmother and late grandfather, Peggy and Jerry Ward; my late grandmother, Yuki Witherow; and my brother, Kevin Witherow. I would like to thank my fiancé, Charles Chiou, for his continuous support and encouragement throughout my studies.

For funding my research activities, I am grateful to the National Science Foundation (NSF) for the NSF Graduate Research Fellowship (Grant Nos. 1753793 and 2139907), to the Batten College of Engineering and Technology for recruitment and retention fellowships, and to Dr. Iftekharuddin for an assistantship, funding to purchase supplies for the research study, and funding for participant incentives. I would also like to thank ODU Research Computing Services for the use of the Wahab and Turing clusters, which are supported by NSF under Grant No. 1828593.

I would like to acknowledge all of the volunteers who participated in the research. Without them, this dissertation would not have been possible.

I would like to give special thanks to Manar Samad, from whom I inherited the Vision Lab 'Autism Project' and who has been an important mentor for me. I would also like to recognize the undergraduate students who worked with me on this project: Abigail Stedman, Winston Shields, Gregory Hubbard, Elija Bullock, and Furkan Ilgin. Finally, I would like to acknowledge the many other mentors, friends, and esteemed colleagues that I have met over the years in the Vision Lab: Mahbubul Alam, Lasitha Vidyaratne, Alexander Glandon, Linmin Pei, Zeina Shboul, Monibor Rahman, Kwame Ampofo, Walia Farzana, Joseph Zalameda, and many others. Thank you to all my colleagues at the Vision Lab for your support, inspiration, and collaboration.

NOMENCLATURE

%Acc	Recognition Accuracy of Participants
%Gaze Face	Percentage of Gaze Duration to the Avatar’s Face
ABC-CT	Autism Biomarkers Consortium for Clinical Trials
ACT	Neural Network Predicted Activation of Participants’ Action Units
ADAM	Adaptive Moment Estimation
ADOS-2	Autism Diagnostic Observation Schedule, Second Edition
Aff-Wild	Affect-in-the-Wild
AIFR	Age-Invariant Face Recognition
AOI	Area of Interest
ASD	Autism Spectrum Disorder
ASYM	Left-Right Asymmetry of Neural Network Predicted Activations of Participants’ Action Units
AU	Action Unit
AUC	Area Under the ROC Curve
BeCoME-Net	Beta-guided Correlation and Multi-task Expression Network
betaMix	Beta-Mixture Method
BVAQ	Bermond-Vorst Alexithymia Questionnaire
CADyFACE	Customizable Avatars with Dynamic Facial Action Coded Expressions
CAFE	Child Affective Facial Expression
CAM	Children’s Alexithymia Measure
CART	Classification and Regression Trees
CD	Customized Avatar with Dynamic Expressions

CDC	Centers for Disease Control and Prevention
ChildEFES	Child Emotion Facial Expression Set
CK+	Extended Cohn-Kanade
CIT	Conditional Inference Tree
CMTF	Cambridge Memory Test of Faces for Children
CNN	Convolutional Neural Network
CS	Customized Avatar with Static Expressions
DISFA+	Extended Denver Intensity of Spontaneous Facial Action
DSM-5	Diagnostic and Statistical Manual of Mental Disorders
DV	Dependent Variable
EEG	Electroencephalography
EM	Expectation-Maximization
ET	Eye Tracking
EU-AIMS	European Autism Interventions—A Multicentre Study for Developing New Medications
EXPR	Neural Network Predicted Softmax Probability Corresponding to the Participants' Ability to Mimic the Avatar's Overall Expression
FACS	Facial Action Coding System
FACE-BE-SELF	FACial Expressions fusing BEtaMix SElected Landmark Features
FDA	Food & Drug Administration
FEA	Facial Expression Analysis
fMRI	Functional Magnetic Resonance Imaging
FN	False Negatives

FP	False Positives
HFA	Hidden Factor Analysis
ID3	Iterative Dichotomiser 3
i.i.d.	Independent and Identically Distributed
IRB	Institutional Review Board
IQ	Intelligence Quotient
JAKE	Janssen Autism Knowledge Engine
KBIT-2	Kaufman Brief Intelligence Test, Second Edition
KNN	K-Nearest Neighbors
LASSO	Least Absolute Shrinkage and Selection Operator
LEAP	Longitudinal European Autism Project
LOOCV	Leave-One-Out Cross-Validation
MAE	Mean Absolute Error
MHFA	Modified Hidden Factor Analysis
MICE	Multiple Imputation by Chained Equations
MIM	Mimicry
MLP	Multilayer Perceptron
MSE	Mean Squared Error
NaN	Not a Number
NT	Neurotypical
OMI	Oculomotor Index of Gaze to Human Faces
PCA	Principal Component Analysis
perc	Percentile of Shadow Feature’s Importance

REC	Recognition
ReLU	Rectified Linear Unit
REMT	Reading the Mind in the Eyes Task
RMSE	Root Mean Squared Error
ROC	Receiver Operating Characteristic
SGD	Stochastic Gradient Descent
SHAP	SHapley Additive exPlanations
SPARK	Simons Foundation Powering Autism Research for Knowledge
TN	True Negatives
TP	True Positives
UD	Uncustomized Avatar with Dynamic Expressions
US	Uncustomized Avatar with Static Expressions
VT	Video Tracking

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

CHAPTER 1

INTRODUCTION

Autism Spectrum Disorder (ASD) is a heterogeneous neurodevelopmental condition. Based on a 2020 surveillance survey, the United States Centers for Disease Control and Prevention (CDC) estimates that ASD affects 1 in 36 children in the United States [1]. This latest report reflects a trend of increasing prevalence (Figure 1) since the first survey in 2000 [2]. There are currently no validated biomarkers for ASD [3, 4]. Instead, diagnosis and ongoing clinical assessment are informed by direct visual observation, parent interviews, and psychological testing [4, 5]. In addition, the United States faces a shortage of ASD specialists, leading to an overworked developmental-behavioral pediatrics workforce and delays in accessing ASD-related care for patients and families [6]. With motivation to improve access to care, reduce wait times, and alleviate stress on the current ASD support system, scalable, automated tools based on computer vision and machine learning have been increasingly applied to ASD research over the past decade [7]. Especially during and since the COVID-19 pandemic, high-speed internet connectivity and the ease of technology have enabled access to children outside of clinical settings, improving access to care and engaging families that may not normally have access to services [8-11]. Computer vision and machine learning enable ecologically valid, precise measurements of ASD-related behaviors at home, school, and in community settings to complement ongoing clinical care, increase access to services, and reduce barriers to research participation [8-11].

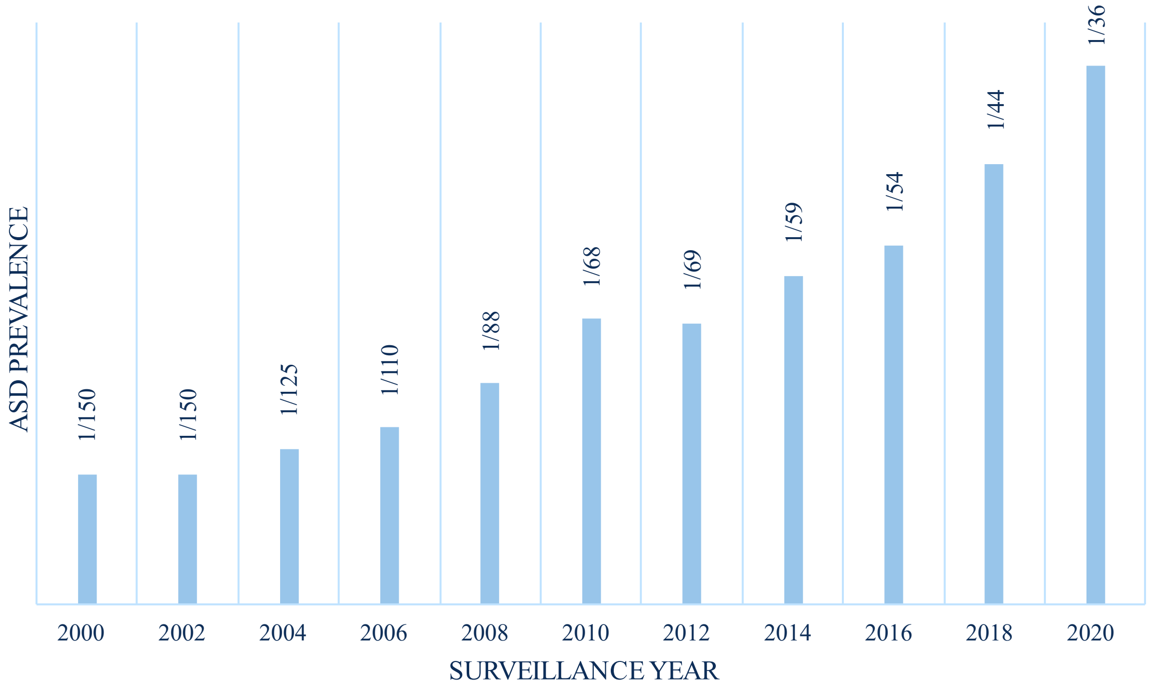


Figure 1. Prevalence of ASD from 2000 to 2020 [2].

Individuals on the autism spectrum may have a wide variety of behavioral symptom profiles and experiences. According to the 5th edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5), symptoms of ASD fall into the two broad categories of "persistent deficits in social communication and social interaction" and "restricted, repetitive patterns of behavior, interests, or activities" [12]. Among differences in social communication and social interaction, some individuals on the spectrum may produce and perceive facial expressions differently than their neurotypical (NT) peers [12]. These differences may be captured and quantified through eye tracking (ET) of gaze to facial stimuli and behavioral video tracking (VT) of facial behaviors [7]. Models for automated ET analysis and facial expression analysis (FEA) may help support autistic individuals and clinicians with rapid and objective assessments of ASD-related symptoms.

High heterogeneity in the perception and production of facial expressions among autistic individuals compared to NT individuals has long resulted in mixed findings regarding the nature of facial expressions in ASD [13]. However, while ASD research has historically focused on the search for diagnostic biomarkers that partition individuals into groups of NT and autistic individuals [13], the focus has recently shifted to the identification of stratification biomarkers that may explain some of the heterogeneity of ASD by identifying internally homogenous subgroups on the spectrum [14]. Such stratification biomarkers may identify prognostic subgroups with different tracks of longitudinal symptom development or treatment subgroups of individuals with clinically relevant difficulties, e.g., in social skills, for selective enrollment in interventions [15]. Globally, several large-scale efforts aimed at ASD biomarker discovery have emerged, including European Autism Interventions—A Multicentre Study for Developing New Medications (EU-AIMS) Longitudinal European Autism Project (LEAP) [16], Janssen Autism Knowledge Engine (JAKE) [17], and the Autism Biomarkers Consortium for Clinical Trials (ABC-CT) [18]. In the United States, ABC-CT has resulted in the first two biomarkers, an EEG N170 measure and an ET measure, for ASD to be accepted for evaluation in the United States Food & Drug Administration (FDA) biomarker qualification program [19].

The highly heterogeneous responses seen in prior studies of facial expressions in ASD suggest that facial expressions may be a meaningful target for stratification. However, there are open challenges. While facial expression perception of individuals diagnosed with ASD has been well-studied via ET [20-22], few studies [23-33] have attempted to quantify facial expression production captured via VT. Among these, most existing studies [23-25, 27-30, 32, 34] analyze facial expression behaviors using off-the-shelf commercial or research tools based on machine learning models for FEA [24, 25, 27, 35-39] that have been developed using data sets of adult facial expressions. However, due to differences in facial growth and motor ability [40, 41], adult facial expressions are not an appropriate ground truth for analysis of child facial expressions, and FEA models trained with data collected from adults have been shown to perform poorly on images collected from children [42-45]. Facial expressions may provide valuable information about a child's development into an adult [46-48], making them a valuable modality for ASD research [13]. Many other applications of FEA including education (e.g., engagement in the classroom [49-51]), healthcare (e.g., monitoring of pain [52, 53], mental health [54, 55]), and entertainment (e.g., video games [56, 57]) also remain relevant from childhood into adulthood. To better support such applications and to better assess communication skills and support life-long care in ASD, there is a need for FEA models that generalize across distinctive expression patterns from early childhood to adulthood. However, few works [58, 59] address FEA across age groups. Furthermore, there has been little interaction between FEA research considering age variations and relevant fields such as facial age estimation and age-invariant face recognition (AIFR), where age variations have been well-studied.

Facial expression recognition and mimicry tasks have been used to elicit behavioral responses associated with facial expression perception and production [13]. Engagement plays a pivotal role in ensuring the validity and efficacy of such tasks in eliciting the intended construct [60]. Over the past twenty years and now entering the age of the metaverse, 3D avatars have become increasingly prominent and effective tools for engaging users in health applications. In clinical settings, 3D avatars have been broadly applied for neuro- and motor-rehabilitation [61] of patients, such as those who have suffered from stroke [62], cerebral palsy [63], brain injury [64], Parkinson's disease [65], Alzheimer's disease and dementia [66, 67]. Furthermore, 3D social avatars have aided the discovery of potential behavioral biomarkers and the development of therapeutic interventions for individuals with social anxiety disorder [68], depression [69, 70], schizophrenia [71], and ASD [32, 72-74]. For ASD research, animated 3D avatars have been shown to evoke higher levels of social engagement among individuals diagnosed with ASD when compared to traditional face-to-face interactions [75]. Given the ubiquity of facial expressions in daily life and their relevance to psychosocial health (e.g., facial palsy [76], depression [69, 70], social anxiety [68], and ASD [32, 72, 73]), facial expressions have become important targets for 3D avatar-based health applications. Thus, important design considerations have been identified to ensure that 3D avatars are valid and engaging for use in such applications [60, 77]. Among these, avatar customization has been shown to improve engagement among both autistic and NT individuals [78]. However, current customizable avatar platforms have not been evaluated by experts in the Facial Action Coding System (FACS) [79] or labeled with FACS action units (AUs), which is important to ensure that avatars accurately depict the target facial expressions. FACS [79] is gold-standard for expression labeling and its taxonomy of AUs describe the individual constituent muscle movements of the face. Using appropriate FEA models, FACS AUs may also be used to quantify the facial muscle activations of research participants in response to the stimuli. This one-to-one correspondence between AUs produced by a 3D avatar and AUs produced by a participant can be used to define the expected NT response for facial expression production based on measurements of individual AUs. Furthermore, this correspondence provides an interesting pathway to study whether participant facial responses follow the elicited configuration of AUs, which may be applicable not only to studies of ASD, but also to healthcare applications for individuals with Alzheimer’s disease, facial palsy, and more [62-64, 66, 67]. Therefore, there is a need for customizable avatar-based facial expression stimuli with FACS AU labels for improved user engagement and corresponding FEA models for measurement of FACS AUs.

There is a critical need for reliable biomarkers to support clinical and behavioral research in ASD [18]. In the United States, ABC-CT has established a thorough and effective framework for evaluation of potential biomarkers for ASD that may serve as a model for successive research studies [18]. With the goal of biomarker qualification, ABC-CT recruits hundreds of participants for multi-day studies including diagnostic confirmation, data acquisition, and deep phenotyping [18, 80]. Numerous candidate measurements, or dependent variables (DVs), are evaluated based on assay validity, data acquisition rates, distributional properties in the NT group, test-retest reliability, and replication in an independent sample [80]. In order to make it to such large scale study, these DVs need to have been discovered in previously published, smaller scale studies and be selected as candidate biomarkers based on ABC-CT's two inclusion criteria: construct validity and group discriminability [80]. Construct validity ensures that the experimental task elicits the intended response in the NT control group, and group discriminability confirms the presence of statistically significant differences between autistic and NT participant groups [80]. According to Shic et al. [81], group discriminability in the context of stratification is not expected to have effect sizes with diagnostic precision but rather, indicate broad group-level (ASD or NT) distributional differences associated with more homogenous subgroups within the ASD group. Few studies have considered facial expression production and perception for stratification biomarker discovery [82-85], and to our knowledge, none have assessed candidate biomarkers based on both construct validity and group discriminability.

Overcoming these open challenges, we hypothesize that DVs related to facial expression perception and production may hold promise for candidate ASD stratification biomarker discovery, as assessed by the ABC-CT criteria. To summarize the current challenges associated with the discovery of candidate ASD stratification biomarkers based on facial expressions, there are needs for 1) models that are appropriate for FEA of both child and adult facial expressions, 2) engaging, customizable avatar-based facial expression stimuli with FACS AU labels and corresponding FEA models for measurement of FACS AUs, and 3) assessment of construct validity and group discriminability for DVs related to facial expression production and perception in NT and ASD groups.

1.1 PROPOSED WORK AND CONTRIBUTIONS

This dissertation addresses the aforementioned challenges in three goals.

The first goal of this dissertation is to obtain a model that is appropriate for both child and adult FEA. For this purpose, a novel deep domain adaptative approach fusing facial landmark features is proposed for concurrent learning of adult and child facial expressions. Source (i.e., adult facial expressions) and target (i.e., child facial expressions) domains are aligned in a unified domain-invariant latent representation. Inspired by facial age estimation and AIFR, facial landmark measurements are fused with deep feature representations for robust expression learning across age groups. These facial landmark features are decomposed based on correlations with expression, domain, and identity factors using a novel facial feature selection method based on a mixture of beta distributions.

The second goal of this dissertation is to design FACS-labeled avatar-based facial expression stimuli for improved user engagement and develop associated models for automatic FACS AU measurement. Working with a certified FACS expert, six avatar models representing different genders and races with customizable hair color, eye color, skin tone, and clothing have been developed, each with FACS-labeled dynamic animations for six facial expressions (‘anger’, ‘disgust’, ‘fear’, ‘happy’, ‘sad’, and ‘surprise’). Corresponding deep models for multi-label AU detection are also developed. For richer representation learning, geometric landmark and deep learning-based texture features are fused while jointly learning AU detection and expression classification tasks. A novel beta-guided correlation loss encourages features to be correlated with AUs while discouraging correlation with subject identity. To study behavioral responses of healthy adult participants in response to the proposed stimuli as measured by proposed AU detection models, an online feasibility study is conducted. Participants complete expression recognition and mimicry tasks with the avatars while their facial webcam video and webcam-based eye-tracking are collected. We define and assess the validity of constructs based on the one-to-one relationship between avatar AUs and participant AUs, as well as a widely known ET construct (face preference [81]).

The third goal of this dissertation is to discover candidate stratification biomarkers based on the analysis of ET and VT data collected during an online pilot study of facial expression production and perception by autistic and NT children and young adults. The study is conducted online and 32 participants (11 with ASD and 21 NT) from across the United States take part. Participants diagnosed with ASD and NT participants are matched on age and gender, yielding 11 matched pairs (11 participants diagnosed with ASD and 11 matched NT participants) in the final cohort. Measurements of participants’ facial VT and webcam-based ET are collected during recognition and mimicry tasks. Construct validity in the NT group is evaluated for each DV (e.g., AUs, percentage of gaze duration to the face) in response to each stimulus (e.g., mimicry of ‘anger’, recognition of ‘surprise’). Then, candidate stratification biomarkers are identified among the DVs with valid constructs based on group discriminability (ASD vs. NT) using the Boruta [86] statistical approach.

In summary, this dissertation proposes computational models for automated FEA of adult and child facial expressions, FACS-labeled 3D avatar-based facial expression stimuli and AU measurements, and human subjects research study for discovery of stratification biomarkers for ASD.

Table 1 summarizes the proposed contributions for each goal. This dissertation has produced three conference proceedings and three journal manuscripts. The proposed contributions in goal 1 have been published in IEEE Transactions on Affective Computing [87] and comparison methods in Proceedings of SPIE Optics & Photonics Applications of Machine Learning 2022 [88], 2020 [44], and 2019 [43]. The proposed contributions in goal 2 are under review by IEEE Transactions on Affective Computing [89]. A journal manuscript reporting the research findings in goal 3 is under review by the Journal of Autism and Developmental Disorders [90].

Table 1. Summary of proposed contributions for this dissertation

Research Goal	**Description**	**Contributions**
1	Deep representation learning of adult and child facial expressions using domain adaptation fusing facial landmark features	Novel representation learning of adult and child facial expressions based on domain adaptation, statistical selection of facial landmark features, and feature fusion
2	Customizable avatars with dynamic facial action coded expressions for improved user engagement	FACS-labeled customizable avatar-based facial expression stimuli, deep learning-based AU measurements, novel beta-guided correlation loss, and construct validity based on feasibility study with healthy adults
3	Pilot study to discover candidate biomarkers for ASD based on perception and production of facial expressions	Online pilot study of individuals diagnosed with ASD and matched NT peers, construct validity and group discriminability for candidate biomarker selection

1.2 ORGANIZATION OF THE DISSERTATION

The remainder of the dissertation is organized as follows. Chapter 2 describes required background, including machine learning paradigms, tree and neural network-based methods, transfer learning and domain adaptation, and analysis of facial expressions and gaze. Chapter 3 proposes a deep domain adaptation and feature fusion model for concurrent learning of adult and child facial expressions with novel facial feature selection based on a mixture of beta distributions. Chapter 4 presents the proposed FACS-labeled customizable 3D avatar-based facial expression stimuli for improved engagement and associated deep learning-based AU measurements, including construct validity based on an online feasibility study with healthy adult participants. Chapter 5 discusses the protocol and findings of the proposed online pilot study involving participants diagnosed with ASD and matched NT peers, including construct validity, group discriminability, and sample size recommendations for future studies. In Chapter 6, the dissertation concludes with a summary and suggestions for future research directions.

CHAPTER 2

BACKGROUND REVIEW

This chapter briefly discusses the techniques and concepts required to understand this dissertation. In Section 2.1, the three main machine learning paradigms are introduced with an emphasis on the formulation of supervised learning problems for Chapters 3-5 and basic concepts from reinforcement learning relevant to Chapter 4. Section 2.2 discusses relevant machine learning methods, including tree-based methods primarily used in Chapter 5 and neural networks used in Chapters 3 and 4. Section 2.3 provides a succinct background of deep transfer learning and domain adaptation, which are foundational to Chapter 3. Section 2.4 briefly describes concepts related to FEA and ET analysis relevant to Chapters 3-5.

2.1 LEARNING PARADIGMS

Machine learning may be partitioned into three main paradigms: unsupervised learning, supervised learning, and reinforcement learning [91]. Unsupervised learning, supervised learning, and reinforcement learning are described in Sections 2.1.1, 2.1.2, and 2.1.3, respectively.

2.1.1 UNSUPERVISED LEARNING

Unsupervised learning seeks to learn patterns from input data $X = \{x_1, \dots, x_N\} \in \mathcal{X}$ without any human-provided labels [91]. For example, if each sample x_i, $i = 1, \dots, N$, represents a vector of p real-valued features, then $\mathcal{X} = \mathbb{R}^p$. If each sample x_i represents an $m \times n$ image, then $\mathcal{X} \subset \mathbb{R}^{m \times n}$. A detailed discussion of unsupervised learning is provided in Chapter 14 of Hastie et al. [92]. Unsupervised learning is concerned with learning the properties of probability density $\mathcal{P}(X)$ of input data X, especially when the dimensionality of p is large [92]. Tasks in unsupervised learning may include dimensionality reduction, i.e., finding lower dimensional manifolds representative of high data density to understand associations among features through methods such as principal component analysis, self-organizing maps, and others [93], and clustering [94], i.e., identification of clusters of data representing the modes of $\mathcal{P}(X)$ that may be associated with different types or classifications of the data [92]. One of the major limitations of unsupervised learning is that evaluating the performance of the methods is often challenging due to a lack of ground truth labels [92]. Instead, evaluation is based on heuristics or, for clustering, model-based evaluation metrics such as the silhouette coefficient [95], Calinski-Harabasz index [96], or Davies-Bouldin index [97] may be used. Such model-based metrics assess how 'well defined' the clusters are based on their distance, dispersion, and/or separation [95-97].

2.1.2 SUPERVISED LEARNING

In supervised learning, the input data $X = \{x_1, \dots, x_N\} \in \mathcal{X}$ have been annotated with associated output labels $Y = \{y_1, \dots, y_N\} \in \mathcal{Y}$, and the goal is to learn a mapping $f: \mathcal{X} \to \mathcal{Y}$ [91]. Considering joint probability density $\mathcal{P}(X, Y) = \mathcal{P}(Y|X) \cdot \mathcal{P}(X)$, supervised learning is concerned with characterizing the conditional density $\mathcal{P}(Y|X)$ [92]. Supervised learning methods are discussed in detail in Chapters 3-7 of Bishop [91] and Chapters 2-13 of Hastie et al. [92]. Supervised learning tasks may be categorized as regression or classification problems [91, 92]. For regression, each sample $x_i, i = 1, \dots, N$ is associated with one or more continuous, quantitative labels y_i, (e.g., $y_i \in \mathcal{Y} \subset \mathbb{R}^1$) [91, 92]. For classification, each x_i is associated with K discrete, qualitative labels called 'classes' (e.g., if $K = 2$, $y_i \in \mathcal{Y} = \{'class1', 'class2'\}$) [91, 92]. For classification, class labels are usually encoded as numerals. For example, a $K = 2$ class problem with classes ‘present’ or ‘absent’ may be represented with ‘1’ and ‘0’, respectively [92]. When there are more than two classes, each class label may be encoded as a ‘one-hot’ vector, i.e., a vector of K elements where the kth class ($k = 1, \ldots, K$) is represented as a ‘1’ in the kth position of the vector, and the remaining positions are ‘0’ (see page 407 of [98]).

Evaluation metrics for supervised learning methods quantify performance through the comparison of model predictions $f(x_i)$ to the ground truth labels y_i. For regression, it is standard practice to report the mean squared error (MSE) or root MSE (RMSE) and the mean absolute error (MAE) [99]. The MSE is defined as [99]:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i))^2. \tag{1}$$

MSE is the averaged form of the Euclidean distance or L2 norm and is optimal for Gaussian errors [99]. RMSE is calculated as the square root of the MSE [99]. RMSE provides a metric with the same units as the y_i’s and for normally distributed errors, represents the standard error [99]. The MAE is the averaged form of the Manhattan distance or L1 norm and is optimal for Laplacian errors [99]. The MAE is defined as [99]:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - f(x_i)|. \tag{2}$$

Lower MSE, RMSE, and MAE are associated with better performance.

For classification, evaluation metrics are defined based on the number of true positives (TP; e.g., ‘class 1’ samples correctly predicted as ‘class 1’), false positives (FP; e.g., samples not in ‘class 1’ incorrectly predicted as 'class 1’), true negatives (TN; e.g., samples not in 'class 1’ correctly predicted as not ‘class 1’), and false negatives (FN; e.g., ‘class 1’ samples incorrectly predicted as not 'class 1') [100]. The 'recall' or 'sensitivity' is the fraction of correctly predicted samples of a particular class divided by the total number of samples of that class [100]:

$$recall = sensitivity = \frac{TP}{TP + FN}. \tag{3}$$

The 'precision' is the fraction of correctly predicted samples of a particular class divided by the total number of predictions for that class [100]:

$$precision = \frac{TP}{TP + FP}. \tag{4}$$

The 'specificity' is the fraction of samples correctly predicted as not of a particular class divided by the total number of samples not of that class [100]:

$$specificity = \frac{TN}{TN + FP}. \tag{5}$$

The 'F1 score' summarizes precision and recall as their harmonic mean [100]:

$$F1\ score = \frac{2 \cdot precision \cdot recall}{precision + recall}. \tag{6}$$

A 'receiver operating characteristic (ROC) curve' is a plot of $sensitivity$ vs. $1 - specificity$ [101]. The 'area under the ROC curve' (AUC) may be used to summarize the sensitivity and specificity metrics for performance evaluation [101]. Higher recall/sensitivity, precision, specificity, F1 score, and AUC are associated with better performance [100, 101].

Examples of supervised learning approaches include decision trees, random forests, and neural networks, which will be discussed in sections 2.2.1, 2.2.2, and 2.2.3, respectively, as well as many other methods [91, 92, 98].

2.1.3 REINFORCEMENT LEARNING

As in supervised learning, reinforcement learning is concerned with the conditional distribution of a response variable given a set of features [102]. However, unlike supervised learning where the optimization targets for the specified task are clear from the labels, reinforcement learning performs its optimization through the reinforcement, or feedback, it receives from the environment at each turn [102]. Although there may be no clear optimal outcome, it is still possible to learn to perform the task well by finding a series of actions that yield better outcomes than others [102]. For example, a gambler may explore and exploit a variety of slot machines with different probabilities of returning a jackpot with the goal of maximizing earnings (without knowing the maximum possible earnings) [102]. More formally, in reinforcement learning, we consider a decision-making 'agent' in an 'environment' that may inhabit one of many possible 'states' [91, 103]. Learning focuses on determining a series of suitable 'actions', i.e., interactions with the environment to transition it from one state to another, for the agent to take in order to maximize the total reward at a particular time step, i.e., the positive or negative cost of a particular action taken in a particular state [91, 103]. The total reward is the summation of the immediate reward and the anticipated future rewards scaled by an exponentiated 'discount factor' [103]. The best action for an agent to choose in each state is determined by a 'policy' that may either be extracted through learning the value of states and actions, as in value-based methods or learned directly through policy-based methods such as policy gradients [103]. A detailed coverage of reinforcement learning is provided in Miller [103].

2.2 MACHINE LEARNING METHODS

This section describes multiple machine learning methods used in this dissertation, including decision trees, random forests, and artificial neural networks. We focus our discussion on these methods as they are applied to supervised classification problems.

For all methods, we consider that the data has been split into independent ‘train’ and ‘test’ sets. The train set is used to fit the model while the test set is reserved for evaluation purposes [104]. The train set may be split again to produce an independent ‘validation’ set, which may be used to assess model performance for, e.g., selecting the best model among multiple models, stopping training early, or choosing ‘hyperparameters’ [104]. Hyperparameters are model parameters that need to be specified prior to training, rather than learned while fitting the model [104]. One method of splitting a data set into train/test or train/validation sets is cross-validation [98]. In k-fold cross-validation, the data set is split into k-independent folds [98]. The first fold is taken as the test (or validation) set and the remaining $k - 1$ folds form the train set [98]. The model is then fit on the train set and evaluated on the held-out test fold [98]. Next, the second fold is taken as the test set and the remaining folds (including the first fold) are used to train the model [98]. This procedure is repeated so that each fold serves as the test set exactly once [98]. A special case of k-fold cross-validation is leave-one-out cross-validation (LOOCV). In LOOCV, k equals the total number of samples in the full data set [98]. Each test set consists of only one sample while the remaining samples are used for training [98].

In the following sections, we denote the input space as $\mathcal{X}$ and output space as $\mathcal{Y}$. The goal of each machine learning method is to learn a function $f: \mathcal{X} \to \mathcal{Y}$. The training data consist of samples $X = \{x_1, \dots, x_N\} \in \mathcal{X}$ with labels $Y = \{y_1, \dots, y_N\} \in \mathcal{Y}$, where N is the total number of training samples. Let individual sample-label pairs be denoted as $(x_i, y_i), i = 1, \dots, N$.

Sections 2.2.1, 2.2.2, and 2.2.3 describe decision trees, random forests, and neural networks, respectively.

2.2.1 DECISION TREE

A decision tree is a supervised learning method that learns a type of directed graph that can be used to perform classification or regression tasks [91, 100]. Chapter 18 of Shalev-Shwartz and Ben-David [100] and Chapter 8 of James et al. [98] provide a detailed discussion of decision trees. This graph consists of an input node known as the 'root', one or more internal nodes known as 'branches', and output nodes known as 'leaves' that each corresponds to a specific label [100]. Figure 2 shows an example of this type of structure. Prediction involves transversing the graph from root to leaf based on partitioning of the input space $\mathcal{X}$ [100]. Rules for such partitioning may be defined based on thresholds τ_j on the value of individual features X_j, where $j = 1, \dots, p$ for p features [100]. These rules are learned during training. There are various algorithms for growing decision trees, including Iterative Dichotomiser 3 (ID3) [105] and Classification and Regression Trees (CART) [106].

We will consider the CART algorithm, which performs binary partitioning of the input space. Let the data at node a be denoted as D_a. For each candidate partitioning rule $\omega = \left(X_j, \tau_a\right)$ based on a feature X_j and threshold τ_a, D_a may be split into two subsets [98, 107]:

$$\begin{aligned} D_a^{left}(\omega\ ) &= \left\{(X, Y)\middle| X_j \le \tau_a\right\}, \\ D_a^{right}(\omega) &= \left\{(X, Y)\middle| X_j > \tau_{a},\right\}. \end{aligned} \tag{7}$$

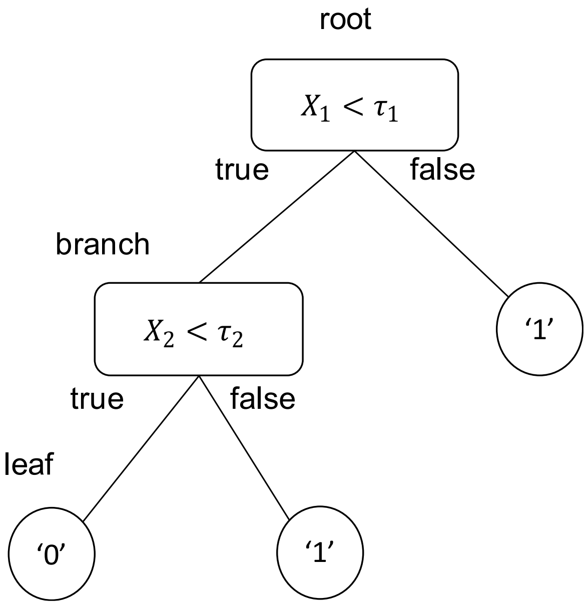


Figure 2. Sample decision tree structure with root, branch, and leaf nodes (representing class '0' or class '1') labeled.

If a subset does not consist of only a single class label, it is called 'impure' [98, 100]. Let $H(\cdot)$ be a function that quantifies the impurity of a subset [98, 107]. A popular choice for $H(\cdot)$ is the Gini index [98]:

$$Gini = \sum_{k=1}^{K} r_{ak}(1 - r_{ak}), \tag{8}$$

where r_{ak} denotes the proportion of samples at a particular node a with the kth class label. Then, the quality of the partitioning rule may be computed as the weighted average of the impurities of the two resulting subsets:

$$G(D_a, \omega) = \frac{n_a^{left}}{n_a}\left(D_a^{left}(\omega\,)\right) + \frac{n_a^{right}}{n_a}\left(D_a^{right}(\omega\,)\right), \tag{9}$$

where n_a, n_a^{left}, and n_a^{right} are the number of samples in D_a, $D_a^{left}(\omega)$, and $D_a^{right}(\omega)$, respectively [98, 107]. Then, we find the partitioning rule that minimizes $G(\cdot)$ [98, 107]:

$$\omega^* = \text{argmin}_\omega\, G(D_a, \omega). \tag{10}$$

This procedure is then recursed for $D_a^{left}(\omega^*)$ and $D_a^{right}(\omega^*)$ [98, 107]. Since it is possible to continue partitioning until all training samples are correctly classified, limitations may be placed on tree depth, i.e., the number of recursions, to reduce 'overfitting' [100]. Overfitting occurs when a machine learning method fits to noise in the training data, resulting in poor generalization to the test set [104]. The maximum depth may be specified as a hyperparameter prior to training. Therefore, recursion continues either until all leaves are 'pure' or the maximum depth has been reached [98, 107]. At test time, the predicted label is the label of the majority of training samples at the leaf node [98].

An extension of CART, called a conditional inference tree (CIT) [108], uses the statistical significance of permutation tests to determine partitioning rules, instead of impurity measures like the Gini index.

2.2.2 RANDOM FOREST

One of the limitations of CART is high variance in the constructed trees, e.g., splitting the train set in half and fitting trees on each part may yield vastly different trees [98]. Bootstrap aggregation, or bagging, may be used to reduce the variance of decision trees [98]. In bagging, *b* bootstrapped training sets are obtained by randomly sampling from the train set [98]. Then, *b* decision trees are fit, one for each of the b bootstrapped train sets [98]. The number of trees b is a hyperparameter set prior to training. To predict the class label using the 'ensemble' of b trees, each tree votes on the class label to predict, and the majority vote is taken as the predicted class label [98]. A random forest is an ensemble of decision trees formed by bagging with the additional requirement that the tth tree in the ensemble only have access to a random subset of the full set of p features, where there are p_t features in each random subset and p_t is typically chosen to be $\sqrt{p}$ [98]. The random sampling of the features has the effect of decorrelating the trees in the ensemble, further reducing the variance and improving the reliability of the approach [98].

2.2.3 NEURAL NETWORKS

An artificial neural network is a computational structure based on a directed graph and biologically inspired by the brain [100]. Neural networks are composed of nodes, called 'neurons' or 'units', that are arranged in disjoint subsets called 'layers' [100]. The edges connecting the nodes in the graph are each associated with a weight that is learned during training [100]. References for neural networks include Chapter 5 of Bishop [91], Chapter 20 of Shalev-Shwartz and Ben-David [100], Chapter 10 of James et al. [98], and Chapters 6-9 of Goodfellow et al. [109].

We concentrate our discussion on feedforward neural networks used in this dissertation. Feedforward neural networks are distinguished by the absence of cycles in the graphical model underlying the neural network [100]. Each edge connects the output of a neuron in the $(l-1)$th layer to the input of a neuron in the lth layer [100]. There are other types of neural networks with cyclic connections, e.g., recurrent neural networks [98].

Every neural network has an input layer, one or more hidden layers, and an output layer [98, 100]. Perhaps the simplest feedforward neural network is a 'multilayer perceptron (MLP)', i.e., where nodes in adjacent layers are fully connected or ‘dense’ [91]. An MLP with one hidden layer is a ‘shallow’ neural network architecture, while neural networks with two or more hidden layers are called ‘deep’ neural networks. ‘Deep learning’ is the field that studies deep neural networks [98]. A type of neural network architecture that is not fully connected is the convolutional neural network (CNN) [98]. CNNs use ‘convolutional layers’ where weights are organized into relatively smaller ‘kernels’, or ‘filters’, that are convolved with the input [98]. CNNs work particularly well on images [98]. The following subsections will discuss fully connected neural networks and CNNs in greater detail.

2.2.3.1 FULLY CONNECTED NEURAL NETWORKS

Consider a fully connected neural network with $L + 1$ layers indexed by $l = 0, \ldots, L$. Figure 3 shows an example of a fully connected deep neural network with four layers. The number of units in the input layer ($l = 0$) is equal to the dimensionality p of the data $X \in \mathbb{R}^p$ plus one neuron called the ‘bias’. Except for the bias node which always outputs 1, the output of each unit in the input layer is just its input, i.e., the jth node in the input layer outputs feature value X_j [100]. Each of the p input units of the input layer is connected to each of the M_1 hidden units in the first hidden layer ($l = 1$) [98]. In general, each hidden unit accepts a ‘fan-in’ of inputs which are the outputs of the nodes in the previous layer times their associated edge weights $w^{(l)} \in W$ [98]. This fan-in of inputs to the neuron is summed and undergoes a nonlinear transformation $g(\cdot)$, called the ‘activation function’, which produces the neuron’s output, called an ‘activation’ [98].

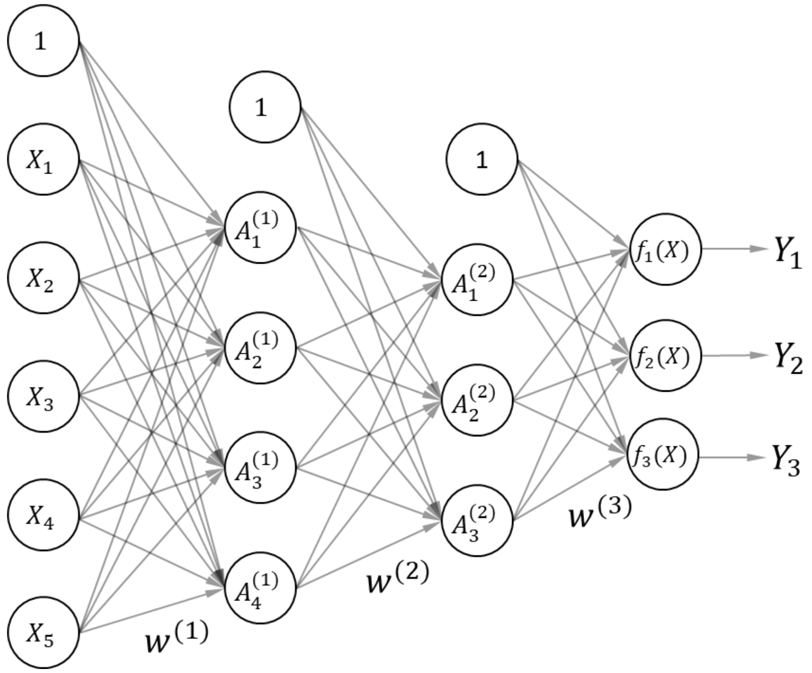


Figure 3. Example of a fully connected deep neural network with an input layer with five non-bias units, two hidden layers with four and three non-bias hidden units, respectively, and an output layer with three output units.

While there are many different activation functions available (see [110] for a survey), the most popular choice of $g(\cdot)$ is the rectified linear unit (ReLU) [98]:

$$g(\xi) = \begin{cases} 0 & if\ \xi < 0 \\ \xi & otherwise \end{cases}, \tag{11}$$

where ξ is a placeholder for the argument of $g(\cdot)$. The first hidden layer ($l = 1$) activations $A_{m_1}^{(1)}$, where $m_1 = 1, \dots, M_1$ hidden units, may be computed as [98]:

$$A_{m_1}^{(1)} = g(w_{m_1 0}^{(1)} + \sum_{j=1}^{p} w_{m_1 j}^{(1)} X_j), \tag{12}$$

where $w_{m_1 0}^{(1)}$ is the weight associated with the bias term, $w_{m_1 j}^{(1)}$ describes the weight between the jth input feature ($j = 1, \dots, p$) and $m_1 th$ hidden unit, X_j is the jth input feature, and $g(\cdot)$ is the activation function. While the first hidden layer is connected directly to the p features in the input layer, subsequent hidden layers ($1 < l < L$) connect to the activations of the previous hidden layer. The activations $A_{m_l}^{(l)}$ of the lth hidden layer ($1 < l < L$) with $m_l = 1, \dots, M_l$ hidden units may be written as [98]:

$$A_{m_l}^{(l)} = g(w_{m_l 0}^{(l)} + \sum_{m_{l-1}=1}^{M_{l-1}} w_{m_l m_{l-1}}^{(l)} A_{m_{l-1}}^{(l-1)}), \tag{13}$$

where $w_{m_l 0}^{(l)}$ is the weight associated with the bias term in the lth layer, $w_{m_l m_{l-1}}^{(l)}$describes the weight between the $m_{l-1}th$ hidden unit ($m_{l-1} = 1, \dots, M_{l-1}$) in the previous (($l - 1)th$) hidden layer and $m_l th$ hidden unit in the current (lth) hidden layer, $A_{m_{l-1}}^{(l-1)}$ is the $m_{l-1}th$ activation of the previous (($l - 1)th$) hidden layer, and $g(\cdot)$ is the activation function. The number of units in the output layer ($l = L$) corresponds to the number of classes K in the classification task. For binary classification ($K = 2$), only one output unit is needed as the labels Y may be encoded as '0' or '1' [98]. For multi-class classification of $K > 2$ different classes, there are K output units to correspond to the one-hot encoded vector of labels Y [98]. An output function $o(\cdot)$ is applied at each output node for final transformation to label space $\mathcal{Y}$ [92]. The selection of $o(\cdot)$ depends on whether the task is binary ($K = 2$) or multiclass classification ($K > 2$). For binary classification ($K = 2$), the output of the single output unit may be written as [98]:

$$f(X) = o(w_{10}^{(L)} + \sum_{m_{L-1}=1}^{M_{L-1}} w_{1m_{L-1}}^{(L)} A_{m_{L-1}}^{(L-1)}), \tag{14}$$

where $w_{10}^{(L)}$ is the weight associated with the bias term in the output $((l = L)th)$ layer, $w_{1m_{L-1}}^{(L)}$describes the weight between the $m_{L-1}th$ hidden unit $(m_{L-1} = 1, \dots, M_{L-1})$ in the last hidden $((L-1)th)$ layer and the single output unit in the output $((l = L)th)$ layer, $A_{m_{L-1}}^{(L-1)}$ is the $m_{L-1}th$ activation of the last hidden $((L-1)th)$ layer, and $o(\cdot)$ is the output function. For binary classification, the typical selection of output function $o(\cdot)$ is the sigmoid function [98]:

$$sigmoid(\xi) = \frac{e^{\xi}}{1 + e^{\xi}} = \frac{1}{1 + e^{-\xi}}, \tag{15}$$

where ξ is a placeholder for the argument of $o(\cdot)$. For multiclass classification $(K > 2)$, the output of the kth output unit (corresponding to the kth class with $k = 1, \dots, K$) in the output layer $(l = L)$ may be computed as [98]:

$$f_k(X) = o(w_{k0}^{(L)} + \sum_{m_{L-1}=1}^{M_{L-1}} w_{km_{L-1}}^{(L)} A_{m_{L-1}}^{(L-1)}), \tag{16}$$

where $w_{k0}^{(L)}$ is the weight associated with the bias term in the output $((l = L)th)$ layer, $w_{km_{L-1}}^{(L)}$describes the weight between the $m_{L-1}th$ hidden unit $(m_{L-1} = 1, \dots, M_{L-1})$ in the last hidden $((L-1)th)$ layer and the kth output unit in the output $((l = L)th)$ layer, $A_{m_{L-1}}^{(L-1)}$ is the $m_{L-1}th$ activation of the last hidden $((L-1)th)$ layer, and $o(\cdot)$ is the output function. For multiclass classification, the usual choice of $o(\cdot)$ is the softmax function [92]:

$$softmax_k(\xi) = \frac{e^{\xi_k}}{\sum_{i=1}^{K} e^{\xi_i}}, \tag{17}$$

where ξ is a placeholder for the argument of $o(\cdot)$. Both sigmoid and softmax output functions produce positive estimates for each class that sum to one, i.e., the class ‘probabilities’ [92].

To train the network, we need to find weights W that minimize a 'loss' function $\mathcal{L}$ that quantifies how well the network classifies the training data at any given training step [98]. Many different loss functions have been proposed for classification tasks [111]. We present the most common, which are based on the cross-entropy between the predicted class probability and ground truth class label [111]. For binary classification ($K = 2$), the cross-entropy may be written as [98, 112]:

$$binary\_crossentropy(X, Y) \; = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(f(x_i)) + (1 - y_i)\log(1 - f(x_i))). \quad (18)$$

For multiclass classification ($K > 2$), categorical cross-entropy generalizes cross-entropy for multiple outputs [98, 112]:

$$categorical\_crossentropy(X, Y) \; = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik}\log\big(f_k(x_i)\big). \quad (19)$$

Given sample-label pairs from the train set, the gradient of the loss function with respect to the weights W is efficiently computed using the backpropagation algorithm (please see Chapter 11.4 of Hastie et al. [92] and Chapter 10.7 of James et al. [98] for detailed discussion of the backpropagation algorithm). Loss functions may also be applied to the validation and test sets to calculate the validation loss and test loss, respectively [98]. However, only the train loss is used to update the weights. The gradients are used by the minibatch stochastic gradient descent (SGD) optimization algorithm to perform updates to the weights W iteratively. Iterations correspond to learning from different portions of the train set, called 'minibatches' or simply 'batches' [98, 104]. Iterating over all the batches, i.e., the entire train set, is called an 'epoch' [104]. Training typically takes place for a fixed number of epochs or until an early stopping condition, e.g., based on the validation loss, is met [98, 104]. At the tth iteration, the weight updates may be computed as [91]:

$$W_{t+1} = W_t - \zeta_t \cdot \nabla_{W_t} \mathcal{L}(x_{t:t+b}, y_{t:t+b}; W_t), \tag{20}$$

where ∇_{W_t} indicates the gradient with respect to weights W_t, $\mathcal{L}(\cdot)$ is the loss function (e.g., binary or categorical cross-entropy), ζ_t is the 'learning rate' (a positive real number and selection of its value is problem-specific), and b is the 'batch size' (a positive integer). The learning rate determines the step size at each iteration, scaling the magnitude of the gradient-based updates [91]. The learning rate and batch size are hyperparameters that need to be selected prior to training [104]. Variants of minibatch SGD have been developed by taking previous weight updates into consideration in addition to the current gradients [104]. One of the most widely used variants is adaptive moment estimation (ADAM), which estimates the first and second moments of the gradients to compute adaptive learning rates for different parameters [113].

Random forests have inspired an efficient form of regularization called 'dropout' that may be used in neural network layers to reduce overfitting [98]. Each epoch, dropout randomly selects $\phi\%$ of units in the layer and sets their activations to 0 [98]. The remaining units participate in the training epoch with their weights scaled by $1/(1-\phi)$ [98]. Further discussion of dropout can be found in Chapter 10.7.3 of James et al. [98].

2.2.3.2 CONVOLUTIONAL NEURAL NETWORK

CNN is a feedforward neural network architecture that uses a special type of hidden layer called a convolutional layer [98]. For further reading on CNNs, please see Chapter 8.3.1 of Berk [102], Chapter 10.3 of James et al. [98], Chapter 5.5.6 of Bishop [91], Chapter 5 of Chollet [104], and Chapter 9 of Goodfellow et al. [109].The weights of a convolutional layer are organized into a kernel that is used to filter the layer input [98, 104]. This involves passing the kernel over the input, multiplying the weights by the corresponding overlapping values of the input, and summing up the products [98, 104]. Then, the activation function $g(\cdot)$, e.g., ReLU, is applied to the filtered input to yield the layer output, which is called a 'feature map' or 'activation map' [98, 104]. Stated in another way, each entry in the feature map is a dot product between the kernel and a region of the layer input, followed by a nonlinear transformation $g(\cdot)$ [98, 104]. Usually, a large bank of kernels is used, producing one feature map per kernel [98, 104]. An example is shown in Figure 4. In Figure 4 (a), the trainable kernel begins scanning the input from the top left. Element-wise multiplication of the kernel and overlapping input region is performed. The products are summed and passed through the activation function to yield the top left element of the feature map. In Figure 4 (b), the kernel is moved to the right. Multiplication, summation, and activation steps are repeated to yield the next entry of the feature map. Finally, in Figure 4 (c) the kernel continues scanning from left to right and top to bottom until the feature map is complete.

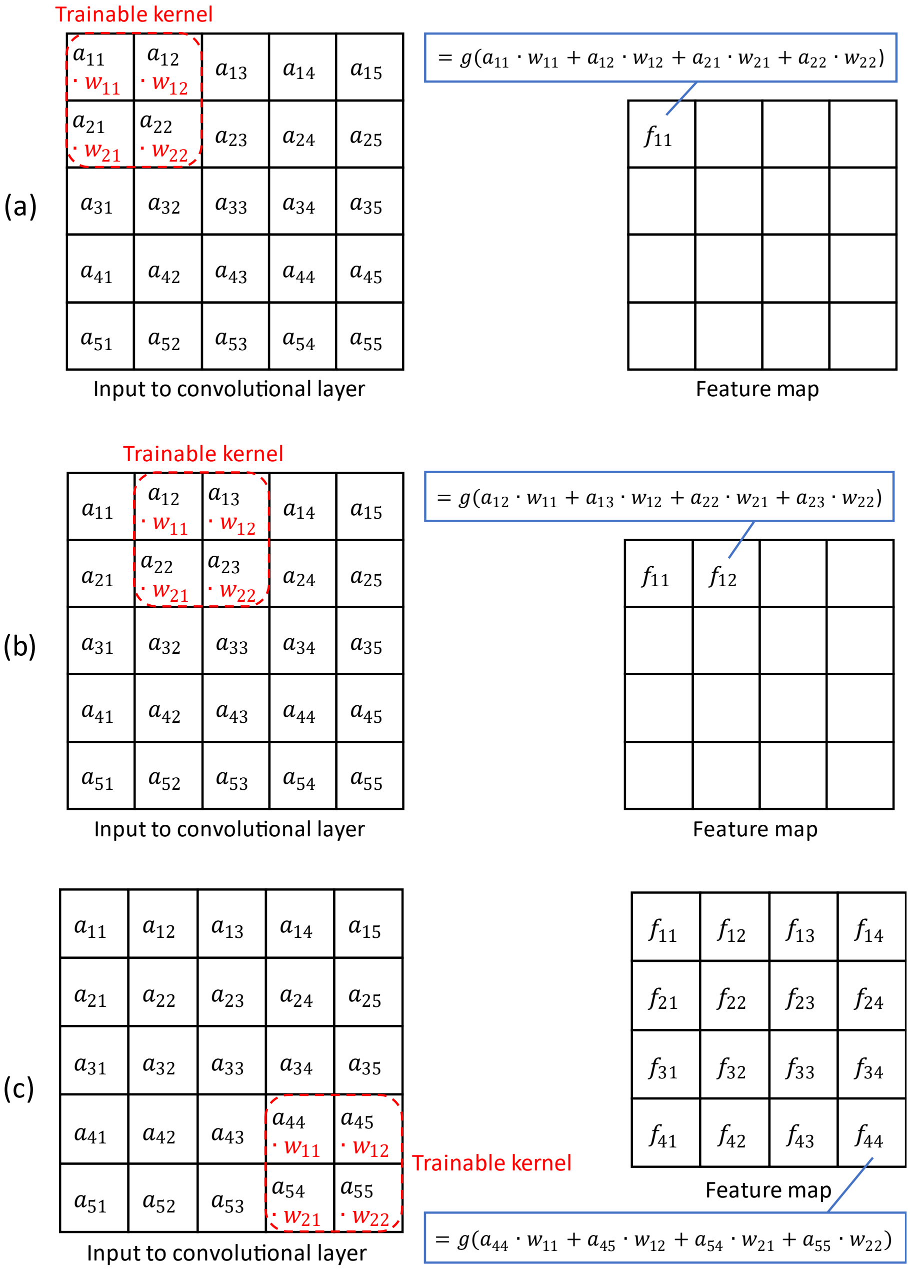

(a)
Trainable kernel
$a_{11} \cdot w_{11}$ $a_{12} \cdot w_{12}$ a_{13} a_{14} a_{15}
$a_{21} \cdot w_{21}$ $a_{22} \cdot w_{22}$ a_{23} a_{24} a_{25}
a_{31} a_{32} a_{33} a_{34} a_{35}
a_{41} a_{42} a_{43} a_{44} a_{45}
a_{51} a_{52} a_{53} a_{54} a_{55}
Input to convolutional layer
$= g(a_{11} \cdot w_{11} + a_{12} \cdot w_{12} + a_{21} \cdot w_{21} + a_{22} \cdot w_{22})$
f_{11}
Feature map
(b)
Trainable kernel
a_{11} $a_{12} \cdot w_{11}$ $a_{13} \cdot w_{12}$ a_{14} a_{15}
a_{21} $a_{22} \cdot w_{21}$ $a_{23} \cdot w_{22}$ a_{24} a_{25}
a_{31} a_{32} a_{33} a_{34} a_{35}
a_{41} a_{42} a_{43} a_{44} a_{45}
a_{51} a_{52} a_{53} a_{54} a_{55}
Input to convolutional layer
$= g(a_{12} \cdot w_{11} + a_{13} \cdot w_{12} + a_{22} \cdot w_{21} + a_{23} \cdot w_{22})$
f_{11} f_{12}
Feature map
(c)
a_{11} a_{12} a_{13} a_{14} a_{15}
a_{21} a_{22} a_{23} a_{24} a_{25}
a_{31} a_{32} a_{33} a_{34} a_{35}
a_{41} a_{42} a_{43} $a_{44} \cdot w_{11}$ $a_{45} \cdot w_{12}$
a_{51} a_{52} a_{53} $a_{54} \cdot w_{21}$ $a_{55} \cdot w_{22}$
Trainable kernel
Input to convolutional layer
f_{11} f_{12} f_{13} f_{14}
f_{21} f_{22} f_{23} f_{24}
f_{31} f_{32} f_{33} f_{34}
f_{41} f_{42} f_{43} f_{44}
Feature map
$= g(a_{44} \cdot w_{11} + a_{45} \cdot w_{12} + a_{54} \cdot w_{21} + a_{55} \cdot w_{22})$

Figure 4. Example of the computations involved in a convolutional layer with a 5x5 layer input and 2x2 kernel.

Pooling is used to perform downsampling of feature maps through aggregation of local feature information [98]. Maximum pooling summarizes local regions of the layer input, e.g., regions of size 2x2, with the maximum value [98, 104]. Similarly, average pooling uses the mean value to summarize each local region [104]. An example of maximum pooling is as follows [98]:

$$\text{Max pool}\begin{bmatrix} 0 & 1 & 2 & 3 \\ 1 & 3 & 8 & 0 \\ 2 & 9 & 5 & 3 \\ 3 & 1 & 2 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 3 & 8 \\ 9 & 5 \end{bmatrix}. \tag{21}$$

A typical CNN architecture begins with an input layer matching the dimensions of images X [98]. The input layer then feeds into a series of blocks consisting of convolutional layers each followed by a pooling operation [98]. Next, the feature maps from the last block are ‘flattened’ into separate units that are fed into fully connected layers [98]. Finally, these fully connected layers feed into the output layer [98]. CNNs may be trained to perform classification tasks in the same way as previously described in Section 2.2.3.1.

2.3 TRANSFER LEARNING AND DOMAIN ADAPTATION

Transfer learning is the use of information learned from a source task T_S on source domain D_S to solve a related target task T_T on target domain D_T [109]. D_S corresponds to an input space $\mathcal{X}_S$ and output space $\mathcal{Y}_S$ [114]. T_S involves learning $f_S\colon \mathcal{X}_S \rightarrow \mathcal{Y}_S$ using source data $X_S = \left\{x_i^S\right\}_{i=1}^{N_S} \in \mathcal{X}_S, x_i^S \sim \mathcal{P}_X^S(x)$ and source labels $Y_S = \left\{y_i^S\right\}_{i=1}^{N_S} \in \mathcal{Y}_S$, where N_S is the number of source samples and $\mathcal{P}^S$ is the source probability distribution. Similarly, T_T involves learning $f_T\colon \mathcal{X}_T \rightarrow \mathcal{Y}_T$ [114]. Target data are denoted as $X_T = \{x_i^T\}_{i=1}^{N_T} \in \mathcal{X}_T, x_i^T \sim \mathcal{P}_X^T(x)$ and target labels are denoted as $Y_T = \{y_i^T\}_{i=1}^{N_T} \in \mathcal{Y}_T$, where N_T is the number of target samples and $\mathcal{P}^T$ is the target probability distribution. A special case of transfer learning where T_S and T_T share label spaces, i.e., $\mathcal{Y}_S = \mathcal{Y}_T$, is called ‘domain adaptation’ [109, 114]. Transfer learning and domain adaptation rely on representation learning of common features that are relevant across domains and/or tasks, such as edges, geometric shapes, lighting, etc., in visual images [109]. For further reading on transfer learning and domain adaptation, please see Chapter 15.2 of Goodfellow et al. [109].

Given the substantial number of samples required for effective deep learning, a common application of transfer learning is the fine-tuning of neural networks that have been pretrained on a large number of samples (e.g., millions) for new tasks for which the number of available labeled training samples is much smaller (e.g., hundreds) [104]. In this context, we assume that $\mathcal{X}_S = \mathcal{X}_T$, i.e., the input spaces are the same, with $\mathcal{P}^S \neq \mathcal{P}^T$ [114]. $\mathcal{Y}_S$ and $\mathcal{Y}_T$ may be the same or different [114]. For example, consider a CNN model that has been trained to classify 1000 classes of objects from images (T_S) [104]. We would like to utilize information the model learned for performing object detection in order to perform a new task (T_T), e.g., binary classification of images of cats and dogs [104]. Since $\mathcal{Y}_S \neq \mathcal{Y}_T$, the minimum modification to the existing network architecture is the removal of the 1000-unit output layer and replacement with a single output unit for binary classification [104]. If desired, we can remove more layers from the existing network (called the ‘base’ network) or add additional layers on top of the base as long as the output layer has the appropriate number of units for task T_T [104]. Next, we ‘freeze’ the base network, i.e., do not allow for updates to the weights [104]. The purpose of freezing the base network is to prevent useful representations in the early layers from being altered by possibly noisy updates based on the target training samples [104]. Then, the newly added layers are trained with data and labels from the target domain [104]. After that, we unfreeze some of the layers at the end of the base layer and train these jointly with the newly added layers, usually using a reduced learning rate [104]. This allows the last few layers of the base layers to specialize their representations for the target task [104].

Domain adaptation methods may be categorized by the availability of target labels Y_T [114]. In supervised domain adaptation, we have access to the target labels Y_T [114]. In unsupervised domain adaptation, no target labels are available [114]. In semi-supervised domain adaptation, a limited number of target labels are available [114]. Domain adaptation requires that assumptions be made on how the joint distribution of X and Y changes [115]. Considering $\mathcal{P}(XY) = \mathcal{P}(X|Y)\mathcal{P}(Y) = \mathcal{P}(Y|X)\mathcal{P}(X)$, the covariate shift assumption attributes the change in the joint distribution to changes in $\mathcal{P}(X)$ only, motivating methods that adapt the feature space to $\mathcal{P}(Y|X)$ which is assumed constant [115]. Other settings may assume label shift, i.e., only $\mathcal{P}(Y)$ changes while $\mathcal{P}(X|Y)$ remains constant, or conditional shift, i.e., only $\mathcal{P}(X|Y)$ changes while $\mathcal{P}(Y)$ remains constant [115].

2.4 ANALYSIS OF FACIAL EXPRESSIONS AND GAZE

Sections 2.4.1 and 2.4.2 below describe concepts related to analysis of facial expressions and gaze, respectively.

2.4.1 ANALYSIS OF FACIAL EXPRESSIONS

A typical FEA pipeline begins with the input data, images of people making facial expressions. Next, face detection is often performed followed by cropping the image to the bounding box of the face. Popular methods for face detection include Haar Cascades [116] and CNNs [117]. Other preprocessing steps, such as normalization, may be applied depending on the data and task needs. Both traditional machine learning and deep learning approaches may be applied to FEA. For traditional machine learning, features must be extracted from the preprocessed input images. Such features may be geometric or texture-based. Geometric features may be computed based on facial landmark points [118-121]. ‘Facial landmarks’ are x and y-coordinates that mark the locations of facial components, such as the nose, eyes, eyebrows, mouth, and jawline [122]. Facial landmark points may be robustly and automatically detected using algorithms such as ensembles of regression trees [122]. Texture-based features may be extracted using traditional methods such as local binary patterns, histograms of oriented gradients, etc., or deep learning, e.g., CNNs [123-125]. Techniques such as filter-based methods based on univariate statistics, forward and backward stepwise selection, Least Absolute Shrinkage and Selection Operator (LASSO), etc., may be used to select a subset of features from the full set to be used for traditional machine learning [98]. Deep learning methods can be applied directly to input images of facial expressions without separate feature extraction and selection steps. Some tasks in FEA are facial expression classification and FACS AU detection.

Facial expression classification tasks focus on classifying images of facial expressions as belonging to one of two or more groups. Examples are positive vs. negative expressions (also called valence) [126-128] and the prototypical expressions (‘anger’, ‘disgust’, ‘fear’, ‘happy’, ‘sad’, ‘surprise’) [129, 130]. More granular analysis of facial expressions may be performed using AUs of FACS, which describe the individual constituent movements of the human face [130]. TABLE 2 provides examples of FACS AUs and associated facial muscles [130]. FACS AU detection involves labeling a facial expression image with multiple AUs based on the configuration of the face [127, 131, 132].

Table 2. FACS AUs and associated muscles

AU	Name	Associated Muscles
1	Inner Brow Raiser	frontalis, pars medialis
2	Outer Brow Raiser	frontalis, pars lateralis
4	Brow Lowerer	depressor glabellae, depressor supercilii, corrugator supercillii
5	Upper Lid Raiser	levator palpebrae superioris, superior tarsal muscle
6	Cheek Raiser	orbicularis oculi, pars orbitalis
7	Lid Tightener	orbicularis oculi, pars palpebralis
9	Nose Wrinkler	levator labii superioris alaeque nasi
10	Upper Lip Raiser	levator labii superioris, caput infraorbitalis
11	Nasolabial Deepener	zygomaticus minor
12	Lip Corner Puller	zygomaticus major
13	Cheek Puffer	levator anguli oris
15	Lip Corner Depressor	depressor anguli oris
17	Chin Raiser	mentalis
20	Lip Stretcher	risorius
23	Lip Tightener	orbicularis oris
24	Lip Pressor	orbicularis oris
25	Lips Part	depressor labii inferioris, or relaxation of mentalis, orbicularis oris
26	Jaw Drop	masseter, relaxed temporalis and internal pterygoid
27	Mouth Stretch	pterygoids, digastric

2.4.2 ANALYSIS OF GAZE

ET involves the tracking of eye movements and the estimation of where a person is placing visual attention with respect to the visual scene [133]. Eye gaze can be divided into ‘fixations’ and ‘saccades’ [133]. Fixations represent a pause in eye movement while the person focuses on a particular area of the visual scene [133]. A saccade is the rapid movement of the eye that connects two fixations [133]. In ET data, fixations are described by x and y-coordinates that indicate where the person placed attention on the visual scene [133]. The visual scene may be divided into different areas of interest (AOIs). For example, in ASD research, a visual scene showing a caregiver playing with a toy may have AOIs associated with the caregiver’s face, the caregiver’s body, and the toy [81]. Thus, fixations may be categorized into different AOIs depending on where they are located in the visual scene [81]. The duration of a fixation is the amount of time (e.g., in milliseconds) that the person spends focusing on that area of the visual scene before the next saccade [133]. Since duration is difficult to interpret on its own [133], it is often reported as a percentage, e.g., the total duration of all fixations to a particular AOI divided by the total duration of all fixations [81].

CHAPTER 3

DEEP REPRESENTATION LEARNING OF ADULT AND CHILD FACIAL EXPRESSIONS USING DOMAIN ADAPTATION FUSING FACIAL LANDMARK FEATURES

This chapter contains materials that have been reprinted, with permission, from [87]. Please find the copyright notice in Appendix A.

3.1 CHAPTER OVERVIEW

From infancy to adulthood, facial expressions are a ubiquitous, information-rich component of human social interactions. Developing models robust to age variations is a challenging problem in FEA [134, 135]. Most existing approaches optimize the FEA performance on data sets representing specific age ranges. There has been limited work on classifying facial expressions across age groups. Furthermore, age variations in facial images have been well-studied in facial age estimation and AIFR, but there has been little cross-pollination among these relevant research areas to improve FEA considering adult-child age variations. In the following sections, we discuss related work on the classification of adult and child expressions and methods from relevant research fields. Then, we propose a novel deep feature adaptation approach to the classification of adult and child expressions inspired by the state-of-the-art domain adaptation learning, facial age estimation, and AIFR literature.

3.1.2 RELATED WORK

This section briefly reviews related work on the classification of adult and child expressions and methods from relevant research fields including facial age estimation and AIFR.

3.1.2.1 CLASSIFICATION OF ADULT AND CHILD FACIAL EXPRESSIONS

Existing off-the-shelf FEA tools and research [36, 38, 39] have been mostly developed using adult benchmark data sets [135-137]. However, facial morphology and kinematics gradually develop throughout childhood [40, 41], resulting in a distribution shift between child and adult expression patterns. For models trained on adult data sets, the distribution shift toward adults poorly generalizes distinctive patterns in child expressions [42-45]. While benchmark data sets of child facial expressions remain limited, they are growing in number [138-140]. Therefore, there has been an emerging trend directed at the classification of child facial expressions [42-45]. Recently, deep transfer learning using CNNs has shown promise for child facial expression classification [43, 44, 141]. However, recent studies focus only on maximizing performance on child facial expression benchmarks, bounded by a limited age range and sample size [135]. Such models tuned for child expressions fail to generalize to adult expressions [45]. To overcome the poor generalization problem across age groups, limited existing work on facial expression classification involving mixed-age groups (child, adult, elderly) suggests two primary approaches: 1) curating a mixed age training set to match the age distribution of the test set [59], and 2) classifying images into age groups to determine the age-appropriate model for subsequent classification [58]. The first approach requires the age distribution of the test set to be known a priori with availability of benchmark data matching. The second approach requires a robust age group classifier to select an appropriate expression classifier model and benchmark data to train expression classifiers for individual age groups. Age group classification is a challenging problem [142, 143] and variations in expression make accurate age estimation even more challenging [58, 144]. Furthermore, developments in both facial structure and muscle movements contribute to visual differences in child and adult expressions. A child’s growth is a gradual and uniquely individual process, making the transition unclear when a child manifests the full spectrum of adult expressions.

Recently, domain adaptation has shown an interesting pathway to adapt an adult expression classification model using few child expression samples [44]. This approach utilizes a dual stream deep CNN architecture and semantically aligns the class conditional distributions of child and adult domains [44]. The underlying framework of this approach [145] is based on learning a domain-invariant latent representation. Such domain-invariant representations have shown to generalize even to unseen domains [145]. We hypothesize that learning a domain-invariant representation of expressions may prove effective for facial expression classification across child and adult domains.

3.1.2.2 RECOGNITION OF AGE-VARYING FACIAL IMAGES

While limited attention has been given to facial expression classification across age groups, facial age estimation [125] and AIFR [124] are active research areas. State-of-art approaches for facial age estimation and AIFR benefit from deep learning and fusion of geometric and texture features [125, 143, 146]. Geometric features derived from facial landmarks capture structural changes associated with childhood development while texture features capture skin artifacts, such as wrinkles, associated with adult aging [143, 146]. Contemporary studies continue to use traditional feature extraction methods, e.g., local binary patterns, histograms of oriented gradients, etc., but recently emphasize deep learning, e.g., CNNs, for texture feature extraction [123-125]. Common geometric landmark features include distances between landmarks, ratios of distances, and areas and angles of triangles formed by landmark triplets [142, 147-149]. Similar landmark features, including pairwise distances between landmarks and areas/angles of facial polygons formed by connecting neighboring landmark points, have also shown to be discriminative for FEA [118-121]. Therefore, we hypothesize that domain-invariant representation learning of adult and child facial expressions can benefit from a fusion of CNN-extracted and landmark-derived features.

The use of the same feature types in both facial age estimation and AIFR suggests a subset of features correlated with and invariant to age. Statistical latent variable models optimized using the Expectation-Maximization (EM) algorithm have been applied to AIFR to decompose feature sets into age and identity factors [124]. This approach identifies a set of discriminative features for identity recognition using the identity factor, representing facial identity features invariant to age [124]. Gong et al. [150] have first proposed this approach using hidden factor analysis (HFA). HFA assumes the independence of age and identity. [124]. However, different people may show signs of aging at different rates. To overcome the independence assumption, the modified HFA (MHFA) approach introduces an additional factor representing age and identity-correlated facial appearance variations [151]. Given that the appearance of facial expressions varies among individuals and age groups, we hypothesize that FEA can benefit from the decomposition of feature sets into those correlated with expression, domain (adult or child), and identity. However, MHFA assumes that data are independent and identically distributed (i.i.d.) following a normal distribution with homogenous variances, which may not be true for real world facial expression data. Furthermore, HFA and MHFA require the optimization of one model per feature, making high dimensional feature vectors computationally prohibitive [150, 151]. Thus, principal component analysis (PCA) has been used for dimensionality reduction prior to HFA or MHFA [150, 151]. While PCA guarantees that the first principal components explain more of the variance than subsequent principal components, such linear data projection method does not guarantee that the PCA feature space will be discriminative for classification. Each principal component is a linear combination of all input features, making it less intuitive to understand the contribution of individual features. Moreover, all features, even those with limited contribution to discriminability, are needed to reproduce the same principal components.

Very recently, the beta-mixture (betaMix) method [152] has been proposed to determine significant correlations among large numbers of variables using a mixture of beta distributions. The method, based on ideas and results from convex geometry, works well even for moderate sample sizes, e.g., $N = 10$ depending on the number of predictors, and does not require assumptions of i.i.d., normality, or homogeneity of variances. The betaMix method detects correlations among all the features at once, so the EM algorithm needs to be applied only once for all features rather than for individual features. Since the betaMix method is appropriate for large feature vectors, dimensionality reduction is not required and the feature correlations may be interpreted directly, allowing for greater understanding of the interaction between the features and domain, identity, and expression factors. The betaMix method has shown promising results across multiple applications, including feature selection and classification [152].

3.1.3 CONTRIBUTIONS

This chapter proposes novel deep domain adaptative FACial Expressions fusing BEtaMix SElected Landmark Features (FACE-BE-SELF) for domain-invariant expression classification. To the best of our knowledge, our proposed deep domain adaptive FACE-BE-SELF approach is the first to perform concurrent adult-child domain adaptation and learn a generalized expression representation that may be used for both child and adult facial expression classification. Our contributions are as follows:

- We fuse facial landmark measurements with deep feature representations for robust expression learning across age groups.
- Our facial landmark features are decomposed based on expression, domain, and identity correlations.
- A novel statistical method based on a mixture of beta distributions is proposed for facial feature selection for deep learning.
- A new variant of concurrent adult-to-child expression learning is performed to yield domain-invariant facial expression classification.
- The proposed domain adaptation method is compared to baseline CNN, transfer learning, and existing domain adaptation methods for facial expression recognition using multiple benchmark data sets.

The remainder of this chapter is organized as follows. Section 3.2 describes the methodology of our approach. Sections 3.3 and 3.4 present the results and discussion, respectively. Section 3.5 discusses limitations and Section 3.6 summarizes.

3.2 METHODS

This section describes and explains benchmark data sets, preprocessing steps, feature extraction, decomposition and selection of landmark features, deep learning models, deep domain adaptation, and experiments.

3.2.1 DATA SETS

We evaluate our proposed method using four data sets of facial expression images: 1) the Extended Cohn-Kanade (CK+) data set [136, 137], 2) the Aff-Wild2 data set [126, 127, 153-

157], 3) the Child Affective Facial Expression (CAFE) data set [138, 139], and 4) the Child Emotion Facial Expression Set (ChildEFES) [140]. We consider both posed and spontaneous data sets. While spontaneous data sets represent most expressions seen in daily life, posed expressions serve a valuable purpose in healthcare applications such as social reciprocity training [158-160] for individuals with ASD and facial rehabilitation exercises for individuals with Parkinson's disease and facial palsy [161, 162].

3.2.1.1 CK+ DATA SET

The CK+ data set [136, 137] consists of 593 image sequences of posed facial expressions, including labeled 'anger', 'disgust', 'fear', 'happy', 'sad', 'surprise', and 'contempt' examples, captured from 123 adult subjects (ages 18 to 50 years). A mixture of color and grayscale sequences are present in the data set. Sequences vary in length, but each sequence begins with the neutral expression and ends with the peak expression frame, which has been coded for AUs from FACS. We assign the last three frames of a sequence with its corresponding expression label and label the first frame of each sequence as 'neutral'. This yields 1,254 samples: 135 'anger', 177 'disgust', 75 'fear', 207 'happy', 327 'neutral', 84 'sad', and 249 'surprise'.

3.2.1.2 AFF-WILD2 DATA SET

The Aff-Wild2 data set [126, 127, 153-157], an extension of the Affect-in-the-Wild (Aff-Wild) [128, 131, 163] data set, consists of 558 YouTube videos with annotations for three behavioral tasks: valance and arousal, FACS AUs, and facial expressions ('anger', 'disgust', 'fear', 'happy', 'neutral', 'sad', 'surprise', and 'other'). The facial expression subset of Aff-Wild2 contains 84 videos with 84 ethnically diverse subjects (42 female). Age labels are not provided. Visually, the subjects appear to be mostly adults with few child subjects, including infants. Labeled frames show a variety of different head poses, occlusions, and illumination conditions. Excluding the frames labeled ‘other’, there are a total of 451,794 samples: 18,940 ‘anger’, 14,545 ‘disgust’, 11,336 ‘fear’, 97862 ‘happy’, 19,7314 ‘neutral’, 80,517 ‘sad’, and 31,280 ‘surprise’.

3.2.1.3 CAFE DATA SET

The CAFE data set [138, 139] consists of 1,192 color photographs of 154 child subjects (ages 2 to 8 years) posing ‘anger’, ‘disgust’, ‘fear’, ‘happy’, ‘sad’, and ‘surprise’ expressions, including ‘neutral’. The data set includes open and closed mouth variations for each expression except ‘surprise’, which is posed with open mouth only. We include the mouth closed variant of all expressions except for ‘surprise’, yielding 707 samples: 119 ‘anger’, 96 ‘disgust’, 79 ‘fear’, 120 ‘happy’, 129 ‘neutral’, 62 ‘sad’, and 102 ‘surprise’. The data usage agreement for the CAFE data set does not allow the republication of the images.

3.2.1.4 CHILDEFES DATA SET

The ChildEFES data set [140] consists of color photos and videos capturing 34 child subjects (ages 4 to 6 years) producing a mixture of spontaneous and posed ‘anger’, ‘disgust’, ‘fear’, ‘happy’, ‘sad’, ‘surprise’, and ‘contempt’ expressions. The expression labels are assigned based upon the agreement of four FACS judges. The expression-labeled videos were cropped to the peak expression. Then, the cropped videos were sampled at 20 frames per second to generate image sequences. Since the photographs are a subset of the image sequences, only the frames sampled from the videos are included. This yields 9,420 (5,107 spontaneous) samples: 1,435

(170) ‘anger’, 1,196 (468) ‘disgust’, 655 (19) ‘fear’, 2,196 (1,535) ‘happy’, 2,445 (2,372) ‘neutral’, 1,053 (450) ‘sad’, and 440 (93) ‘surprise’. The data usage agreement for the ChildEFES data set does not allow for the use of the images in publications.

3.2.1.5 NOTATION

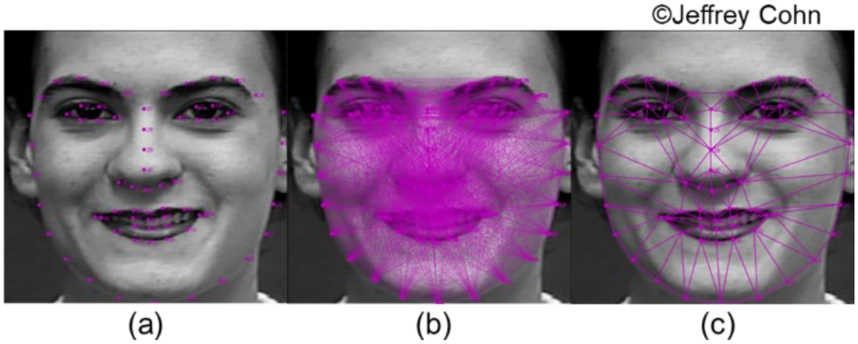
Let input space $\mathcal{X}$ represent the set of all possible facial images and features. Output space $\mathcal{Y} = \{1, \ldots, K\}$ is the set of $K = 7$ expression class labels (‘anger’, ‘disgust’, ‘fear’, ‘happy’, ‘neutral’, ‘sad’, ‘surprise’). $\mathcal{X}$ and $\mathcal{Y}$ are related by a function $f: \mathcal{X} \to \mathcal{Y}$. We consider adult facial expressions (CK+, Aff-Wild2) as the source domain and child facial expressions (CAFE, ChildEFES) as the target domain. We represent each source data set as $D_S = \{(x_i^S, y_i^S) \mid x_i^S \in \mathcal{X}, y_i^S \in \mathcal{Y}\}_{i=1}^{N_S}, x_i^S \sim \mathcal{P}_X^S$ where N_S is the total number of samples and $\mathcal{P}^S$ is the source probability distribution. We represent each target dataset as $D_T = \{(x_i^T, y_i^T) \mid x_i^T \in \mathcal{X}, y_i^T \in \mathcal{Y}\}_{i=1}^{N_T}, x_i^T \sim \mathcal{P}_X^T$ where N_T is the total number of samples and $\mathcal{P}^T$ is the target probability distribution.

3.2.2 PREPROCESSING

Data sets are preprocessed following [43]. The dlib (http://dlib.net/) library is used to detect the face in each image and extract landmark coordinates on the face. The landmarks are used to center and rotate the face so that the eyes are level. The images are cropped in such a way that the left eye is located 30% of the image width in pixels from the left edge. Images are resized to 256 by 256 pixels, converted to grayscale, and normalized to range [0, 1].

3.2.3 FEATURE EXTRACTION

Using the dlib library, we extract landmark points located at and around facial features such as the nose, eyes, mouth, and eyebrows as well as the perimeter of the face. These landmark locations are used to derive geometric features from FEA and AIFR literature based on pairs of landmarks and triplets of landmarks. Inter-landmark distance features [121, 143, 148, 149] are measured as the Euclidean distance between pairs of landmarks. Facial triangles [120, 147, 148] are extracted based on a Delaunay triangulation over the landmark locations. Each triangle is represented by a landmark triplet and has four associated features: the area of the triangle and its three angles expressed in radians. Figure 5 shows examples of the extracted features.



© 2023 IEEE

Figure 5. Sample image overlaid with: (a) facial landmark points, (b) inter-landmark distance features, (c) Delaunay triangulation of the face.

3.2.4 LANDMARK FEATURE DECOMPOSITION AND SELECTION

We fit the betaMix method [152] to find significant correlations between the extracted features from adult-child data and three experimental factors taken from the labeled data: expression, domain, and identity. Based on given data, the betaMix method automatically learns a threshold that represents significant correlations among pairwise landmark features (predictors) and factors (domain, expression, and subject identity). The extracted features for the source and target data sets are concatenated to form a matrix of p predictors and N samples, where $p > N$. Expression, domain, and identity are also entered into the model, yielding a $(p + 3) \times N$ data matrix. We assume the data as $p + 3$ points in $\mathbb{R}^N$. Subspaces of $\mathbb{R}^N$ lie on the Grassmann manifold (see [164] for a detailed exposition of the Grassman manifold). The Grassmann manifold $\mathbb{G}_{N,d}$ is used to study d-dimensional subspaces of $\mathbb{R}^N$ [152, 164]. For principal angles $(\theta_1, \dots, \theta_d)$ between subspaces in $\mathbb{R}^N$, $\mathbb{G}_{N,d}$ has an invariant measure that can be used to compute the volume and probability of their sets [152]. These principal angles can be used to determine canonical correlations $(\rho_1, \dots, \rho_d)$ as $\rho_j = \cos \theta_j$ with pairs of canonical variables lying on $\mathbb{G}_{N,d}$ [152]. $\mathbb{G}_{N,1}$ corresponds to lines through the origin of Euclidean space [152]. The line is a natural choice of projection due to its computational ease and interpretability. Furthermore, with $\mathbb{G}_{N,1}$, the random variable $\sin^2 \theta_j$ has the following beta distribution [152]:

$$\lambda \stackrel{\text{def}}{=} \sin^2 \theta_j \sim beta\left(\frac{N-1}{2}, \frac{1}{2}\right). \tag{22}$$

Thus, we consider that the predictors and factors lie on $\mathbb{G}_{N,1}$ and define θ_j as the angle between the jth pair of predictors/factors, with a total of $((p + 3)(p + 2))/2$ pairs of predictors/factors, and $j = 1, \dots., ((p + 3)(p + 2))/2$ [152]. We let $\lambda_j = \sin^2 \theta_j$. A predictor-predictor or predictor-factor pair is considered ‘null’ if it corresponds to randomly sampled (uncorrelated) points in $\mathbb{R}^N$. As shown in [152], pairs of null predictors/factors are expected to be mutually perpendicular with high probability, even for moderate values of N. In relation to Equation (16), a mixture of beta distributions may be used to determine if the pair of predictors/factors represented by each λ_j are null (uncorrelated) or nonnull (correlated). Then, the betaMix model is defined as:

$$l(\lambda_j) = \iota_{0_j} d_0(\lambda_j) + \left(1 - \iota_{0_j}\right) d(\lambda_j), \tag{23}$$

where $d_0(\lambda_j)$ is the null distribution, $d(\lambda_j)$ is the alternative distribution and $\iota_{0_j} \sim Ber(r_0)$ with probability of the null component r_0 is a random indicator that equals one if the jth pair of predictors/factors corresponding to λ_j are null. The null component of the mixture model is defined by the beta distribution:

$$d_0(\lambda_j) = \frac{1}{beta(\frac{s-1}{2}, \frac{1}{2})} \lambda_j^{\frac{s-1}{2}-1} (1 - \lambda_j)^{-\frac{1}{2}}, \tag{24}$$

where $s \leq n$ is the estimated effective sample size. The nonnull component of the mixture model is defined as:

$$d(\lambda_j) = \frac{1}{beta(\alpha, \beta)} \lambda_j^{\alpha-1} (1 - \lambda_j)^{\beta-1}, \tag{25}$$

where $\alpha, \beta > 0$. The latent mixture variables (α, β, s) are estimated using the EM algorithm.

The E-step updates ι_{0_j} with the posterior mean:

$$\hat{\iota}_{0_j} = \frac{r_0 d_0(\lambda_j)}{r_0 d_0(\lambda_j) + (1 - r_0) d_0(\lambda_j)}, \tag{26}$$

and r_0 is updated with its maximum likelihood estimate, $\hat{r}_0 = \mathbb{E}(\hat{\iota})$. The M-step obtains the maximum likelihood estimates of α, β, and s by solving the following equations:

$$\psi(\alpha) - \psi(\alpha + \beta) = \frac{\sum_{j=1}^{((p+3)(p+2))/2} \left(1 - \iota_{0_j}\right) \log(\lambda_j)}{\sum_{j=1}^{((p+3)(p+2))/2} \left(1 - \iota_{0_j}\right)}, \tag{27}$$

$$\psi(\beta) - \psi(\alpha + \beta) = \frac{\sum_{j=1}^{((p+3)(p+2))/2} \left(1 - \iota_{0_j}\right) \log\left(1 - \lambda_j\right)}{\sum_{j=1}^{((p+3)(p+2))/2} \left(1 - \iota_{0_j}\right)}, \tag{28}$$

$$\psi\left(\frac{s-1}{2}\right) - \psi\left(\frac{s}{2}\right) = \frac{\sum_{j=1}^{((p+3)(p+2))/2} \iota_{0_j} \log\left(\lambda_j\right)}{\sum_{j=1}^{((p+3)(p+2))/2} \iota_{0_j}}, \tag{29}$$

where $\psi(\cdot)$ is the digamma function. The E- and M-steps are repeated iteratively to update the parameters until convergence. Pairs of predictors/factors are considered nonnull if the posterior null probability under d_0 is smaller than threshold τ, $\hat{\iota}_{0_j} < \tau$. We denote the maximum λ_j that satisfies $\hat{\iota}_{0_j} < \tau$ as Q. Then, the screening rule for nonnull pairs may be written as $\lambda_j < Q$. Since $\lambda_j = \sin^2 \theta_j$ and $\rho_j = \cos \theta_j$, pairs with a correlation of at least $\rho = \cos\left(\sin^{-1}(Q^{1/2})\right)$ are considered significant. Figure 6 summarizes the betaMix method.

Based on the fitted beta distribution, a graphical model is built where each node is a predictor or factor. An edge connects each nonnull predictor-predictor pair or factor-predictor pair. These edges represent a significant correlation between the connected nodes (predictors or factors). A subgraph formed by a factor and its adjacent predictor nodes captures the subset of predictors that are significantly correlated with the factor. Using these subgraphs, we decompose the feature vector into sets correlated with expression, domain, and identity. For our proposed FACE-BE-SELF approach, we select the features in the expression subgraph and prune features that also appear in the domain or identity subgraphs. The resulting selection of features is used in subsequent feature fusion.

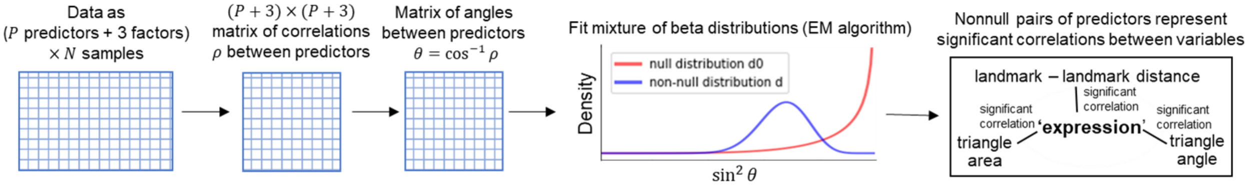


Figure 6. Overview of the betaMix method.

3.2.5 DEEP LEARNING MODELS

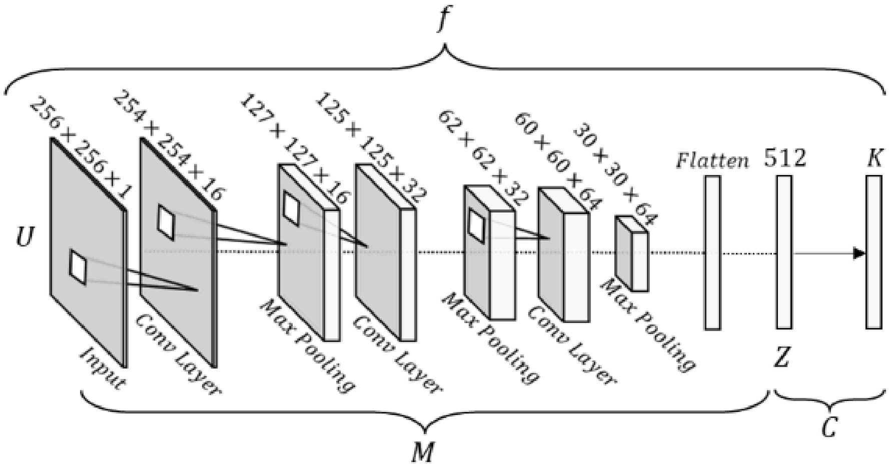
We model supervised classification as the following inverse problem:

$$Y = f(X; W) \tag{30}$$

where $f(\cdot)$ is a neural network model parameterized by weights W, $X \in \mathcal{X}$ are the model inputs, and $Y \in \mathcal{Y}$ are the associated class labels. We partition $f(\cdot)$ into feature extractor $M: \mathcal{X} \to \mathcal{Z}$ and classifier $C: \mathcal{Z} \to \mathcal{Y}$ such that $f = C \circ M$ with latent feature space $\mathcal{Z}$. Using this notation, we define multiple architectures: MLP, CNN, and feature fusion model including MLP and CNN components.

For the MLP, we consider $X = V$, where feature set $V \in \mathcal{V} = \mathbb{R}^{p_{beta}}$ and p_{beta} is the number of betaMix-selected features based on significant correlations with expression. The MLP has one hidden layer with 512 hidden units, ReLU activation, and dropout with a probability of 0.5. We consider a latent feature vector $Z \in \mathcal{Z} = \mathbb{R}^{512}$ produced by the hidden layer of the MLP. The hidden layer is followed by a softmax output layer of $K = 7$ $(number\ of\ classes)$ nodes. Uniform initialization is applied to all of the MLP weights.

For the CNN, we consider $X = U \in \mathcal{U} = \mathbb{R}^{256\times256}$ and define $M(\cdot)$ as a sequence of three convolutional blocks, each consisting of a convolutional layer with 3x3 filter kernels followed by a 2x2 max pooling, and a fully connected neural network with 512-dimensional hidden layer. This hidden layer also yields a latent feature vector $Z \in \mathcal{Z} = \mathbb{R}^{512}$. The uniform distribution was used to initialize all weights. Dropout with a probability of 0.5 is applied to the 512-dimensional hidden layer. We define $C(\cdot)$ as a K-dimensional fully connected layer with softmax mapping from $\mathcal{Z}$ onto $\mathcal{Y}$. This CNN architecture is shown in Figure 7.



© 2023 IEEE

Figure 7. CNN architecture. The model is partitioned into a feature extractor that maps from input to latent feature space and classifier from the latent feature space to the output space.

For the proposed FACE-BE-SELF feature fusion model, we define $X = (U, V)$, where $U \in \mathcal{U} = \mathbb{R}^{256\times256}$ and $V \in \mathcal{V} = \mathbb{R}^{p_{beta}}$, where p_{beta} is the number of betaMix-selected features based on significant correlations with expression. Feature extractor $M(\cdot)$ is made up of CNN model $G: \mathcal{U} \rightarrow \mathcal{Z}_G, \mathcal{Z}_G = \mathbb{R}^{512}$ and the MLP model $H: \mathcal{V} \rightarrow \mathcal{Z}_H, \mathcal{Z}_H = \mathbb{R}^{512}$. We define the concatenation of Z_G and Z_H spaces as $Z \in \mathcal{Z} = \mathbb{R}^{1024}$. Then, we define $C(\cdot)$ as a K-dimensional fully connected layer with softmax mapping from $\mathcal{Z}$ onto $\mathcal{Y}$. The architecture of the feature fusion model is shown in Figure 8.

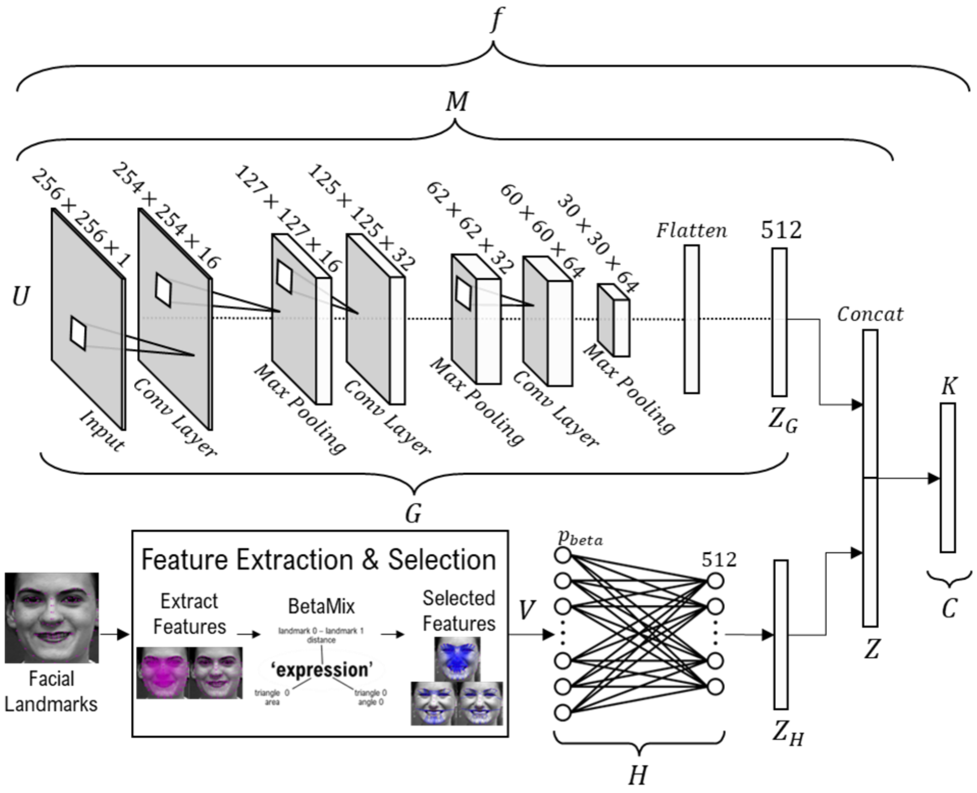
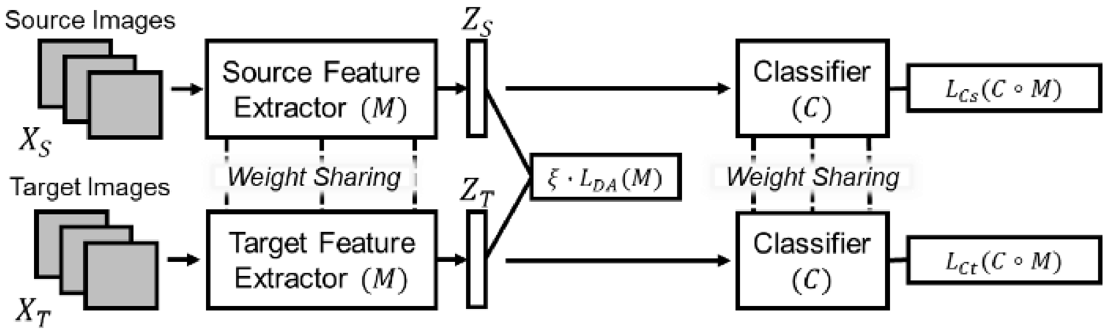


Figure 8. Feature fusion architecture.

3.2.6 DEEP DOMAIN ADAPTATION

Rather than maximizing the performance on a target domain, our goal for deep domain adaptation is to optimize the model for maximum performance on both the source and target domains. We assume that the distribution shift between source and target domains can be attributed to covariate shift $\mathcal{P}_X^S(x) \neq \mathcal{P}_X^T(x)$ and assume $\forall x \in \mathcal{X},\ \mathcal{P}^S(Y \mid X = x) = \mathcal{P}^T(Y \mid X = x)$. We adopt a dual stream architecture (Figure 9) consisting of parallel feature extractors $M_S(\cdot)$ and $M_T(\cdot)$ for source and target distributions, respectively. Weights are shared between the two branches such that $M(\cdot) = M_S(\cdot) = M_T(\cdot)$. Paired source and target examples X_S and X_T are passed into their respective feature extractors to yield source and target latent representations, i.e., $Z_S = M(X_S)$ and $Z_T = M(X_T)$. Parallel classifiers $C(\cdot)$, which also share weights, are trained with Z_S and Z_T to optimize performance on both source and target domains.



© 2023 IEEE

Figure 9. Domain adaptation framework. Source-target pairs are passed into parallel feature extractors. Resulting latent distributions are aligned by the domain alignment loss. Parallel classifiers are supervised by source and target classification losses.

The model is optimized using three supervised loss functions: source classification loss $\mathcal{L}_{Cs}(f)$, target classification loss $\mathcal{L}_{Ct}(f)$, and domain alignment loss $\mathcal{L}_{DA}(M)$. We define $\mathcal{L}_{Cs}$ and $\mathcal{L}_{Ct}$ as the categorical cross-entropy loss given our multiclass expression classification problem. To address class imbalance in the training sets, we scale each sample's contribution to the overall loss by the frequency of its associated class in the training set. We define $\mathcal{L}_{DA}$ as the contrastive alignment loss [145]:

$$\mathcal{L}_{DA}(M) = \sum_{a=1}^{K} \sum_{i,j} d\left(M\left(x_i^S | y_i^S = a\right), M\left(x_j^T | y_j^T = a\right)\right) + \sum_{a,b|a \neq b}^{K} \sum_{i,j} k\left(M\left(x_i^S | y_i^S = a\right), M\left(x_j^T | y_j^T = b\right)\right), \tag{31}$$

with $d(\cdot)$ and $k(\cdot)$ defined as:

$$d(M\left(x_i^S\right), M(x_j^T)) = 0.5\left\|M\left(x_i^S\right) - M(x_j^T)\right\|_F^2 \tag{32}$$

and

$$k(M\left(x_i^S\right), M(x_j^T)) = \frac{1}{2}(max(0,\, m - \left\|M\left(x_i^S\right) - M\left(x_j^T\right)\right\|_F))^2, \tag{33}$$

where $\|\cdot\|_F$ is the Frobenius norm and margin $m = 1$ [145]. The effect of $\mathcal{L}_{DA}$ is to minimize the distance between samples of the same class from different domains, and the similarity between samples of different classes and domains. The overall loss is:

$$\mathcal{L} = (\mathcal{L}_{Cs} + \mathcal{L}_{Ct}) + \xi \mathcal{L}_{DA}, \tag{34}$$

where $0 < \xi < 1$ is a scaling parameter for balancing the contribution of domain alignment loss.

3.2.7 EXPERIMENTS

We perform preprocessing of the CAFE, ChildEFES, CK+, and Aff-Wild2 data sets following Section 3.2.2. To evaluate our proposed FACE-BE-SELF method, we consider data sets in two source/target pairs: CK+/CAFE (posed expressions only) and Aff-Wild2/ChildEFES (majority spontaneous expressions). We split each data set into multiple train, validation, and test sets using a 5x2 nested cross-validation design. In the outer 5-fold cross-validation loop, the data is split into train (4 folds) and test (1 fold) sets. In the inner loop, the train set is divided into 2 to yield inner train and validation sets for hyperparameter selection. The validation performance metrics are averaged across the two folds to yield the best hyperparameters. These hyperparameter selections are then used to train the model with the recombined outer loop training set and evaluate on the held-out test fold. This procedure is repeated a total of 5 times, such that each sample appears in one of the 5 test sets. To avoid inflation of performance estimates based on subject-specific features, we generate the cross-validation folds such that each subject appears in one fold only and no subject appears in both train and test sets [43, 44].

We fit the betaMix method on the train sets for each source/target pair. The fitted mixture model identifies nonnull (significantly correlated) pairs of predictors/factors which are used to build a graph with predictors/factors represented as nodes and significant correlations represented as edges. By examining the subgraphs of each factor node and its adjacent predictor nodes, we report the mean number of significantly correlated features for each factor and the overlap of features appearing in multiple factor subgraphs. To select features for subsequent fusion, we consider the expression subgraph, pruning features that also appear in the domain and/or subject subgraphs. We assess the discriminability of our data-driven feature selection compared to that of features selected based upon a range of correlation thresholds (0.1, 0.2, …, 1.0).

The average overall F1 performance on the inner 2-fold cross-validation loop is used to select a value for the loss balancing parameter ξ (Equation (34)) for each of the outer 5-fold cross-validation training sets. Other studies [165, 166] have found that ξ is problem-specific and consider values in the range $(0.00, 1.00)$. Due to high computational costs, we choose among representative low (0.01), moderate (0.3), and high (0.8) values in $(0.00, 1.00)$. To better understand the contributions of CNN, betaMix-selected landmark features, and domain adaptation to our proposed FACE-BE-SELF approach, we perform an ablation study.

Then, we evaluate the performance of our proposed domain adaptation with FACE-BE-SELF approach on two source/target data set pairs and compare against four baseline models: 1) CNN trained on source data (source CNN) [43], 2) CNN trained on target data (target CNN) [43], 3) three transfer learning approaches (pretraining on source data then, a, training on target data [43], b, fine-tuning on the target data [43], or c, fine-tuning on a mixture of source and target data), and 4) two existing domain adaptation approaches [44, 145].

For all experiments, we train deep models using the ADAM optimizer with a triangular learning rate policy [167] cycling between a minimum learning rate of $\zeta_{min} = 10^{-5}$ and a maximum learning rate of $\zeta_{max} = 10^{-3}$. We use a batch size of 32.

3.3 RESULTS

This section describes the results of the feature extraction; selection of landmark features for expression, domain, and identity factors; and domain adaptation.

3.3.1 FEATURE EXTRACTION

We extract 68 landmark points on the face as shown in Figure 5 (a) and use these to measure inter-landmark distances. Because the 68×68 Euclidean distance matrix is symmetric with zeros (self-distance) in diagonal entries, the total number of inter-landmark distance features is $(68\ landmarks\ \times 68\ landmarks) - 68/2 = 2278$. Figure 5 (b) overlays all possible inter-landmark distance features on the face. The Delaunay triangulation over the landmark locations results in a set of 106 triangles on the face. For each facial triangle, the area and three internal angles are computed, resulting in $106\ triangles\ \times (4\ features/triangle) = 424$ triangle-based features. Figure 5 (c) visualizes the Delaunay triangulation on the face.

3.3.2 SELECTION OF LANDMARK FEATURES FOR EXPRESSION, DOMAIN, AND IDENTITY FACTORS

We fit the betaMix method on each of the 5-fold cross-validation training sets for both source/target data set pairs. For a representative CK+/CAFE training set, betaMix learns the screening rule $\lambda_j = \sin^2(\theta) < Q = 0.83$. This is equivalent to an angle of $65.7°$ or less between the pairs of factors/features, or a correlation coefficient of at least $\rho = \cos(65.7°) = 0.412$. Averaging over the 5 training sets, the mean correlation threshold for CK+/CAFE is 0.414±0.025. Similarly, for a representative Aff-Wild2/ChildEFES training set, betaMix learns the screening rule $\lambda_j = \sin^2(\theta) < Q = 0.99$. This is equivalent to an angle of $84.3°$ or less between the pairs of factors/features, or a correlation coefficient of at least $\rho = \cos(84.3°) = 0.100$. For Aff-Wild2, the mean correlation threshold is 0.045±0.031. Figure 10 shows the mean number of features correlated with ‘expression’, ‘domain’, and ‘identity’, as well as the number correlated with two out of three and all three factors.

Considering the CK+/CAFE pair of data sets, Figure 11 compares the performance of an MLP trained on features selected by the data-driven correlation threshold learned by betaMix and those selected based upon a range of correlation thresholds (0.1, 0.2, …, 1.0). There is not any feature with a correlation coefficient of 0.6 or greater for the expression factor.

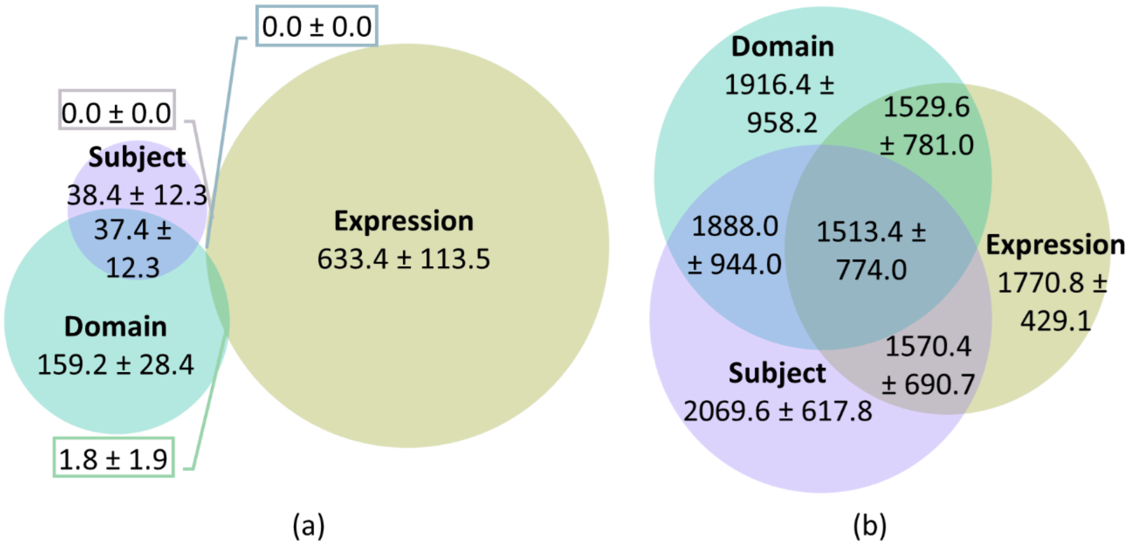


Figure 10. Mean number of features correlated with expression, domain, and identity for (a) CK+/CAFE and (b) Aff-Wild2/ChildEFES.

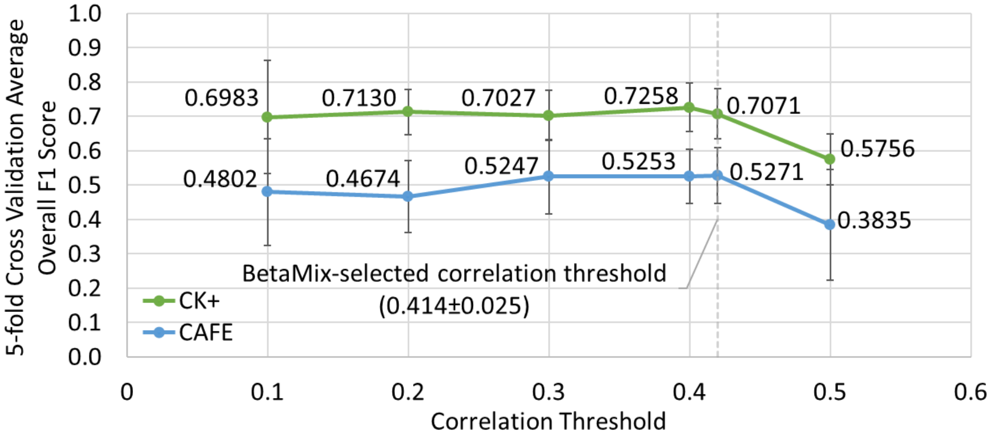


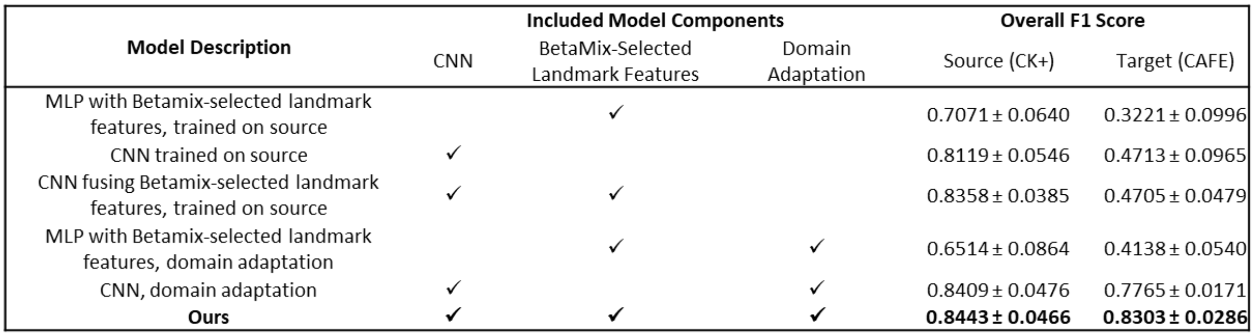
Figure 11. CK+/CAFE 5-fold cross-validation average overall F1 scores for MLP trained on expression-correlated feature selections at various thresholds.

3.3.3 DOMAIN ADAPTATION

For each train/test split in the outer 5-fold cross-validation loop, we select ξ based on the overall F1 score averaged over the 2 validation sets of the inner 2-fold cross-validation loop. For CK+/CAFE, $\xi = 0.01$ is selected for all 5 train/test splits of the outer cross-validation loop. For Aff-Wild2/ChildEFES, $\xi = 0.01$ is selected once, $\xi = 0.3$ is selected twice, and $\xi = 0.8$ is selected twice.

Ablation study results for FACE-BE-SELF are presented in Table 3. The proposed model is compared with variants that selectively remove one or two of the following model components: CNN, betaMix-selected landmark features, and domain adaptation.

Table 3. Ablation study for the proposed FACE-BE-SELF model



Model Description	**Included Model Components**			**Overall F1 Score**	
	CNN	BetaMix-Selected Landmark Features	Domain Adaptation	Source (CK+)	Target (CAFE)
MLP with Betamix-selected landmark features, trained on source		✓		0.7071 ± 0.0640	0.3221 ± 0.0996
CNN trained on source	✓			0.8119 ± 0.0546	0.4713 ± 0.0965
CNN fusing Betamix-selected landmark features, trained on source	✓	✓		0.8358 ± 0.0385	0.4705 ± 0.0479
MLP with Betamix-selected landmark features, domain adaptation		✓	✓	0.6514 ± 0.0864	0.4138 ± 0.0540
CNN, domain adaptation	✓		✓	0.8409 ± 0.0476	0.7765 ± 0.0171
Ours	✓	✓	✓	**0.8443 ± 0.0466**	**0.8303 ± 0.0286**

Figure 12 compares the 5-fold cross-validation performance of our proposed FACE-BE-SELF with multiple baselines for the CK+/CAFE and Aff-Wild2/ChildEFES source/target pairs, including CNNs trained on a single domain [43], transfer learning [43], and domain adaptation approaches [44]. Since transfer learning performance on the source domain is expected to deteriorate after fine-tuning on target data only, we also compare with transfer learning finetuned on a mixture of source and target data.

Figure 13 plots one-versus-rest multiclass ROC curves and reports the AUC metrics for the proposed FACE-BE-SELF approach.

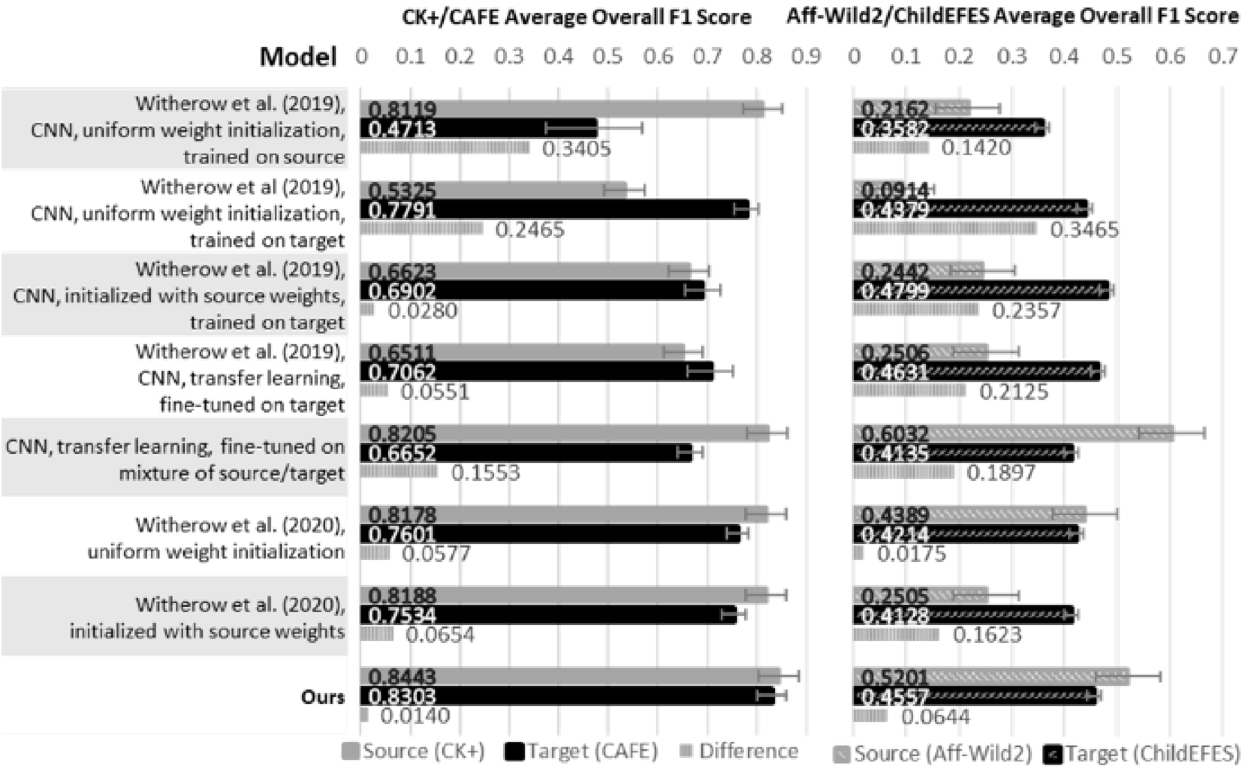


Figure 12. 5-fold cross-validation overall F1 score for comparison models.

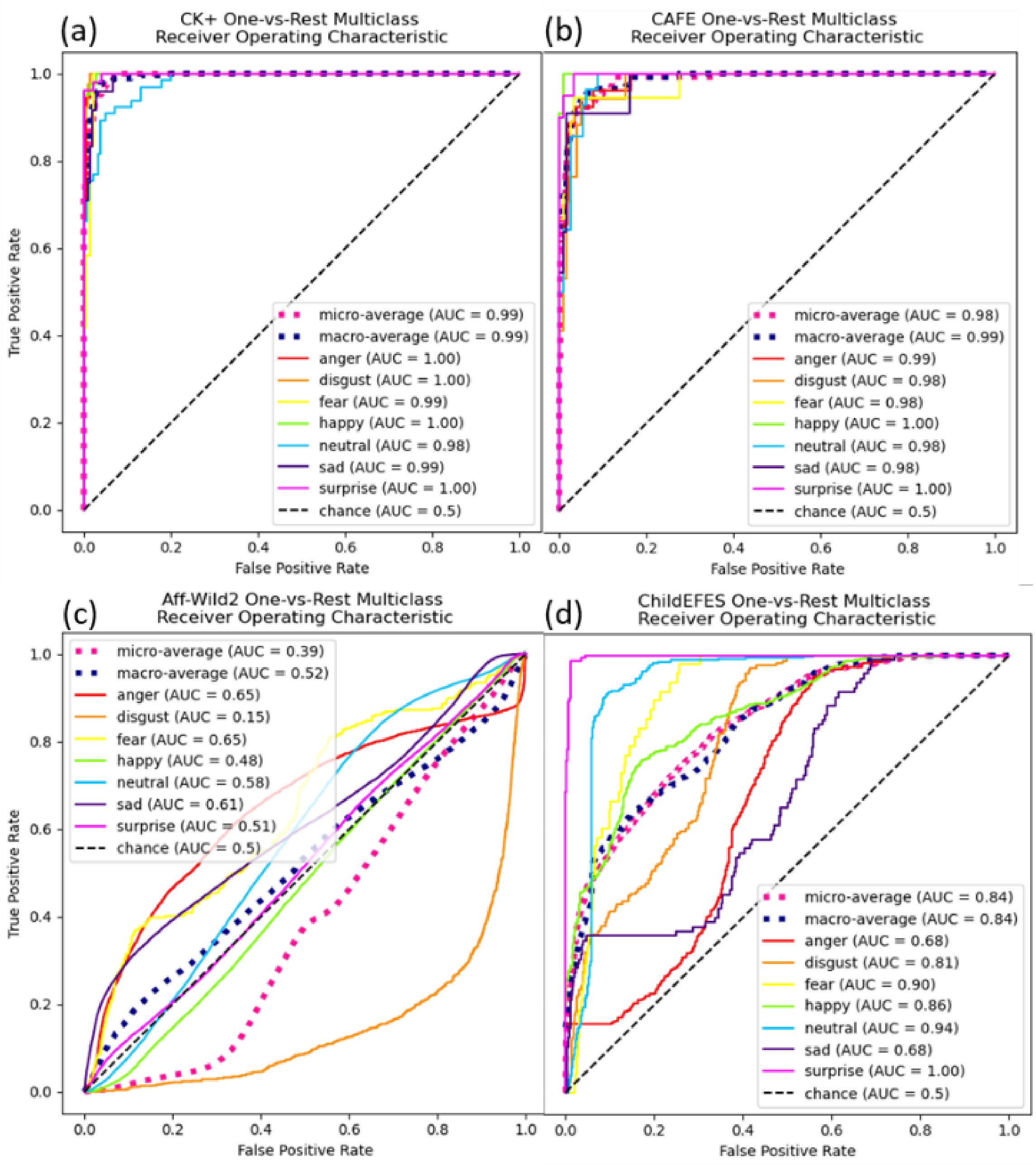


Figure 13. FACE-BE-SELF ROC Curves for various data sets.

3.4 DISCUSSION

This chapter presents FACE-BE-SELF for classification of adult and child facial expressions through deep domain adaptation and the fusion of facial landmark features correlated with expressions. Our experiments on four data sets and comparison of eight facial expression classification methods have revealed four important findings as follows. First, the decomposition of landmark features for expression, domain, and identity factors based on the data-driven threshold learned by betaMix reveals very little overlap in the subgraphs of different factors for CK+/CAFE (Figure 10 (a)) while the factor subgraphs of Aff-Wild2/ChildEFES share a substantial number of adjacent feature nodes (Figure 10 (b)). Features concurrently correlated with expression and domain factors indicate the presence of domain shift in the landmark feature space $\mathcal{V} = \mathbb{R}^{p_{beta}}$. While the CNN feature space is known to exhibit adult-child domain shift [42-45], our results suggest the domain shift to be dependent on the domain data set pair. The underlying data dependency (differences in overlap regions of Figure 10 (a) and Figure 10 (b)) may be attributed to differences in sample size and demographics, age ranges, and/or mixture of posed/spontaneous expressions [134, 135]. Second, a parsimonious feature selection is obtained from the expression subgraph after eliminating features significantly correlated with the other factors (Figure 11). Third, our ablation study shows that fusing these selected landmark features and CNN-extracted image features improves the expression classification performance for both child and adult data (Table 3). Fourth, our proposed FACE-BE-SELF method outperforms all baseline models for the posed data sets (CK+/CAFE) and performs competitively for the data sets with spontaneous expressions (Aff-Wild2/ChildEFES). The sections to follow provide detailed discussions in addition to and expanding upon these four key findings.

3.1.1 SELECTION AND FUSION OF FACIAL LANDMARK FEATURES

Our comparison of the proposed data-driven feature selection and a range of correlation thresholds reveals that our data driven betaMix approach yields the largest correlation threshold prior to substantial performance degradation (Figure 11). This threshold corresponds to a parsimonious selection of highly correlated features that preserve useful complementary information for expressions that is discarded at higher thresholds. Furthermore, fusing CNN-extracted features with the selected landmark features improves the classification performance of child and adult facial expressions (Table 3). Like age estimation and AIFR, facial expression classification also benefits from the fusion of geometric landmark and texture features [124, 125]. Given that the feature fusion model outperforms CNN features only, our selected landmark features provide complementary information representative of expressions beyond that learned by the CNN (Table 3). The effectiveness of selected features in classification suggests that the proposed betaMix correlation coefficient threshold is an effective metric in optimizing feature selection for facial expression classification.

3.1.2 DOMAIN ADAPTATION FOR EXPRESSION LEARNING

Our findings suggest that domain adaptation methods provide robust representation learning of adult and child facial expressions (Figure 12). During adaptation, source and target performance are jointly optimized via $\mathcal{L}_{Cs}$ and $\mathcal{L}_{Ct}$ while the class conditional distributions are aligned using $\mathcal{L}_{DA}$. This optimization procedure ensures balanced performance on both domains. Our findings also confirm that supervision on both domains (as in transfer learning) or a method of domain alignment is required for effective classification (Figure 12). For both CK+/CAFE and Aff-Wild2/ChildEFES source-target pairs, we observe poor cross domain performance for CNNs trained on a single domain (Figure 12). This poor cross domain performance is indicative of distribution shift and replicates the findings of multiple prior studies [43-45].

Our proposed FACE-BE-SELF method yields higher source and target average overall F1 scores for CK+/CAFE than all baseline models, with similar average overall F1 scores for source and target of 0.8443 and 0.8303, respectively, with a difference of 0.0140 (Figure 12). Spontaneous expression classification (Aff-Wild2/ChildEFES) is more challenging than classification of posed facial expressions such as CK+ and CAFE. For example, Aff-Wild2 is the most challenging of the four data sets that we use to evaluate our approach. Current state-of-the-art performance on the official test set for Aff-Wild2 is an overall F1 score of 0.3587, achieved by the best performing team at the recent 3rd Affective Behavior Analysis in-the-wild Competition [168]. Please note that our results are not directly comparable as we perform cross-validation rather than use the official test set. For Aff-Wild2/ChildEFES, the best performing models are FACE-BE-SELF and fine-tuning on a mixture of source and target data (Figure 12). Compared to fine-tuning on a mixture of source and target data, FACE-BE-SELF performs better on ChildEFES (average overall F1 score 0.4557 > 0.4214) and worse on Aff-Wild2 (average overall F1 score 0.5201 < 0.6032) but has a smaller difference in source and target performance (0.0644 vs 0.1897). Thus, despite poorer performance on Aff-Wild2, FACE-BE-SELF offers better target (ChildEFES) performance and more balanced performance between source and target.

The ROC curves for CK+/CAFE (Figure 13 (a)(b)) reveal that despite class imbalance during training, FACE-BE-SELF learns to recognize all classes with AUCs near unity, indicating high sensitivity and specificity. For ChildEFES, the ROC curves show that all classes perform better than chance (Figure 13 (d)). ‘Surprise’, with its distinctive open mouth appearance achieves an AUC of unity while negative expressions ‘anger’ and ‘sad’ prove more difficult. The ROC curves for Aff-Wild2 (Figure 13 (c)) reflect the challenging nature of the data set with an overall average AUC of 0.52, close to chance level (0.50). The best performing classes are ‘anger’ (AUC 0.65) and ‘fear’ (AUC 0.65), while ‘disgust’ performs worst (0.15).

3.4.4 EXPLAINING FEATURE CONTRIBUTIONS

We perform additional analysis using SHapley Additive exPlanations (SHAP) [169] to explain the contributions of different features to the classification of child and adult expressions. To quantify the contributions of both betaMix-selected landmark features and CNN features in the feature fusion model, we use the expected gradients method [170] as implemented in the SHAP library (https://github.com/slundberg/shap) to obtain and visualize the (approximate) SHAP values for both landmark and CNN features. Figure 14 visualizes the SHAP values for source (CK+) and target (CAFE) domains. We use the same image from the CK+ data set for all visualizations. Figure 14 (a) shows the SHAP values associated with source and target image inputs to the CNN feature extractor. For both source and target, areas of the input with the greatest (positive or negative) contribution to expression classification are those involved in producing facial expressions: the eyebrows, eyes, nose, and mouth. Figure 14 (b) and Figure 14 (c) visualize the SHAP values of the top ten most important landmark features for source and target, respectively, ranked based on their mean absolute SHAP value. Figure 14 (d) plots these top ten features. Nine out of the ten features are the same for the source and target sets. The top four features, which are ranked in the same order for both source and target, are areas of triangles located at the right corner of the lips (2 features), left corner of the lips (1 feature), and between the left eye and eyebrow (1 feature). The symmetric features (the second area at the left corner of the lips and area between the right eye and eyebrow) are also among the top ten most important features, but in different orders of importance. In addition to these six triangle area features, three inter-landmark distance features are ranked among the top ten for both domains. These features represent distances between the mouth and eyes (2 features) and the mouth and nose (1 feature). As with the image input, the top ten landmark features represent important areas of the face for producing expressions: the eyebrows, eyes, and mouth.

3.5 LIMITATIONS

Although the betaMix method is robust to dependence among samples, the high degree of similarity among faces (compared to other types of data) and universality of expressions may yield a small effective sample size. Even with a small effective sample size, betaMix is shown to capture significantly correlated landmark features. However, there may be features that are useful for classification of expressions but are not significantly correlated with expression based on the betaMix-learned minimum correlation coefficient. Furthermore, the data dependency of betaMix feature selections may affect performance on unseen data sets. An additional adaptation or fine-tuning step may be required for these models to address possible data dependency. The age ranges studied cover 2 to 8 years for CAFE, 4 to 6 years for ChildEFES, and 18+ years for CK+. Aff-Wild2 does not report specific age ranges. Further research is required to determine if the adapted models are capable of generalizing to participants in other age groups, e.g., teens and pre-teens.

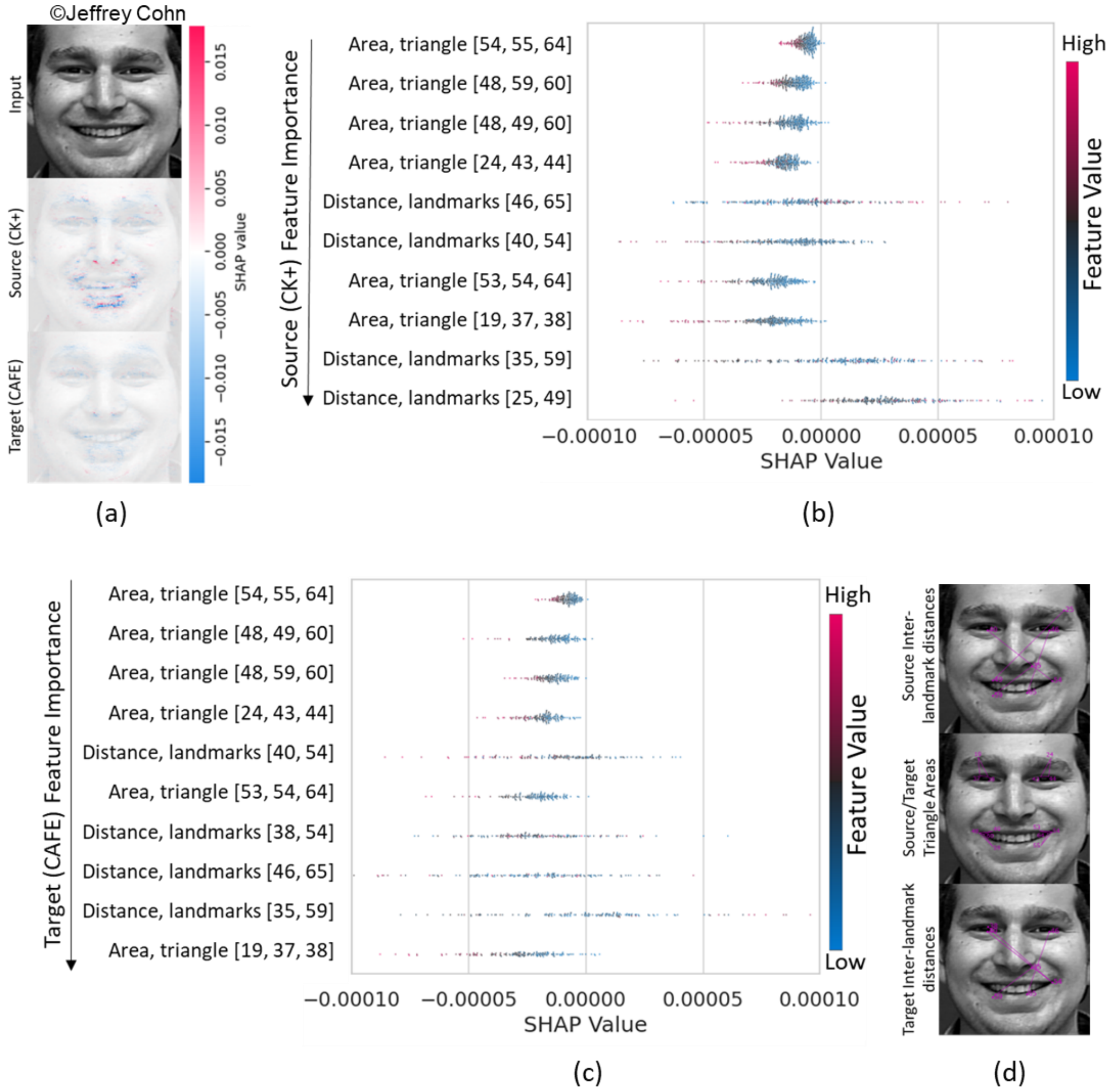


Figure 14. Visualization of SHAP values for FACE-BE-SELF: (a) image input to CNN, (b) top 10 source landmark features, (c) top 10 target landmark features, (d) plotted top 10 landmark features for source and target.

3.6 SUMMARY

In this chapter, a novel deep domain adaptative FACE-BE-SELF for concurrent learning of adult and child facial expressions is proposed. FACE-BE-SELF yields a meaningful and effective selection of features that are correlated with expressions. The explanation and visualization of SHAP values corroborate the facial expression classification performance of our method. The superiority of our method over existing transfer learning and domain adaption methods satisfies the need for a systematic feature selection, feature fusion, and domain adaptation to perform domain-invariant classification. In future work, we plan to investigate the generalizability of this approach to other age groups and data acquisition pipelines. We hope that this approach may be used to yield automated, objective assessments of age or domain varying patterns in other applications.

CHAPTER 4

CUSTOMIZABLE AVATARS WITH DYNAMIC FACIAL ACTION CODED EXPRESSIONS FOR IMPROVED USER ENGAGEMENT

4.1 CHAPTER OVERVIEW

Facial expressions of 3D avatars are often used as stimuli in studies of intervention efficacy or behavioral biomarker discovery [32, 69, 70, 73, 74]. Such studies incorporate tasks to elicit and measure constructs related to facial expressions. The typical setting involves eliciting a response using the 3D avatar-based stimuli, capturing the response with a sensor, and extracting relevant measurements from the raw sensor data. To capture perception and production of facial expressions, the applicable sensing modalities are ET and VT, respectively [32, 69, 70, 73, 74]. Measures such as the percentage duration of gaze fixations to AOIs within the stimuli have been used to study perception [81]. To assess production, FACS [79] provides a taxonomy of AUs that describe the individual constituent movements of the face. Machine and deep learning approaches may be used to detect AUs from video frames of the face [132, 171-177]. Finally, evaluating the construct validity of these tasks, i.e., whether the intended construct is elicited and measured, is an important precursor for well-designed studies of intervention efficacy or behavioral biomarker discovery [81, 178].

In the sections to follow, we review related work on important design considerations for 3D avatar-based facial expression stimuli, automatic detection of FACS AUs, and construct validity. Then, we propose 1) dynamic, FACS-labeled stimuli for perception and production of facial expressions, rendered on customizable 3D avatars, 2) a new deep learning-based AU detector for measurement of subjects' facial responses, and 3) construct validity of the proposed stimuli and measurements based on two tasks (recognition and mimicry) completed by 20 healthy adult volunteers.

4.1.1 RELATED WORK

This section discusses related work on design considerations for 3D avatar-based facial expression stimuli, automatic detection of FACS AUs, and construct validity.

4.1.1.1 DESIGN CONSIDERATIONS FOR 3D AVATAR-BASED FACIAL EXPRESSION STIMULI

Securing and maintaining user engagement is a key challenge for avatar-based health applications [60]. Recently, avatar customization has been identified as an effective means of improving engagement [60]. Avatar customization has been shown to increase engagement and enjoyment in social [78, 179], procedural [180], creative [180], and cognitive tasks [60], including interventions for physical [181, 182] and mental health [60, 179]. Avatar realism is another important factor influencing engagement. Hyper-realistic avatars may trigger the uncanny valley effect, a phenomenon where objects with increasingly realistic human appearances evoke uneasiness or revulsion, causing users to disengage [77]. Furthermore, several studies find that users prefer to interact with semi-realistic avatars [77, 183]. Avatars may embody a humanoid form to varying degrees from ‘talking heads’ to full body representations. Full body representations have been shown to improve dyadic interactions with avatars [184]. Facial expression stimuli may be rendered statically as still images or dynamically as animations from neutral to peak expression. While some studies [72, 185] continue to use static facial expressions due to accessibility of widely used, validated stimuli sets [186, 187], it has been established from both neuroimaging and behavioral perspectives that dynamic expressions are more salient than static expressions, and show increased activity in face processing regions of the brain [188]. Thus, dynamic facial expressions play a pivotal role in assessing relevant differences between control individuals and individuals with a diagnosis in biomarker discovery studies (e.g., depression [189], Moebius syndrome [190], ASD [188]).

To study the effect of 3D avatar-based stimuli on perception or production of facial expressions, it is important to ensure that the avatar accurately renders the target expressions by having the expressions evaluated and labeled with AUs by FACS experts. This labeling may be especially critical for studies of expression production, where the construct may be defined based on a one-to-one correspondence between the avatar's AUs and the participant's AUs. While several methods for transferring AUs to arbitrary avatar faces have emerged, e.g., [191, 192], these methods are not guaranteed to accurately reproduce the target AUs. Therefore, avatar models and avatar-generation platforms that have been evaluated by FACS experts, such as MiFace [193], HapFACS [194], FACSGen [195], FACSHuman [196], and García et al.'s avatars [197], are preferred. While all of these existing avatars and avatar-generation platforms support dynamic animations, they are limited in that they either lack customization capabilities [193, 197], rely on commercial software [194, 195], and/or are rendered as a disembodied floating head or face [193, 195, 196], which may break immersion. Additionally, García et al.'s avatars [197] are hyper-realistic, which may trigger the uncanny valley effect [77]. Given these limitations, there still exists a need for customizable, dynamic 3D avatars and avatar-based facial expression stimuli that have been evaluated and labeled with AUs by FACS experts.

4.1.1.2 AUTOMATIC DETECTION OF FACS AUS

While ET based measures of facial expression perception require only straightforward mathematical operations [81], automatic AU detection from images is more challenging. Facial expressions consist of multiple AUs occurring simultaneously in various localized areas of the face. Thus, AU detection is a multi-label problem, where each facial image is assigned one or more AUs. AU detection methods either train individual binary classifiers to detect the presence or absence of each AU or train a single model to detect multiple AUs at once [172-175, 177]. The latter approach, referred to as multi-label learning, is considered superior due to its computational efficiency and ability to take relationships between AUs into account [172-175, 177]. In addition to modeling the relationships between AUs, state-of-the-art multi-label learning approaches often incorporate methods for focusing on relevant areas of the input or features using saliency maps [177], attention [176], or patch/region learning [171, 175]. Multi-label learning approaches may also benefit from multi-task learning of other tasks related to the face (e.g., landmark prediction [176], facial expression classification [198], valence-arousal estimation [198]) and from feature fusion (e.g., saliency maps [177], geometric features [199]).

A drawback of state-of-the-art multi-label AU detection approaches is that they do not independently predict left and right activations of bilaterally located AUs, which may be useful for health applications. For example, Dell'Olio et al. [200] recently proposed FaraPy, an augmented reality mirror therapy for patients with facial paralysis. Asymmetrical AU activation is characteristic of facial palsy or paralysis, e.g., due to stroke, Parkinson's, Bell's Palsy, etc. [200], and has also been observed among individuals diagnosed with ASD [201-203].

Recently, Bar and Wells [152] present an approach based on convex geometry for identifying significant correlations among a large number of features using a mixture of beta distributions (betaMix). The betaMix approach relies upon Theorem 1.1 from [204], which shows that the sine squared of the angles between randomly drawn features in high dimensional space follows a beta distribution. In Chapter 3, we apply the betaMix approach [152] in the context of facial expression classification [87] for decomposing geometric landmark-based features (e.g., distances between pairs of landmarks) into sets associated with expressions, identity, and age groups (children and adults). As described in Section 3.2.4 [87], learning takes place in two separate steps. First, betaMix [152] is fit using the EM algorithm to learn correlations between already extracted landmark-based features and three factors (expressions, identity, and age groups). The resulting graph is used to select expression-correlated features that are invariant to age and identity. Then, in the second step, the betaMix-selected features are fused with deep learning-based features to fit the expression classifier. Given our success with facial expression classification, we anticipate that feature extraction, selection, and learning steps may be further optimized for AUs by training end-to-end with supervision from the AU labels. We hypothesize that the aforementioned Theorem 1.1 from [204] may be adapted into a loss function for simultaneous, end-to-end learning of correlations among AUs and features, while discouraging dependence on identity, which is not addressed by present multi-label AU detection approaches.

4.1.1.3 CONSTRUCT VALIDITY

Construct validity may be determined by assessing whether the expected response is elicited in a healthy control group. For example, this approach has been used to evaluate candidate ET biomarkers for ASD [81]. Since NT individuals are known to prefer to attend to faces when viewing social stimuli, Shic et al. [81] test for face preference (percentage of gaze duration to the face AOI vs. random gaze) among NT controls to determine construct validity.

4.1.2 CONTRIBUTIONS

To address the limitations of currently available 3D avatar-based facial expression stimuli, we propose Customizable Avatars with Dynamic Facial Action Coded Expressions (CADyFACE) for user engagement. To detect AUs elicited by CADyFACE, we propose a deep neural network for novel Beta-guided Correlation and Multi-task Expression learning (BeCoME-Net). We further conduct a pilot study to evaluate the construct validity of CADyFACE and BeCoME-Net AU measurements. Our contributions are as follows:

- CADyFACE incorporates six avatar models representing different genders and races with customizable hair color, eye color, skin tone, and clothing. For each CADyFACE model, six facial expressions (‘anger’, ‘disgust’, ‘fear’, ‘happy’, ‘sad’, and ‘surprise’) have been posed and labeled by a certified FACS expert with over 600 hours of coding experience.
- We propose a novel beta-guided correlation loss for BeCoME-Net that encourages features to be correlated with AUs while discouraging correlation with subject identity. For richer representation learning, BeCoME-Net fuses geometric landmarks and deep learning-based texture features while jointly learning AU detection and expression classification tasks. We consider variants of BeCoME-Net for bilateral and unilateral AU detection. We compare BeCoME-Net with state-of-the-art AU detection methods on two benchmark data sets.
- We conduct an online pilot study of 20 healthy adult participants to evaluate the construct

validity of the proposed CADyFACE stimuli and BeCoME-Net AU measurements. Participants complete two facial expression-related tasks, recognition and mimicry, while facial video and webcam-based eye-tracking data are collected.

The remainder of this chapter is organized as follows. Section 4.2 describes the proposed methods. Section 4.3 presents the results and discussion. Section 4.4 discusses limitations. Section 4.5 summarizes.

4.2 METHODS

This section describes the design and development of the proposed CADyFACE stimuli, BeCoME-Net for multi-label AU detection, and pilot study, tasks, and constructs.

4.2.1 DESIGN AND DEVELOPMENT OF CADYFACE STIMULI

The design and development of CADyFACE stimuli involves multiple steps including avatar generation, avatar customization, FACS-annotation of the avatars' facial expressions, dynamic animation of the facial expressions, and review by clinical team members.

4.2.1.1 AVATAR GENERATION

We generate 3D avatars for CADyFACE using free, open-source tools including the 3D modeling software Blender (https://www.blender.org/) and ManuelBastioniLAB 1.6.1a (https://github.com/animate1978/MB-Lab), a character creation plugin for Blender. ManuelBastioniLAB 1.6.1a includes six human prototypes: African female, African male, Asian female, Asian male, European female, and European male. We obtain one 3D avatar for each of these six prototypes using the default settings. Each avatar includes a face rig with 75 blendshapes for facial animation. We dress the avatars in pants, a shirt, and a jacket. Clothing assets are obtained from [192].

4.2.1.2 AVATAR CUSTOMIZATION

We develop the CADyFACE avatar customization application using the free Unity game engine (https://unity.com/). Users are shown their current avatar on the left side of the screen and a selection of customization options on the right side of the screen. Users navigate between screens of options using ‘next’ or ‘back’ buttons, which also update a progress bar. As users select different customization options, the updates are rendered on the avatar. An example customization screen is shown in Figure 15.

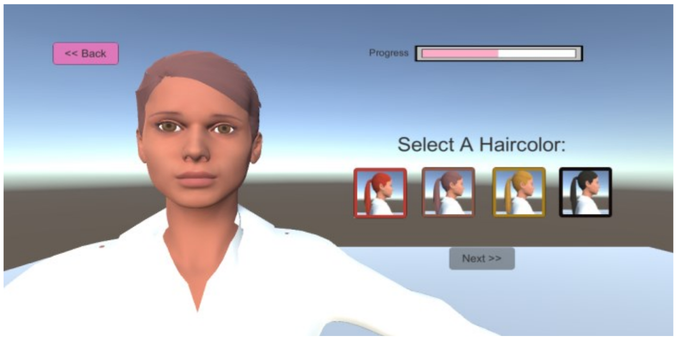


Figure 15. Example customization screen for hair color.

There are 49,152 different possible combinations based on the selection of one of each of the following: six different avatar models, three skin tones, four eye colors, four hair colors, eight jacket colors, eight shirt colors, and eight pants colors. All customization options are summarized in Figure 16.

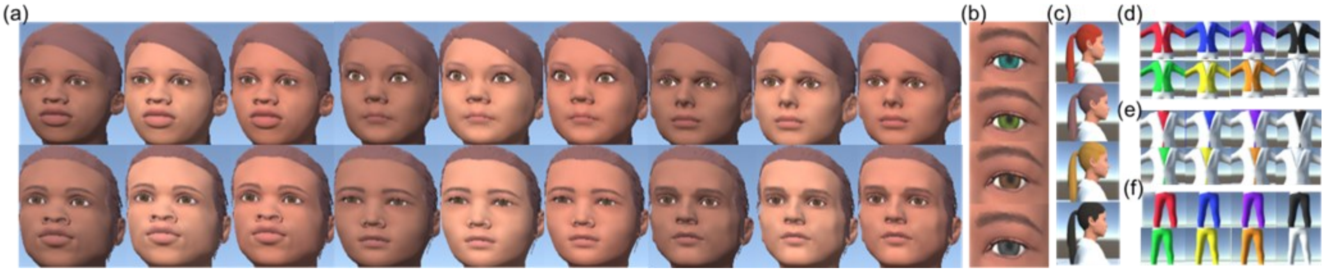


Figure 16. Avatar customization options: (a) all avatar model and skin tone combinations, (b) eye color options, (c) hair color options, (d) jacket color options, (e) shirt color options, and (f) pants color options.

4.2.1.3 FACS-ANNOTATED FACIAL EXPRESSIONS

Within Unity, we develop a software application to visualize expressions on each of the six prototype avatars and to adjust the appearance of each expression. Using this software, a member of our team who is a certified FACS expert with over 600 hours of coding experience has tuned the blendshapes for each of the six prototype avatars to render six different facial expressions (a total of 36 sets of 75 blendshapes). The AUs representing each expression are selected based upon their definitions in the FACS Investigator's Guide [79]. The specific AUs present in each expression and their intensities are reported in Table 4. Examples of each expression are shown in Figure 17.

Table 4. CADyFACE AU intensities (A= low to E= high)

Expressions	**Action Units**															
	1	**2**	**4**	**5**	**6**	**7**	**10**	**11**	**12**	**15**	**17**	**20**	**23**	**25**	**26**	**27**
Anger			E	E		C	D						D	B	C	
Disgust							E				D					
Fear	E	C	D	E								C		C		B
Happy					C				E							
Sad	E		D					B		E						
Surprise	D	D		B										C		C

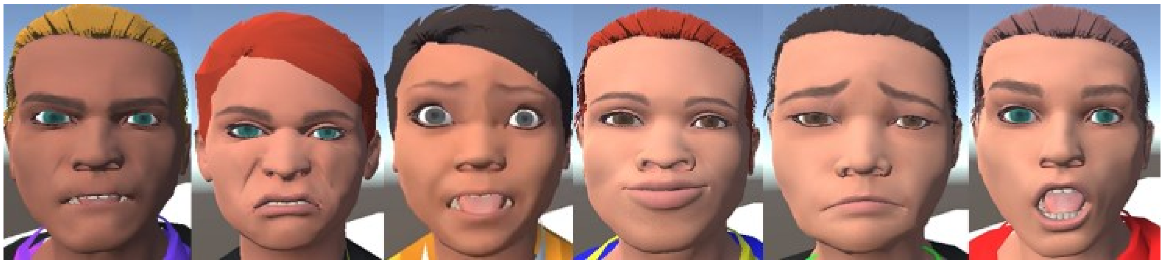


Figure 17. FACS-annotated expressions in CADyFACE (left to right): ‘anger’, ‘disgust’, ‘fear’, ‘happy’, ‘sad’, and ‘surprise’.

4.2.1.4 DYNAMIC ANIMATION OF FACIAL EXPRESSIONS

To generate the facial expression animations for CADyFACE, we linearly interpolate the blendshape values from ‘0’ (neutral) to the values associated with the AU labels defined for the target expression and avatar prototype. We animate each expression over 25 frames with a delay of 50 milliseconds between frames.

4.2.1.5 REVIEW BY CLINICAL TEAM MEMBERS

We have developed CADyFACE as a part of our Institutional Review Board (IRB)-approved study for behavioral biomarker discovery among children and young adults with ASD. Two team members who are clinicians with expertise in ASD have reviewed and provided feedback on CADyFACE throughout its development to ensure suitability for the study and appropriateness for individuals with ASD.

4.2.2 BECOME-NET FOR MULTI-LABEL AU DETECTION

This section describes the proposed BeCoME-Net including benchmark data sets, preprocessing, notation, bilateral and unilateral AU detection, backbone architecture, beta-guided correlation loss, multi-label learning framework, and experiments.

4.2.2.1 BENCHMARK DATA SETS

To train and evaluate BeCoME-Net for detecting the AUs present in CADyFACE, we consider the CK+ [137, 205] data set. The CK+ data set comprises 593 image sequences of 123 adult subjects ages 18 to 50 years posing facial expressions including ‘anger’, ‘disgust’, ‘fear’, ‘happy’, ‘sad’, and ‘surprise’. Each sequence begins with a neutral expression frame and ends with the peak expression frame, which has been annotated with AU labels. CK+ includes 30 different AUs, including the 16 AUs in CADyFACE: AUs 1, 2, 4, 5, 6, 7, 10, 11, 12, 15, 17, 20, 23, 25, 26, and 27. We refer to this subset of CK+ AUs as 16AU-CK+. We also use CK+ to compare BeCoME-Net with existing state-of-the-art approaches. However, since some of the AUs in CK+ appear with low frequency, existing state-of-the-art approaches report results on 12 or 13 AU subsets. We follow established literature [177] to define the 12 AU subset (12AU-

CK+) as AUs 1, 2, 4, 5, 6, 7, 9, 12, 17, 23, 24, and 25. Then, the 13 AU subset (13AU-CK+) is defined as 12AU-CK+ and AU 27 [177]. The frequencies of AUs present in at least one of these three subsets are reported in Table 5. Additionally, we follow the same procedure as [87] to obtain expression-labeled samples of CK+ to train the model to perform the expression classification task as a part of the proposed multi-task learning. The distribution of expression labels is 135 'anger', 177 'disgust', 75 'fear', 207 'happy', 327 'neutral', 84 'sad', and 249 'surprise' samples.

In addition to our primary data set, CK+, we also benchmark our approach on the Extended Denver Intensity of Spontaneous Facial Action (DISFA+) [206, 207] data set. DISFA+ consists of image sequences of 9 adult subjects posing 42 facial expressions including individual AUs, combinations of AUs, and the 6 prototypical expressions ('anger', 'disgust', 'fear', 'happy', 'sad', and 'surprise'). Each sequence begins with a neutral expression, moves to the peak expression, and ends with a neutral expression. All frames have been annotated for 12 different AUs: 1, 2, 4, 5, 6, 9, 12, 15, 17, 20, 25, and 26. The frequencies of these AUs are reported in Table 5. We also extract the samples with expression labels for use in multi-task learning, including 324 'anger', 2,779 'disgust', 1,276 'fear', 3,881 'happy', 24,793 'neutral', 436 'sad', and 1,345 'surprise' samples.

Table 5. Frequency of AUs in CK+ and DISFA+ data sets

AU	Description	CK+ Frequency	DISFA+ Frequency
1	Inner Brow Raiser	117	9353
2	Outer Brow Raiser	117	7982
4	Brow Lowerer	194	12036
5	Upper Lid Raiser	102	9208
6	Cheek Raiser	123	9839
7	Lid Tightener	121	--
9	Nose Wrinkler	75	3993
10	Upper Lip Raiser	21	--
11	Nasolabial Deepener	34	--
12	Lip Corner Puller	131	10371
15	Lip Corner Depressor	94	3956
17	Chin Raiser	202	5689
20	Lip Stretcher	79	4854
23	Lip Tightener	60	--
24	Lip Pressor	58	--
25	Lips Part	324	11442
26	Jaw Drop	50	--
27	Mouth Stretch	81	7487

4.2.2.2 PREPROCESSING

We follow the same preprocessing pipeline as in [87], which yields 256×256-pixel grayscale images of centered faces, rotated such that the eyes are level and the left eye is 30% of the image width from the left edge. The images are min-max normalized to the range [0,1]. For each image, we use the dlib library (http://dlib.net/) to extract 68 landmark points on the face and normalize x- and y-coordinates to [0,1].

4.2.2.3 NOTATION

BeCoME-Net is a deep learning model of the form $f\colon \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ is the input space of facial images and landmarks and $\mathcal{Y}$ is the output space of AU labels for the AU detection task and expression labels for the expression classification task. We define input $X = (X_{img}, X_{lmk}) \in \mathcal{X}$, where $X_{img} \in \mathcal{X}_{img} \subseteq \mathbb{R}^{m \times n}$ are $m \times n$ facial expression images and $X_{lmk} \in \mathcal{X}_{lmk} \subseteq \mathbb{R}^{l \times 2}$ are pairs of l x- and y- landmark coordinates. We define output $Y \in \mathcal{Y}$ separately for multi-task learning of AU detection and expression classification tasks. For the AU detection task, we define $Y = Y_{AU} \in \mathcal{Y}_{AU}$, where $\mathcal{Y}_{AU}$ represents the set of all binary label vectors indicating the presence or absence of K_{AU} (12, 13, or 16) different facial action units. For the expression classification task, we define $Y = Y_{EXPR} \in \mathcal{Y}_{EXPR}$ where $\mathcal{Y}_{EXPR}$ is the set of K_{EXPR}-dimensional one-hot encoded vectors representing the $K_{EXPR} = 7$ expression labels. We denote the vector for subject identity as g, which will be used in the design of the proposed beta-guided correlation loss to discourage feature correlations with identity. We partition f into backbone network M and task head H such that $f = H \circ M$, $M: X \rightarrow Z$, and $H: Z \rightarrow Y$, where Z is a latent space of p features.

4.2.2.4 BILATERAL AND UNILATERAL AU DETECTION

We define two variants of BeCoME-Net with different input shapes for bilateral and unilateral detection of AUs. BeCoME-Net-F is designed to process $(m \times n = 256 \times 256)$-pixel grayscale images and $l = 68$ landmark points extracted from the full facial image for bilateral AU detection. For unilateral AU detection, we predict AUs on the left and right sides of the face independently and define BeCoME-Net-H for $(m \times n = 256 \times 128)$-pixel grayscale images of the left or right side of the face and $l = 39$ landmark points (29 from the same side of the face and 10 located along the center line of the face).

4.2.2.5 BACKBONE ARCHITECTURE

The architecture for BeCoME-Net begins with a backbone $M(\cdot)$, consisting of convolutional, pooling, and fully connected layers for feature extraction. Figure 18 presents the backbone architecture for BeCoME-Net-F. The backbone incorporates two branches for processing images X_{img} and landmarks X_{lmk}, respectively. For the image branch, we consider the same model architecture as in Chapter 3 [87]: three blocks of a 2D convolutional layer with 3x3 kernel size followed by 2x2 maximum pooling yielding 16, 32, and 64 feature maps, respectively, and a final fully connected layer of 512 hidden units. Convolutional and fully connected layers use the ReLU activation function. Dropout is applied with a probability of 0.5 at the final fully connected layer. For the landmark branch, we input the x, y-coordinates of the l landmark points directly into a 1D convolutional layer with a kernel size of 1 to yield 16 feature maps, which are flattened prior to a final 512-unit fully connected layer. We use ReLU in the convolutional and fully connected layers and apply dropout with a probability of 0.5 at the fully connected layer. Compared to Chapter 3 [87], in which we perform feature engineering and selection based on the landmarks prior to learning, the 1D convolutional layer with kernel size 1 serves to aggregate the 2D coordinate information so that the network may learn relevant features from the normalized landmark positions directly. We concatenate the outputs of the image and landmark branches to form $(p = 1024)$-dimensional feature vector Z.

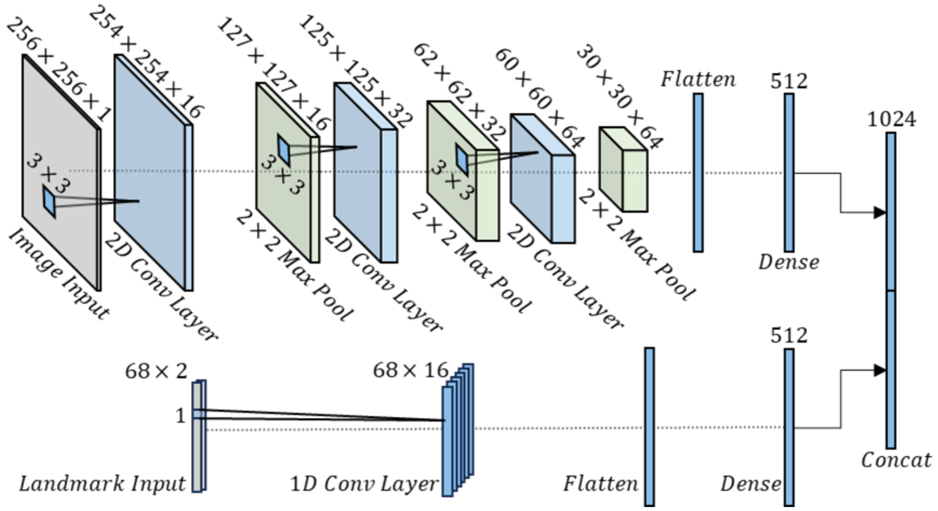


Figure 18. BeCoME-Net-F backbone architecture.

4.2.2.6 BETA-GUIDED CORRELATION LOSS

We are interested in modeling significant correlations between the features in Z, labels in Y, and subject identity g during training. Let b represent the batch size. Consider the space $\mathbb{R}^b$. From [87, 152, 204], the sine squared of the angle θ between two random lines drawn from $\mathbb{R}^b$ follows the beta distribution:

$$\lambda \overset{\text{def}}{=} \sin^2\theta \sim beta\left(\frac{b-1}{2},\frac{1}{2}\right) \tag{35}$$

Random or 'null' pairs will be approximately perpendicular for even moderate values of b, e.g., $b = 10$, meaning that the probability of two random lines being correlated by chance is very small [152]. This result may be used to build a graphical model where the nodes are features, labels, or identity, and the edges represent significant correlations. We denote the number of nodes as w. For the AU detection task, $w\ nodes = p\ features + K_{AU}\ AUs +$

$1\ for\ subject\ identity$. For the expression classification task, $w\ nodes = p\ features + K_{EXPR}\ expressions + 1\ for\ subject\ identity$.

To build the graph, we employ a frequentist inferential procedure to screen for edges ('non-null' pairs or significant correlations) among the features in Z, labels in Y, and subject identity g. We denote the η-quantile of $beta\left(\frac{b-1}{2}, \frac{1}{2}\right)$ as Q_η. Pairs λ_e (e.g., feature-feature, feature-label, feature-identity) are considered significantly correlated if $\lambda_e = \sin^2 \theta_e < Q_\eta$, where $e = 0,1, \dots, t$ and the total number of possible edges $t = 0.5w(w-1)$. The selection of η may be used to control the Type I error rate. For each possible edge λ_e, we consider the null hypothesis $H_0: \lambda_e \geq Q_\eta$ (i.e., no edge) and the alternative $H_a: \lambda_e < Q_\eta$ (i.e., edge in the graph). We conduct a total of t individual hypothesis tests to determine the presence/absence of all possible edges. Using the Bonferroni correction, we divide $\alpha = 0.05$ by the total number of hypothesis tests t to set $\eta = \frac{\alpha}{t}$. The screening rules associated with the null and alternative hypotheses may be implemented using mirrored and translated Heaviside functions:

$$Heaviside(Q_\eta - \lambda_e) = \begin{cases} 0, & \lambda_e > Q_\eta \\ 1, & \lambda_e < Q_\eta \end{cases}. \tag{36}$$

However, due to the discontinuity at $\lambda_e = Q_\eta$, equation (30) is not differentiable. Sigmoid functions may be used to provide a smooth approximation for the Heaviside functions [208]. Therefore, we consider the following sigmoid function for differentiable implementation of the screening rules:

$$\sigma(\lambda_e) = 1 - \frac{1}{1+e^{-\gamma(\lambda_e - Q_\eta)}}, \tag{37}$$

where γ adjusts the sharpness of the transition from 1 to 0 at Q_η.

To construct the predicted graph adjacency matrix A, we apply (3) for each λ_e to yield the edge connection between each e^{th} pair of nodes (features, labels, or identity) and assign these to the upper triangle of A in row-major order. We fill the diagonal (representing self-connection) with 1's. The lower triangle of A is the upper triangle mirrored over the diagonal. We construct A such that the first p rows and columns represent the p features. The next K_{AU} or K_{EXPR} rows and columns represent the AU or expression labels, respectively. The last row and column represent identity. Then, we propose the beta-guided correlation loss $\mathcal{L}_{BGC}$ as:

$$\mathcal{L}_{BGC}(A) = \frac{1}{w^2}\sum_{i}^{w}\sum_{j}^{w}(S_{ij} \cdot A_{ij}), \tag{38}$$

where S is a $w \times w$ sign matrix (consisting of -1's, 0's, and 1's) that we use to encourage features to be correlated with the labels, discourage feature correlations with subject identity, and encourage feature diversity by discouraging correlations among the features themselves. We set the diagonal of S to 0's as self-connection will be unchanging and have no impact on the loss. Similarly, labels and identity will not be updated during learning. Only the features will be affected by the gradient updates. Therefore, we multiply the entries of A associated with label-label and label-identity pairs by 0's in S so that they do not contribute to the loss. Since we minimize the loss during learning, rows and columns representing edges between the labels and features are multiplied by -1 to maximize feature correlations with the labels. The remaining entries of S are filled with 1's to discourage correlations with subject identity and among the features. The entries are multiplied by the corresponding entries of A. To aggregate the individual loss contributions into a single number, we sum over all entries. Then, we divide by the total number of entries w^2 so that the scale of the loss does not change for different numbers of features or labels.

Equation (38) bears some similarity to reinforcement learning. Considering the policy gradient theorem without discounting [209], the loss at each time step is defined as the immediate reward times the predicted action (e.g., the one-hot encoding for the current action times the log of predicted probabilities for each action). Analogously, our A_{ij}'s encode the predicted presence or absence of an edge in the graph and the S_{ij}'s encode the associated rewards. However, rather than simply using cross-entropy for edge predictions, our equation (37) has several advantages. Figure 19 shows three key regions of equation (37). A particular λ_e will fall within region (a) if the associated pair of nodes is significantly correlated and will receive the full reward (or penalty) based on the associated S_{ij}. For example, a λ_e representing a feature-identity pair with $\sigma(\lambda_e) \approx 1$ will receive a penalty ≈ 1, while a λ_e representing a feature-label pair with $\sigma(\lambda_e) \approx 1$ will receive a reward ≈ 1 (penalty ≈ -1). Region (b) represents the boundary between significantly correlated and uncorrelated nodes. For λ_e's falling within region (b), the reward will be weighted by $\sigma(\lambda_e)$. Finally, region (c) represents uncorrelated nodes where $\sigma(\lambda_e) \approx 0$, so uncorrelated nodes will have a very small contribution to the loss. Due to equation (37), $\mathcal{L}_{BGC}$ focuses more on significantly correlated pairs while ignoring uncorrelated pairs. Therefore, individual features are allowed to specialize, i.e., to be highly correlated with one or several specific AUs (or expressions), without being penalized for having low correlation with other AUs (or expressions). This property is especially suitable for AU and expression learning as many of the constituent muscle actions of the face cannot or rarely occur concurrently (e.g., AU 24 ‘lip pressor’ and AU 25 ‘lips part’ cannot occur together) while others often occur together (e.g., AU 1 ‘inner brow raiser’ and AU 2 ‘outer brow raiser’).

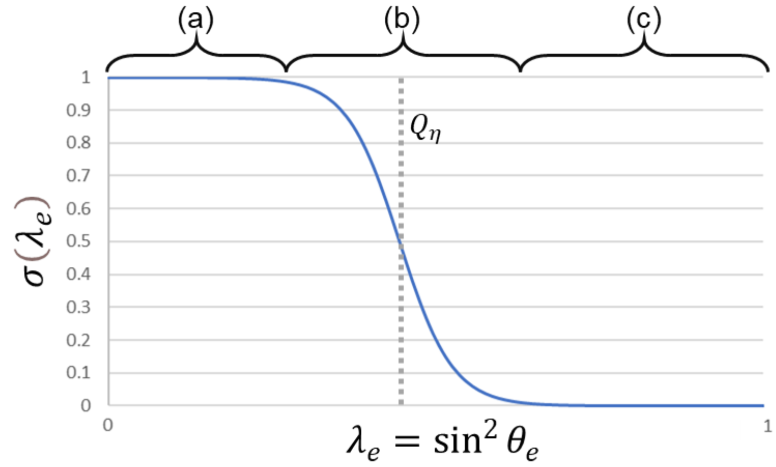


Figure 19. Plot of equation (37) showing regions where (a) there is a significant correlation between nodes, (b) transition between significantly correlated and uncorrelated nodes, and (c) nodes are uncorrelated.

4.2.2.7 MULTI-TASK LEARNING FRAMEWORK

While our primary goal is AU detection, training BeCoME-Net to perform the related task of expression classification is expected to improve representation learning and AU detection performance. We define the multi-label AU detection head as a fully connected output layer of K_{AU} units with sigmoid activation, where K_{AU} is the number of target AUs. We define the expression classification head as a fully connected softmax output layer with K_{EXPR} units for the K_{EXPR} facial expression classes. For efficient learning of both tasks, we duplicate the backbone and connect one head to each copy, as shown in Figure 20. Weights are shared between the two copies of the backbone for simultaneous training on both AU detection and expression classification tasks. We supervise the learning of the AU detection and expression classification tasks with the weighted multi-label cross-entropy loss $\mathcal{L}_{WMCE}$ [210] and weighted categorical cross-entropy loss $\mathcal{L}_{WCCE}$ [211], respectively. We choose the weighted variants of both losses to address imbalance in the label distributions. We use the beta-guided correlation loss $\mathcal{L}_{BGC}$ in both tasks ($\mathcal{L}_{BGC\_AU}$ for AU detection and $\mathcal{L}_{BGC\_EXPR}$ for expression classification) to encourage features to be correlated with the labels while discouraging correlation with subject identity. The loss for the AU detection task is $\mathcal{L}_{AU} = \mathcal{L}_{WMCE} + \mathcal{L}_{BGC\_AU}$, and $\mathcal{L}_{EXPR} = \mathcal{L}_{WCCE} + \mathcal{L}_{BGC\_EXPR}$ for the expression classification task. The overall loss is $\mathcal{L} = \mathcal{L}_{AU} + \mathcal{L}_{EXPR}$.

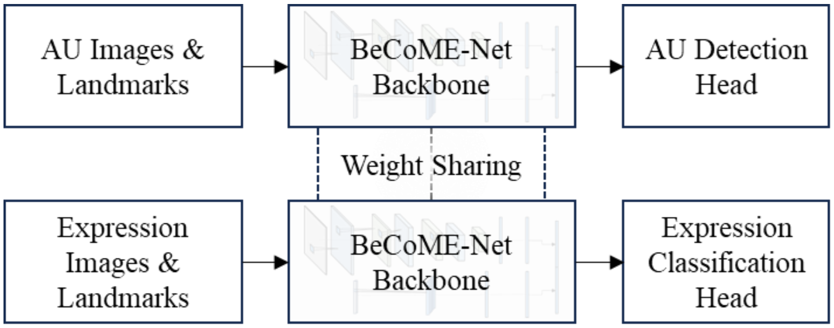


Figure 20. BeCoME-Net multi-task learning framework.

4.2.2.8 EXPERIMENTS

Following established literature [177], we perform 3-fold subject-independent cross-validation for all experiments and report F1 scores calculated over all test folds. To study the effect of multi-task learning and the proposed beta-guided correlation loss, we perform an ablation study on 16AU-CK+ for both bilateral and unilateral AU detection models. Then, we benchmark BeCoME-Net-F and BeCoME-Net-H on 16AU-CK+ and report the performance for each AU. Next, we train and test BeCoME-Net-F and BeCoME-Net-H on 13AU-CK+, 12AU-CK+, and DISFA+ for comparison with state-of-the-art approaches. For 13AU-CK+, we compare with BGCS [174], HRBM [172], and LNDSM [177]. For 12AU-CK+, we compare with JPML [173], DSCMR [132], and LNDSM [177]. For DISFA+, we compare with DRML [171], AU R-CNN [175], JÂA-Net [176], and LNDSM [177]. LNDSM [177] is the newest state-of-the-art approach with which we compare.

For all experiments, we train using the ADAM optimizer with a triangular learning rate policy [167] cycling the learning rate between 10^{-5} and 10^{-3} until convergence. We use a batch size of $b = 32$ and set γ =100.

4.2.3 PILOT STUDY, TASKS, AND CONSTRUCTS

This section describes IRB approval, participants, online stimuli presentation and data collection, tasks, and constructs.

4.2.3.1 IRB APPROVAL

This pilot study has been conducted as a part of our larger IRB-approved study to discover behavioral biomarkers for children and young adults with ASD and has been approved by the IRBs at Old Dominion University (Application No. 1424272) and Eastern Virginia Medical School (Application No. 19-06-EX-0152). Approval letters from ODU and EVMS IRBs may be found in Appendices B and C, respectively. All participants have provided informed consent and have not received compensation for their participation.

4.2.3.2 PARTICIPANTS

We have recruited 20 healthy adult volunteers (ages 21 to 35, 4 female) for an online pilot study of CADyFACE. Inclusion criteria includes being at least 20 years of age at the time of enrollment and having access to an Internet-connected personal computer with a webcam. For privacy, each participant has been assigned a unique subject identifier.

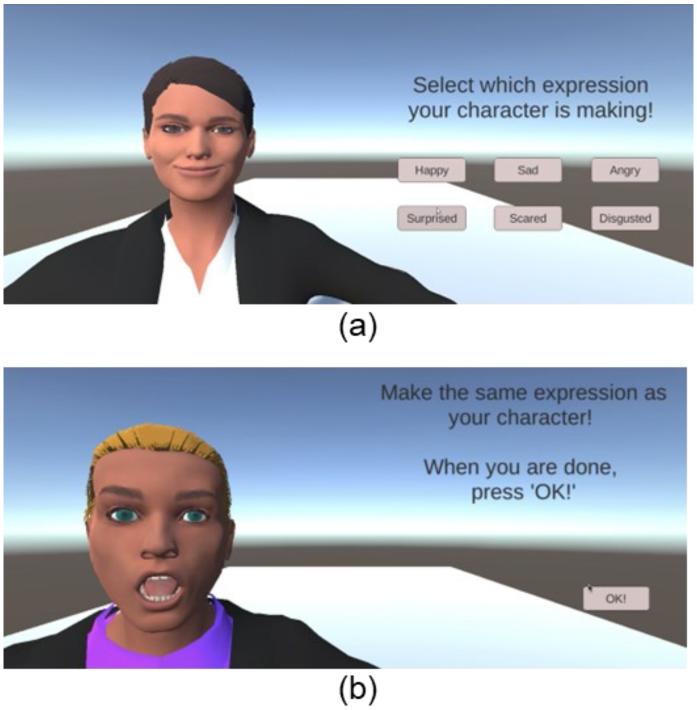
4.2.3.3 ONLINE STIMULI PRESENTATION AND DATA COLLECTION

A Unity Web-GL application has been developed to present the CADyFACE stimuli in each participant's web browser. The application is embedded into a webpage hosted on the visionlab.odu.edu domain, which is served by a secure web server located at Old Dominion University. Participants must enter their unique subject identifier to access the webpage.

As the participants interact with the Unity Web-GL application, the WebGazer.js (https://webgazer.cs.brown.edu/) [212] JavaScript library and its self-calibrating ET model are used to collect VT and webcam-based ET coordinates from the participants' webcams. As these VT frames and ET coordinates are collected, they are recorded to the secure web server.

4.2.3.4 TASKS

Participants complete two tasks developed using the CADyFACE stimuli: recognition and mimicry. In the recognition task, participants are asked to select the expression shown by clicking the button labeled with the name of the expression. In the mimicry task, participants are asked to make the same facial expression as the avatar. Each task consists of six trials, one for each of the six FACS-annotated expressions in CADyFACE ('anger', 'disgust', 'fear', 'happy', 'sad', or 'surprise'). Each participant customizes their own avatar that is used in both tasks. The expression and mimicry tasks are shown in Figure 21.


Select which expression your character is making!

Make the same expression as your character!

When you are done, press 'OK!'

Figure 21. (a) Recognition and (b) mimicry tasks.

4.2.3.5 CONSTRUCTS

During the recognition task, participants are expected to attend their gaze to the avatar's face to determine the expression. Therefore, we consider the face preference construct [81]. Following [81], we measure the construct as the percentage of gaze duration to the avatar's face (%Gaze Face) and test construct validity using a one-sample t-test of %Gaze Face against the percentage of the scene taken up by the avatar's face, which is the expected %Gaze Face given random gaze. For our recognition task, the avatar's face occupies 15.0% of the scene. We test the construct validity of all six expressions.

During the mimicry task, participants are expected to pose the same facial expression as the avatar. Since CADyFACE has AU labels, we consider constructs based upon the activation of the same AUs by the participants. To measure the constructs, we use BeCoME-Net-F and

BeCoME-Net-H to detect the AUs in peak expression frames of each of the participants' mimicked expressions. We use both BeCoME-Net-F and BeCoME-Net-H to measure the construct with both bilateral and unilateral AU detectors. For each AU, we test construct validity using a one-sample t-test against 0 (no activation). We test the construct validity of all AUs in all six expressions.

4.3 RESULTS AND DISCUSSION

This section presents and discusses the ablation study, BeCoME-Net performance on the AUs in CADyFACE, comparison of BeCoME-Net with state-of-the-art AU detectors, and construct validity for BeCoME-Net AU measurements based on the pilot study group's responses to the CADyFACE stimuli.

4.3.1 ABLATION STUDY

To understand the impact of multi-task learning and the beta-guided correlation loss, we perform an ablation study for both bilateral and unilateral AU detection on 16AU-CK+. As shown in Table 6, the best performance is achieved for bilateral and unilateral models when both multi-task learning and the beta-guided correlation loss are considered. The inclusion of multi-task learning or the beta-guided correlation loss alone result in small improvements in mean F1 score (less than 1%) for bilateral and unilateral models. Including both multi-task learning and beta-guided correlation achieves an improvement of 1.81% and 2.86% in mean F1 score for bilateral and unilateral models, respectively. These results suggest that the use of the beta-guided correlation loss in the secondary expression classification task yields better representation learning for AU detection. These results also show that all bilateral models perform better than their unilateral counterparts, which we expect is due to the bilateral models having access to information from the entire face.

Table 6. Ablation study

AU Detection	**Input Size (image, landmarks)**	**Multi-Task Learning**	**Beta-Guided Correlation Loss**	**Mean F1 Score**
Unilateral	256x128, 39			61.16%
Unilateral	256x128, 39	X		61.83%
Unilateral	256x128, 39		X	61.77%
Unilateral	256x128, 39	***X***	***X***	***64.02%***
Bilateral	256x256, 68			64.51%
Bilateral	256x256, 68	X		64.78%
Bilateral	256x256, 68		X	65.16%
Bilateral	256x256, 68	***X***	***X***	***66.32%***

4.3.2 BECOME-NET PERFORMANCE FOR CADYFACE AUS

The precision, recall, and F1 scores for each AU based on 3-fold cross-validation of 16AU-CK+ for BeCoME-Net-F and BeCoME-Net-H are reported in Figure 22. FIGURE 23 reports precision, recall, and F1 scores based on the expression classification task. Both models follow similar patterns of performance. The best performing AUs with the F1 scores of over 80% are AUs 2, 12, 17, 25, and 27. As shown in Table 5, AUs 2, 12, 17, and 25 are some of the most frequent AUs in 16AU-CK+. While being a less frequent AU, AU 27 (mouth stretch) is associated with the distinctive open mouth appearance seen in the fear and surprise expressions. The worst performing AUs with F1 scores less than 50% are the four least frequent AUs in 16AU-CK+: AUs 10, 11, 23, and 26.

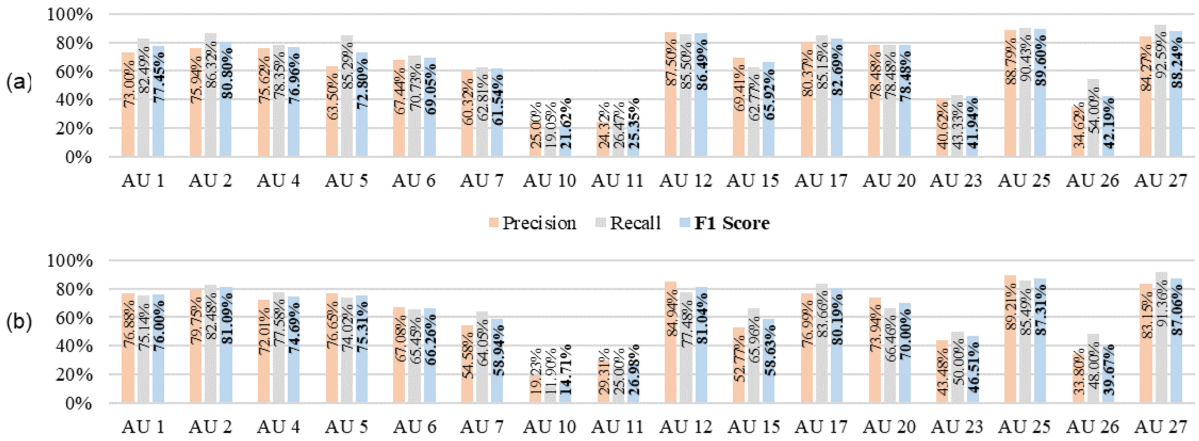


Figure 22. AU detection precision, recall, and F1 scores based on 3-fold cross-validation of 16AU-CK+ for (a) BeCoME-Net-F and (b) BeCoME-Net-H.

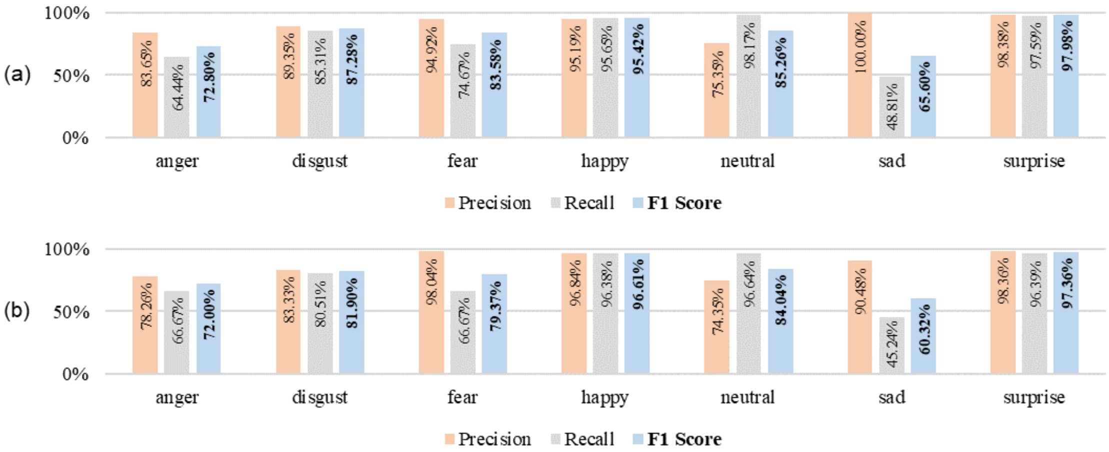


Figure 23. Expression classification precision, recall, and F1 scores based on 3-fold cross-validation of 16AU-CK+ for (a) BeCoME-Net-F and (b) BeCoME-Net-H.

4.3.3 COMPARISON WITH STATE-OF-THE-ART AU DETECTORS

We compare our proposed BeCoME-Net with state-of-the-art approaches for multi-label AU detection using the CK+ (AU13-CK+ and AU12-CK+) and DISFA+ data sets. Table 7 shows that BeCoME-Net-F achieves performance on par with the best performing and most recent state-of-the-art method LNDSM [177] for AU13-CK+. BeCoME-Net-H performs second best after BeCoME-Net-F and LNDSM. For AU12-CK+, BeCoME-Net-F achieves the highest performance, slightly outperforming LNDSM while BeCoME-Net-H performs equally well to LNDSM. We note that for each prediction, LNDSM requires a reference image of the neutral face for the same subject. Our proposed BeCoME-Net performs competitively on CK+ without requiring reference images of the neutral face.

Table 7. F1 scores for BeCoME-Net compared with state-of-the-art approaches for multi-label AU Detection on CK+

Method	**AU**													**Mean**
	1	**2**	**4**	**5**	**6**	**7**	**9**	**12**	**17**	**23**	**24**	**25**	**27**	
13 AU Detection														
BGCS	0.71	0.68	0.64	0.60	0.58	0.47	0.52	0.73	0.77	0.22	0.24	0.79	0.58	0.58
HRBM	**[0.87]**	**0.86**	0.73	0.72	0.62	0.55	0.86	0.73	0.82	**[0.57]**	0.35	**0.93**	0.88	0.73
LNDSM	**0.86**	**[0.88]**	**[0.81]**	0.74	**[0.70]**	**0.62**	0.89	**[0.87]**	**[0.86]**	0.46	0.46	**[0.94]**	**0.90**	**[0.77]**
BeCoME-Net-F	*0.82*	*0.85*	***0.80***	***[0.76]***	***0.66***	***[0.63]***	***[0.93]***	*0.82*	***0.83***	*0.45*	***[0.58]***	*0.91*	***[0.91]***	***[0.77]***
BeCoME-Net-H	*0.81*	*0.82*	*0.76*	***0.75***	***0.66***	*0.58*	***0.90***	***0.83***	***0.83***	***0.49***	***[0.58]***	*0.90*	***0.90***	***0.76***
12 AU Detection														
JPML	0.50	0.40	0.72	0.53	0.58	0.24	0.55	0.75	0.82	0.42	0.31	0.76	--	0.55
DSCMR	0.54	0.64	0.61	0.42	**[0.68]**	0.36	0.54	0.80	**[0.90]**	**[0.75]**	0.36	0.86	--	0.62
LNDSM	**[0.88]**	**[0.86]**	**[0.82]**	0.75	**[0.68]**	0.56	0.90	**[0.87]**	**0.85**	0.33	0.43	**[0.91]**	--	**0.74**
BeCoME-Net-F	*0.80*	***0.83***	***0.80***	***0.76***	***0.67***	***[0.61]***	***[0.94]***	***0.85***	*0.82*	***0.46***	***[0.59]***	***0.90***	--	***[0.75]***
BeCoME-Net-H	***0.81***	***0.83***	*0.78*	***[0.78]***	*0.65*	***0.60***	***0.91***	*0.82*	*0.81*	*0.44*	***0.54***	***0.90***	--	***0.74***

*Bold with brackets indicates the best score. Bold without brackets indicates the second-best score.

To show how BeCoME-Net performs on a data set other than CK+, we compare our performance on the DISFA+ data set in Table 8. While BeCoME-Net-F and BeCoME-Net-H both outperform DRML [171] and AU R-CNN [175], the best performance is achieved by LNDSM [177] followed by JÂA- Net [176]. Both LNDSM and JÂA-Net methods exhibit greater complexity and depth than ours. LNDSM is twice as deep as BeCoME-Net with six convolutional blocks compared to our three [177]. LNDSM also benefits from using neutral reference images to generate saliency maps that are fused at several intermediate layers of the network [177]. JÂA-Net involves multiple sub-networks for face alignment, global feature learning, local AU feature learning, and attention refinement [176]. Furthermore, BeCoME-Net may be less competitive on DISFA+ due to the beta-guided correlation loss, which discourages correlation between the learned features and subject identities. Given that DISFA+ contains only 9 subjects (compared to 123 in CK+), some discriminative features may be spuriously correlated with subject identity.

Table 8. F1 scores for BeCoME-NET compared with state-of-the-art approaches for multi-label AU detection on DISFA+

Method	**AU**												**Mean**
	1	**2**	**4**	**5**	**6**	**9**	**12**	**15**	**17**	**20**	**25**	**26**	
DRML	0.27	0.22	0.51	0.36	0.56	0.32	0.39	0.23	0.27	0.16	0.57	0.42	0.36
AU R-CNN	0.48	0.43	0.56	0.48	0.43	0.24	0.47	0.24	0.06	0.29	0.53	0.39	0.38
JÂA-Net	0.84	0.81	0.79	0.78	0.78	0.68	0.85	0.55	0.60	0.49	0.85	0.69	0.73
LNDSM	0.83	0.80	0.78	0.74	0.82	0.74	0.84	0.56	0.65	0.50	0.88	0.77	0.74
BeCoME-Net-F	*0.75*	*0.71*	*0.70*	*0.68*	*0.72*	*0.64*	*0.81*	*0.53*	*0.60*	*0.41*	*0.74*	*0.60*	*0.66*
BeCoME-Net-H	*0.73*	*0.74*	*0.67*	*0.69*	*0.72*	*0.61*	*0.73*	*0.39*	*0.49*	*0.37*	*0.69*	*0.55*	*0.61*

4.3.4 CONSTRUCT VALIDITY

As shown in Table 9, the recognition task's face preference construct is valid for all expressions. Two participants are excluded due to ET track loss. Table 10 reports construct validity for the mimicry task based on BeCoME-Net-F and BeCoME-Net-H AU predictions. For some of the tests (indicated by an asterisk*), we are unable to compute a t-statistic due to there being no predictions of the AU among any of the participants. These AUs are AU 10, AU 11, AU 23, and AU 26, which are the four least frequent AUs in the 16AU-CK+ training set. The following unilateral and bilateral constructs are valid: AUs 4 and 25 for ‘anger’; AU 17 for ‘disgust’; AUs 1, 2, 5, 25, and 27 for ‘fear’; AU 12 for ‘happy’; all AUs (1, 4, 11, 15) except AU 11 for ‘sad’; and all AUs (1, 2, 5, 25, 27) for ‘surprise’. Failing to pass the test of construct validity for some AUs may be attributed to one of two reasons: BeCoME-Net did not detect a present AU or the CADyFACE stimuli did not successfully elicit the AU.

4.4 LIMITATIONS

As with other 3D avatar models, the ManuelBastioniLAB models that we use in this work are limited by the fidelity and quality of their blendshapes. While AUs 9 and 10 are both listed as potential core components of a prototypical disgust face in the FACS Investigator’s Guide [79], AU 9 (‘nose wrinkler’) is more common, as reflected by the relative frequencies in the CK+ data set. However, since the ManuelBastioniLAB models are unable to perform nose wrinkling, we opt for AU 10. Using AU 9 instead of AU 10 may have yielded better results for the construct validity of the disgust expression. Furthermore, for AU 11 (‘nasolabial deepener’), the ManuelBastioniLAB models are only able to render a low level of activation. More conspicuous representation of AU 11 may have had a positive impact on the construct validity of AU 11 within the sad expression.

Table 9. Construct validity for recognition task

Expression	Construct	%Gaze Face			
		df	t-statistic	p-value	validity
Anger	Face Preference	17	3.523	0.001	✓
Disgust	Face Preference	17	2.291	0.018	✓
Fear	Face Preference	17	3.320	0.002	✓
Happy	Face Preference	17	3.123	0.003	✓
Sad	Face Preference	17	3.113	0.003	✓
Surprise	Face Preference	17	2.566	0.010	✓

Table 10. Construct validity for mimicry task

Expression	Construct	BeCoME-Net-F (Bilateral)				BeCoME-Net-H (Unilateral)			
		df	t-statistic	p-value	validity	df	t-statistic	p-value	validity
Anger	AU 4 Activation	19	3.199	0.005	✓	19	4.359	<0.001	✓
	AU 5 Activation	19	1.831	0.082		19	1.831	0.083	
	AU 7 Activation	19	1.453	0.163		19	3.199	0.005	✓
	AU 10 Activation	19	--*	--		19	2.517	0.021	✓
	AU 23 Activation	19	--*	--		19	--*	--	
	AU 25 Activation	19	5.339	<0.001	✓	19	3.943	<0.001	✓
	AU 26 Activation	19	--*	--		19	1.453	0.163	
Disgust	AU 10 Activation	19	1.000	0.330		19	1.453	0.163	
	AU 17 Activation	19	4.819	<0.001	✓	19	2.854	0.010	✓
Fear	AU 1 Activation	19	7.550	<0.001	✓	19	8.718	<0.001	✓
	AU 2 Activation	19	13.077	<0.001	✓	19	8.718	<0.001	✓
	AU 4 Activation	19	1.831	0.083		19	2.517	0.021	✓
	AU 5 Activation	19	3.943	<0.001	✓	19	3.943	<0.001	✓
	AU 20 Activation	19	1.000	0.330		19	1.831	0.083	
	AU 25 Activation	19	10.376	<0.001	✓	19	8.718	<0.001	✓
	AU 27 Activation	19	3.199	0.005	✓	19	2.854	0.010	✓
Happy	AU 6 Activation	19	2.179	0.042	✓	19	1.000	0.330	
	AU 12 Activation	19	2.854	0.010	✓	19	2.854	0.010	✓
Sad	AU 1 Activation	19	3.943	<0.001	✓	19	4.818	<0.001	✓
	AU 4 Activation	19	4.359	<0.001	✓	19	5.338	<0.001	✓
	AU 11 Activation	19	--*	--		19	--*	--	
	AU 15 Activation	19	5.339	<0.001	✓	19	2.517	0.021	✓
Surprise	AU 1 Activation	19	4.359	<0.001	✓	19	5.339	<0.001	✓
	AU 2 Activation	19	4.819	<0.001	✓	19	5.940	<0.001	✓
	AU 5 Activation	19	2.854	0.010	✓	19	3.943	<0.001	✓
	AU 25 Activation	19	7.550	<0.001	✓	19	10.376	<0.001	✓
	AU 27 Activation	19	3.560	0.002	✓	19	3.199	0.005	✓

*AU not detected

4.5 SUMMARY

This chapter proposes the CADyFACE stimuli, customizable 3D avatars with FACS labels for improved user engagement. Additionally, we propose BeCoME-Net for multi-label AU detection of AUs elicited by CADyFACE. We conduct an online feasibility study with 20 adult volunteers who complete recognition and mimicry tasks based on CADyFACE while their expressions and eye-gaze are recorded. We report the construct validity of these tasks using a well-known eye-tracking measure and the BeCoME-Net AU predictions. In the next chapter, CADyFACE and BeCoME-Net are used in a study aimed at discovering behavioral biomarkers for children and young adults diagnosed with ASD.

CHAPTER 5
PILOT STUDY TO DISCOVER CANDIDATE BIOMARKERS FOR AUTISM BASED ON PERCEPTION AND PRODUCTION OF FACIAL EXPRESSIONS

5.1 CHAPTER OVERVIEW

Identifying stratification biomarkers based on perception and production of facial expressions may help explain the heterogeneity of ASD with aims to improve patient care and individualize treatment strategies [14]. Behavioral VT and ET enable low-cost, unobtrusive acquisition of quantitative behavioral measurements of the face and eyes in naturalistic settings [18, 80], which may be subsequently studied using robust statistical tools for biomarker identification [213-215]. Online research studies may play a major role in ASD research by providing online access to large numbers of participants and longitudinal follow up [216]. The design of pilot studies compatible with the online model is an important step towards scalability for future large-scale replication and validation of research findings.

In the next section, we discuss related work on facial expression perception and production in ASD within the context of biomarker discovery as well as online studies and associated challenges. Then, we describe the contributions of this chapter, including the online pilot study of facial expression perception and production among ASD and NT groups of participants; assessment of VT and ET derived DVs based on ABC-CT criteria [18] of construct validity and group discriminability; and identification of one candidate biomarker and several more DVs of interest for future study.

5.1.1 RELATED WORK

The role of ASD in the perception and production of facial expressions has been studied for decades with conflicting findings, in part due to high heterogeneity in the responses of individuals diagnosed with ASD. The majority of studies report that autistic individuals have greater difficulty (e.g., lower accuracy, higher electroencephalography (EEG) N170 latency) compared to NT controls when completing facial expression recognition tasks, while others report no group differences [13]. Production of facial expressions has been studied via one of two modes of elicitation: 1) spontaneity, i.e., natural elicitation of expressions, or 2) mimicry, i.e., imitation of expressions. Spontaneous expressions of autistic individuals have been reported to be less frequent, shorter in duration, and lower in quality as rated by NT observers [13]. The intensity of spontaneous expressions has been reported to be the same [13, 217], lesser [23, 82], or greater (more exaggerated) [82, 85, 218, 219] in individuals diagnosed with ASD compared to NT controls. Some studies [202, 220, 221] have also reported greater asymmetry in the spontaneous facial expressions of individuals diagnosed with ASD. Regarding facial expression mimicry ability, some studies find no group differences, while others report that expressions posed by individuals diagnosed with ASD are less congruent or less accurate than NT controls [13].

A more nuanced view of facial expressions in ASD may be revealed through stratification of such heterogeneous behavioral responses. In a study of facial expression recognition with autistic and NT participants ages 14-55 years, Loth et al. [83] report that 63% of their sample of individuals diagnosed with ASD performed more than two standard deviations below the NT mean, while a smaller subgroup of about 15% of autistic participants performed as well as the NT group. As a part of the EU-AIMS LEAP, Meyer-Lindenberg et al. [84]

investigate facial expression recognition as a candidate stratification biomarker for ASD. In their sample (participants ages 6-30 years and NT or diagnosed with ASD, and/or mild intellectual disability), partitioning individuals diagnosed with ASD into low and high performing subgroups accounts for 9-14% of the variance in ASD-related traits, adaptive behavior, and severity of social difficulties. Furthermore, Meyer-Lindenberg et al. [84] report neurofunctional differences, i.e., significantly lower functional magnetic resonance imaging (fMRI) activation in the amygdala and fusiform gyrus for the low performing subgroup, between subgroups. Differences in ET to social images, including facial expressions, as a characteristic of ASD has been well documented [222]. As a part of ABC-CT, Shic et al. [81] present the Oculomotor Index of Gaze to Human Faces (OMI), an aggregated measure of ET to faces in videos of social scenes, as a candidate stratification biomarker for ASD. OMI has recently been accepted into the FDA’s biomarker qualification program [19]. While OMI focuses on social scenes overall and not on facial expressions in particular, recent ET studies on facial expression recognition have also found reduced visual attention to the face by participants diagnosed with ASD compared to NT controls, as well as greater difficulty in recognizing negative or complex emotions such as ‘anger’, ‘disgust’, and ‘fear’ [223, 224].

For measuring facial expression production, FACS [130] is considered the gold standard. Quinde-Zlibut et al. [85] use FACS AUs to measure spontaneous facial expression production among adult (ages 18-59 years) autistic and NT participants for subgroup identification. For the ASD group, they report that increased expressiveness is associated with poorer facial expression recognition performance [85]. Moreover, Quinde-Zlibut et al. [85] identify a subgroup of autistic adults in their sample that show heightened expressiveness (‘engagement’ summary score from the commercially available iMotions AFFDEX software (https://imotions.com/products/

imotions-lab/modules/fea-facial-expression-analysis/), computed as the average activation of upper and lower face AUs). Bangerter et al. [82] study the spontaneous facial responses of autistic and NT individuals ages 6 to 63 in response to funny videos. Using the iMotions FACET software (no longer commercially available), two constituent AUs of a smile / 'happy' expression are measured: AU 6 and AU 12 [82]. They report that, on average, individuals diagnosed with ASD show less activation of AUs 6 and 12 as compared to the NT control group [82]. However, within the ASD group, they identify "over-responsive" and "under-responsive" subgroups, which display more intense and less intense responses to the stimuli, respectively, as compared to the NT group [82]. Using JAKE, Manfredonia et al. [23] study facial expression mimicry in individuals ages 6-54 years, including an ASD group (n = 144) and an NT group (n = 41). They study activation of two AUs, as measured using the iMotions FACET software, per each of six facial expressions: 'anger' (AUs 4 and 23), 'disgust' (AUs 9 and 10), 'fear' (AUs 4 and 5), 'happy' (AUs 12 and 20), 'sad' (AUs 1 and 15), and 'surprise' (AUs 5 and 26) [23]. Manfredonia et al. [23] report statistically significant differences between autistic and NT individuals' portrayals of 'happy' (AU 12), 'fear' (AU 5), 'surprise' (AU 5), and 'disgust' (AU 9). They also find significant negative correlations between some AUs ('happy' AU 12 and 'fear' AU 5) and social communication scores [23]. Although the purpose of their study is not to identify subgroups, Manfredonia et al. [23] report that in some cases, more activation of AUs corresponded to a greater severity of symptoms, which may suggest a subgroup of individuals with more exaggerated or intense expressions. Drimalla et al. [217] investigate recognition and mimicry of facial expressions in a sample of autistic and NT individuals ages 18 to 62 years. They report significantly higher recognition accuracy and faster response times for the NT group compared to the ASD group [217]. To automatically recognize AUs during expression mimicry of AU-labeled images of actors posing expressions, Drimalla et al. [217] use the open-source OpenFace 2.0 [225] software. Drimalla et al. [217] report that the imitated expressions of participants diagnosed with ASD are significantly more different than the stimulus expressions, with significantly more variance in intensity, and require significantly more time to pose when compared to the NT control group. Furthermore, more accurate imitation of facial expressions is found to be positively associated with better recognition performance [217].

In addition to individual heterogeneity across the spectrum, task design and participant characteristics play an important role in the elicitation and measurement of ASD-related behaviors. Facial expression stimuli may be presented as either static, i.e., still images, or dynamic, i.e., videos or animations. Both static and dynamic expressions have been shown to elicit differential responses, e.g., reduced recognition accuracy, in individuals diagnosed with ASD as compared to NT controls [13]. Keating and Cook [13] point out a need for both static and dynamic expression stimuli to be used in studies of ASD, as it is unclear to what extent autistic individuals may rely on static features (e.g., configuration of the face) versus dynamic features (e.g., order and speed of moving facial muscles) during processing of facial expressions. Effective task design also requires consideration of participant engagement. It has been noted that many individuals on the spectrum have an affinity and interest in technology, including increased engagement with 3D avatar characters [226, 227]. In an ET study on the visual processing of real and avatar faces by children diagnosed with ASD, Pino et al. [75] found that participants show increased interest and more visual exploration of the avatar faces compared to the real faces. Furthermore, customization of avatars has been recommended as an important design consideration for the development of interactive technologies for individuals on the spectrum [227, 228] and has been shown to increase task engagement and enjoyment among both NT and autistic individuals [78]. Individual participant characteristics such as age, gender, and intelligence quotient (IQ) may also affect behavioral responses [13]. Alexithymia, a subclinical personality trait characterized by difficulty in describing one's own emotions and with a prevalence of approximately 50% of the autistic population [229] and 10% of the NT population [230], has also been found to affect perception and production of facial expressions in both NT and autistic individuals [13].

Studies such as the Simons Foundation Powering Autism Research for Knowledge (SPARK) [216] demonstrate the recent success of online ASD research. SPARK now has over 100,000 people diagnosed with ASD and 175,000 of their family members [231] sharing medical and behavioral information online through questionnaires and mailing in saliva for genetic analysis. The rise of online research has also motivated platforms for remote data collection, such as Apple ResearchKit (apple.com/researchkit) which has been applied to ASD research for collection of behavioral data in the Autism & Beyond research study [8]. There may be potential challenges associated with online research. Recently, it has been reported that online qualitative research studies (e.g., focus groups and interviews) have seen a rise in 'scammer participants' attempting to pose as autistic individuals or their parents [232]. Some characteristics of suspect participants include appointment booking data suggesting that they are in different countries than they claim to be in, keeping cameras off during Zoom/Teams interviews, brief and vague responses, discrepancies in responses (e.g., names and ages changing), frequent inquiries about payment, etc. [232]. However, numerous strategies, including careful screening over telephone or videoconferencing, requiring the webcam to be turned on at the beginning of the research session, and checks to ensure that participants are within the geographic limits of the study (e.g., checking the time zone of the appointment booking) may help safeguard data integrity of such studies [232]. Another challenge of online research is diagnostic confirmation. For example, SPARK requires its autistic participants to have received a lifetime professional diagnosis of ASD. Although diagnoses in the SPARK cohort are not independently verified, the study of the validity of self- and caregiver-reported diagnoses has shown good agreement with diagnosis of ASD based on electronic medical records [233]. While telehealth research reports that autistic individuals tend to be more comfortable and relaxed interacting with clinicians from their homes, there may be a need for more parent involvement when interacting with young children [234]. It has also been noted that during online videoconferencing, some autistic individuals may find their own webcam video distracting [234]. Screening participants for suitability and informing participants/parents of what to expect beforehand has been recommended as some children with moderate or severe challenging behaviors may be less likely to stay engaged during online interactions with clinicians [234].

5.1.2 CONTRIBUTIONS

Motivated by recent progress into the discovery and qualification of stratification biomarkers for ASD, we have conducted an IRB-approved pilot study on facial expression perception and production among children and young adults diagnosed with ASD compared to age- and gender-matched NT controls. Prompted by successful online research studies in the literature [8, 216, 231] and the global COVID-19 pandemic, this pilot study has been conducted online with steps taken to ensure the integrity of the data and comfort of participants [234]. Participants have completed recognition and mimicry tasks using previously validated static and dynamic stimuli based on customizable 3D avatars [235] while their webcam captures ET and VT of the face. Since the avatar stimuli are labeled with AUs, we are able to define constructs for expression mimicry based on the avatar AUs. It has been shown that facial expression analysis models that are trained using adult expressions (such as OpenFace 2.0 and iMotions) may perform poorly on child facial expressions [87, 236]. Therefore, we use state-of-the-art deep neural network models [87, 235], that have undergone domain adaptation for use in our age group (children and young adults, ages 8 to 20 years) [87, 237], to extract facial expression and AU labels from the webcam images for behavioral VT. Furthermore, we measure the asymmetry of facial expressions in ASD, which has been investigated by few prior studies [202, 220, 221]. We evaluate our DVs (e.g., participants' facial expressions, activations and asymmetry of AUs, ET measurements, etc. in response to different avatar-rendered facial expressions) using ABC-CT's criteria of construct validity and group discriminability in order to identify candidate stratification biomarkers for future study. Given the large number of parameters designed to capture the AUs, and their complex structures, the use of statistical methods becomes useful to determine the functional forms of the interactions and build group classifications, e.g., ASD vs NT, for group discriminability. The methods we propose are built from the Boruta algorithm models [86, 213, 238] that circumvent unrealistic assumptions of normality and independence in order to capture and showcase class behaviors. We identify one candidate biomarker plus fourteen additional DVs that may be of interest for future research.

The remainder of this chapter is organized as follows. Section 5.2 describes methodology including the study protocol and data analyses. Sections 5.3 and 5.4 present results and discussion, respectively. Section 5.5 discusses limitations, and Section 5.6 concludes with a brief summary.

5.2 METHODS

This section describes the protocol of the proposed study and analytic plan, including IRB approval, groups of participants and inclusion/exclusion criteria, recruitment and screening, informed consent and assent, protection of participant privacy, phenotypic measures, remote experiments and collection of data, data acquisition rates, derivation of DVs, constructs, imputation of missing DVs, application of Boruta methods to group discriminability, and power analysis.

5.2.1 IRB APPROVAL

The IRBs at ODU and EVMS have reviewed and approved the proposed study. The EVMS and ODU IRB approval numbers are 19-06-EX-0152 and 1424272-21, respectively. Approval letters from ODU and EVMS IRBs may be found in Appendices B and C, respectively.

5.2.2 GROUPS OF PARTICIPANTS, INCLUSION AND EXCLUSION CRITERIA

Participants include English-speaking children and young adults between ages 8 to 20 years and residing in the United States. All participants are required to be generally healthy, with either no diagnosis of mood disorders or no change in medication regimen for six months, and IQ of 70 or above. Two groups have been recruited: an NT group and a group of individuals who have received a diagnosis of ASD (henceforth, ASD group). Inclusion and exclusion criteria for each group follow.

5.2.2.1 ASD GROUP

The ASD group includes English-speaking individuals (ages 8 to 20) diagnosed with ASD, residing in the United States, generally healthy, with either no diagnosis of mood disorders or stable mood disorder (off medications and stable OR on the same dose of medication(s) for six months), IQ 70 or above, and the ability to sit in a chair, attend to a computer monitor, and use a mouse to interact with computer software for one hour. The following individuals are excluded from the ASD group: individuals age <8 years or ≥21 years at time of study recruitment, not English-speaking, not residing in the United States, not diagnosed with ASD, have history of severe and chronic illnesses affecting general health, have diagnosis of mood disorders without stable medication regimen for six months, IQ <70, and/or who will be unable to sit in a chair, attend to a computer monitor, and use a mouse to interact with computer software for one hour.

5.2.2.2 NT GROUP

The NT group consists of generally healthy, NT individuals residing in the United States that are age-matched and gender-matched with the individuals in the group with ASD. These individuals have no history of severe and chronic diseases affecting general health. Participants with mood disorders have reported that they are stable without medications or on a stable dose of medication(s) for six months.

5.2.3 RECRUITMENT AND SCREENING

Participants for the ASD group have been recruited via flyers distributed locally and nationwide through the CHKD General Academic Pediatrics and Developmental Pediatrics Clinics; Autism Society Tidewater Virginia and other affiliates of the Autism Society across the continental United States (Acadiana in Louisiana, Central Virginia, Colorado, Iowa, Minnesota, Oregon, Texas, Inland Empire in California); Parents of Autistic Children of Northern Virginia; the Autism Science Foundation; the Organization for Autism Research; and Autism-related Facebook groups (Autism Parents Hampton Roads, Autism Research Study Database, and Autism – All Across the Spectrum). Participants have been recruited for the NT group through flyers posted at CHKD General Academic Pediatrics and ODU. Flyers for both groups have been posted to the ODU Vision Lab website and posted at local public libraries and community centers.

To confirm that all eligibility criteria are met and to safeguard against possible scammer participants, a phone screening interview has been conducted with each participant or their parent/guardian via their United States-based telephone number. Following screening, participants are asked to provide their time zone during scheduling. The time zones are confirmed based on timestamps in scheduling emails. Given the sensitive nature of diagnostic records [232], participants are invited, but not required, to provide the research team with documentation of diagnosis. During screening, all participants confirm that they meet the following minimum technology requirements: 1) access to a personal computer with Internet, 2) webcam, 3) headphones or speakers.

5.2.4 INFORMED CONSENT AND ASSENT

Consent/assent has been obtained electronically as follows. For potential participants who meet the screening criteria, the participant (or parent/guardian if the participant is <18 years of age) is asked to provide an email address to which the Zoom session links and consent/assent documents are sent. The potential participant or parent/guardian reviews the consent/assent documents prior to the first virtual study visit. During the first virtual study visit, a member of the research team explains the informed consent form to the participant or parent/guardian. During the session, the participant or parent/guardian is given the opportunity to review the consent form and ask any questions. The study is explained to the participant at an age-appropriate level. To collect the signatures, the participant (if age 18 or older) or parent/guardian is asked to sign the consent form by either printing, signing, and scanning, OR by providing their electronic signature. Similarly, child assent (for participants under age 18) is obtained either by having the child type their name on the signature line or the parent/guardian may choose to print a copy that the child signs and the parent scans. The signed electronic copies are returned to the investigator via email before moving to the next step of the study.

5.2.5 PROTECTION OF PARTICIPANT PRIVACY

Only IRB-approved investigators for this study have access to the data. To ensure participant confidentiality, data is identified only by a coded participant number. A key with the non-identifying participant number linked to the participant's name is kept separate from the database in a password protected electronic file.

5.2.6 PHENOTYPIC MEASURES

To measure IQ, all participants complete the Kaufman Brief Intelligence Test, Second Edition (KBIT-2). The KBIT-2 has been administered over Zoom by an investigator trained in the administration of this assessment. To measure alexithymia traits, participants 18 or older complete the Bermond-Vorst Alexithymia Questionnaire (BVAQ). For participants under 18, a parent/guardian completes the Children's Alexithymia Measure (CAM) for their child.

Participants are also asked to complete two psychological research measures and a well-known multi-sensory integration task:

- Reading the Mind in the Eyes Task (REMT) (http://www.midss.org/content/reading-mind-eyes-test)
- Cambridge Memory Test of Faces for Children (CMTF) (https://www.ccd.edu.au/services/tools/CFMTC/index2.html)
- McGurk Effect (https://openwetware.org/wiki/Beauchamp:Stimuli#McGurk_and_Control_Audiovisual_Speech_Syllables)

5.2.7 REMOTE EXPERIMENTS AND COLLECTION OF DATA

Motivated by successful online research studies [8, 216, 231], we have designed an online study protocol to conduct during the COVID-19 pandemic. Experiments have been conducted over a Zoom call between the participant (and participant's parent/guardian if the participant is under 18 years old) and research team. The call begins with the webcam turned on in Zoom. A researcher checks that the participant is centered in front of their webcam and instructs the participant or parent/guardian to navigate to a web URL where the experimental tasks are hosted. Once at the web URL, the webcam is disconnected from Zoom and the participant or parent/guardian is instructed to give the website permission to access the webcam feed. To prevent participants from being distracted by their own webcam video [234], no visual webcam feed is shown during the experimental tasks. Participants complete recognition (REC) and mimicry (MIM) tasks using the validated, customizable avatars as previously described in Chapter 4. During the REC task, the participant is asked to click the button for which of six expressions (‘anger’, ‘disgust’, ‘fear’, ‘happy’, ‘sad’, or ‘surprise’) they recognize as being shown on the avatar’s face. During the MIM task, the participant is asked to pose the same expression as the avatar in front of their webcam. Given possible variations in engagement and responses to static or dynamic stimuli, each task is completed under four conditions: uncustomized avatar with static expressions (US), uncustomized avatar with dynamic expressions (UD), customized avatar with static expressions (CS), and customized avatar with dynamic expressions (CD). The order of tasks and conditions is REC-US, MIM-US, REC-UD, and MIM-UD, followed by an avatar customization screen, and then REC-CS, MIM-CS, REC-CD, and MIM-CD. Conditions involving the uncustomized avatar are completed first to avoid biasing participants’ responses to the uncustomized avatar (e.g., due to disappointment) after having created their customized avatar. For each task and each condition, each of the six expressions (‘anger’, ‘disgust’, ‘fear’, ‘happy’, ‘sad’, ‘surprise’) is shown twice for a total of 12 trials. The order of expressions is randomized within each task and condition. WebGazer.js (https://webgazer.cs.brown.edu/) [212] is used to record video frames for facial VT and webcam-based ET fixation coordinates from the participants’ webcams. Participants are offered breaks in between each task and informed that they may at any time request to take as many additional breaks as needed. For each participant, the entire visit including breaks is approximately one hour in duration. Upon completion of the study tasks, each participant is compensated with a $10.00 Visa gift card (valid only in the United States), emailed to the participant’s email address on file.

5.2.8 DATA ACQUISTION RATES

To study acquisition rates for different types of data (e.g., VT, ET) and patterns of possible data loss (e.g., due to participants moving out of frame, ET track loss, etc.), we report the percentage of missing values for each group per task, condition, and expression.

5.2.9 DERIVATIONS OF DVS

Our DVs may be described using (stimulus, measurement) pairs. Stimulus refers to a particular expression ('anger', 'disgust', 'fear', 'happy', 'sad', 'surprise') under a particular stimulus condition (US, UD, CS, or CD) presented during a particular task (REC or MIM). The measurements are defined based on ET, VT, and button click data collected from the participants. There are 312 DVs in total as follows. During the REC task, two types of measurements are collected:

1. The participants' recognition accuracy (%Acc) is calculated based on the percentage of correct responses (clicking the button labeled with the name of the expression that is shown on the avatar).
2. Following Shic et al. [81], we compute %Gaze Face as the percentage of participants' ET fixation duration gazing at the avatar's face.

The total number of DVs for the REC task is $2\ for\ \%Acc\ and\ \ \%Gaze\ Face \times 6\ expressions\ \times 4\ stimulus\ conditions = 48$. During the MIM task, four different types of measurements are collected as follows:

1. The age-appropriate facial expression classification model described in Chapter 3 [87] is used to predict the participants' mimicked expressions from VT. The model outputs softmax probabilities (range 0 to 1) for each expression ('anger', 'disgust', etc.). The

softmax probability corresponding to the stimulus expression is used as a measurement of the participants' ability to mimic the avatar's overall expression (EXPR measurement).

2. The participants' ability to mimic each AU presented by the avatar is quantified by the predicted AU activation (ACT measurement, range 0 to 1) in VT frames using an AU model we adapt from Chapter 4 [235].
3. The left-right asymmetry of the participants' AU activations (ASYM measurement) is computed as the difference of left-right activations for the AU in VT frames as predicted by the AU model we adapt from Chapter 4 [235].
4. We compute %Gaze Face based on ET in the same way as for the REC task.

The total number of DVs for the MIM task is $(2\ for\ \ EXPR\ and\ \%\ Gaze\ Face\ \times 6\ expressions + 2\ for\ ACT\ and\ ASYM \times (7\ AUs\ in\ 'anger'\ + 2\ AUs\ in\ 'disgust'\ + 7\ AUs\ in\ 'fear'\ + 2\ AUs\ in\ 'happy'\ + 4\ AUs\ in\ 'sad'\ + 5\ AUs\ in\ 'surprise')) \times 4\ stimulus\ conditions\ \ = 264.\ 48\ REC\ DVs + 264\ MIM\ DVs = 312\ total\ DVs.$ We use Ganin and Lempitsky [237]'s unsupervised domain adaptation method to adapt the AU model from Chapter 4 [235] for our age group by finetuning the network on facial expression samples collected in this study for 50 epochs with a leaning rate of 1e-7. Using the adapted model, ACT and ASYM values are obtained for 16 AUs (AUs 1, 2, 4, 5, 6, 7, 10, 11, 12, 15, 17, 20, 23, 25, 26, 27). For each task, each measurement is computed for each of the six expressions and four stimulus conditions. Then, the DVs are specified as (stimulus, measurement) pairs such as (MIM-US 'Happy', VT ACT AU 6) or (REC-CD 'Sad', ET %Gaze Face).

5.2.10 CONSTRUCTS

Following ABC-CT [18, 80, 81], we evaluate the construct validity of each DV based on whether the expected response is elicited in the NT group and use one sample t-tests to test for validity. The construct for %Acc based DVs during REC is intact expression recognition, with null hypothesis H_0: μ=16.7% and alternative H_a: μ >16.7%, i.e., the NT mean for %Acc is greater than chance (1 clicked expression button / 6 total expression buttons = 16.7%). The construct for DVs measuring %Gaze Face during REC and MIM is gaze preference to the face, with H_0: μ =15.0% and H_a: μ >15.0%, i.e., the NT mean for %Gaze Face is greater than random gaze (the face occupies 15.0% of the visual scene) [81]. Intact expression mimicry is the construct for DVs measuring EXPR and ACT of AUs during the MIM task. For EXPR, H_0: μ =16.7% and H_a: μ >16.7% (greater than chance). For ACT of AUs, H_0: μ =0 and H_a: μ >0 (AU is present). The expected response in the NT group for ASYM of AUs is symmetrical AU activation. We consider H_0: μ =0 (symmetrical activation) and H_a: μ ≠0 (asymmetrical activation). For all DVs except those based on ASYM of AUs, the construct is valid if H_0 is rejected. For ASYM of AUs, we consider the construct valid if the corresponding construct for ACT of the AU is valid and we fail to reject H_0.

5.2.11 IMPUTATION OF MISSING DVS

Especially when working with children, it is possible for data to be lost due to participants moving out of their calibrated positions. To impute any missing DVs, we evaluate the performance of five different imputation methods on our data set using samples with no missing values. These five methods include simple imputation with 1) the mean or 2) the median value of the DV, multiple imputation by chained equations (MICE) [239, 240] using 3) Bayesian ridge regression or 4) random forest regression, and 5) k-nearest neighbors (KNN) imputation [241].

We follow prior studies [240, 242] to design an experiment to compare among the five imputation methods. From the full data set, we determine the set of DVs that are missing for one or more samples. Next, we identify a reduced data set of samples with no missing DVs. Then, we repeat the following procedure for each target DV in the set of all missing DVs:

1. We consider LOOCV of the reduced data set (samples with no missing values) to obtain train/test splits of the data. In LOOCV, each sample serves as the test set once and remaining samples form the train set.
2. For each split, we save the ground truth value of the target DV from the test sample. Then, we assign 'not a number' ('NaN', denotes missing) to the target DV in the test sample.
3. For each sample in the train set, we randomly assign DVs from the set of all missing DVs to 'NaN' with a probability of 50%. This results in a train set with 50% missing values among DVs from the set of missing DVs.
4. We perform imputation with each of the methods.
5. We repeat steps 2-4 for each of the train/test splits. Using the stored ground truth values and imputed values, we compute evaluation metrics: MSE, root RMSE, and MAE.

We use Scikit-learn (https://scikit-learn.org/) for our implementation. For MICE, we use mean imputation as the initial strategy and obtain the imputed values by averaging over ten repeated imputations. We use the default settings for Bayesian ridge regression. For random forest, we use ten trees in the ensemble. For KNN, we consider the five nearest neighbors. We average the evaluation metrics over all missing DVs to obtain aggregated metrics for performance comparison among imputation methods. We use the best performing method to impute missing values in the full data set.

5.2.12 GROUP DISCRIMINABILITY

Among the DVs with valid constructs, we use the Boruta method [86] to find all DVs that are relevant in discriminating between the NT and ASD groups. To make use of the Boruta method, we consider all DVs with valid constructs as features and group (ASD or NT) as the classification labels. Originally developed with genetics research in mind, Boruta is an 'all-relevant' feature selection method that is appropriate for handling correlated features [86], as is expected to be the case with our DVs. For example, ACT AU 6 and ACT AU 12 are expected to occur together during the mimicry of a 'happy' expression [130]. Boruta offers numerous advantages: identification of all discriminative features given a specified Type I error rate α, high stability of feature selections, and because it is based on trees, no assumption on the distribution of the data [86, 213].

Boruta is implemented as a wrapper around the random forest classification algorithm, an ensemble method comprising decision trees independently developed on different bootstrapped samples of the data [86]. Each tree in the forest assigns an importance value to each feature in the tree based on its contribution to the classification loss [86]. To calculate the importance of each feature in the forest, the average loss for the feature among all trees in which it is present may be divided by its standard deviation to generate a Z score [86]. These Z scores are used to measure feature importance in the Boruta algorithm [86]. Then, Boruta works as follows [86]:

1. The information system is extended by making copies of all of the features, called ‘shadow features’. To remove correlations with the classification labels, the sample values within each of the shadow features are permuted.
2. The random forest algorithm is run on the extended information system to determine the importance values for the features and shadow features.
3. The percentile ($perc$) of the shadow features’ importance is used as a reference value. Features that have a higher importance than the reference value are assigned a ‘hit’. In standard Boruta, $perc = 100$, so features must have higher importance than the most important shadow feature to be assigned a ‘hit’.
4. Steps 1-3 are repeated for a specified number of iterations or until all features are deemed ‘important’ or ‘unimportant’. After each iteration, p-values are computed using the binomial distribution, e.g., a feature is assigned a ‘hit’ k times in n iterations (Bernoulli trials) with the null hypothesis that the probability of a ‘hit’ p is 0.5, i.e., H_0: $p = 0.5$. Two one-tailed binomial tests are performed: a test of rejection (H_a: $p < 0.5$) and a test of confirmation (H_a: $p > 0.5$). In the test of confirmation, we consider features with a Bonferroni-corrected p-value of less than $\alpha = 0.05$ as ‘important’. Similarly, features with a Bonferroni-corrected p-value of less than $\alpha = 0.05$ per the test of rejection are considered ‘unimportant’.

When the number of samples is small, there is a greater likelihood that the permuted shadow features are correlated with the classification labels by chance [238]. To address this issue, a modified version of Boruta, called r-Boruta [238], that adjusts $perc$ to account for this chance correlation may be used. The new value of $perc$ is determined by generating a large number of random features (100,000 in our case), computing the correlation coefficients between the random features and classification labels, and taking the maximum absolute value times 100 as $perc$ [238]. We use the BorutaPy library (https://github.com/scikit-learn-contrib/boruta_py) implementation of Boruta and r-Boruta. Following BorutaPy's documentation, we set the maximum tree depth to 5, allow BorutaPy to dynamically adjust the number of trees, and set the maximum number of iterations to 1000.

5.2.13 POWER ANALYSIS

We use Acharjee et al. [213]'s PowerTools framework for power analysis of candidate biomarkers. PowerTools uses the observed effect size for each DV to generate multiple synthetic data sets following a series of sample sizes (e.g., 22, 44, 88, …). The associated statistical power for each sample size is reported.

5.3 RESULTS

This section presents findings including participant characteristics, data loss, construct validity, imputation study results, group discriminability, and power analysis.

5.3.1 PARTICIPANT CHARACTERISTICS

Twenty-two participants (11 in each group) are included in our final analysis. ASD and NT groups are selected from a total of 32 volunteers (11 diagnosed with ASD and 21 NT) who completed the study. The ASD group includes all 11 participants diagnosed with ASD. The NT group of 11 participants is formed by matching NT participants with participants in the ASD group on age (±1 year) and gender. Eight out of eleven participants in the ASD group have provided documentation of diagnosis. CAM [243] and BVAQ [244] alexithymia scores are standardized based on the population values reported in their respective publications. We obtain alexithymia scores for all participants except two in the ASD group. For two participants in the ASD group, have not received the REMT, and for one participant in the ASD group, we do not have a CMTF score. Participant characteristics are summarized in Table 11.

Shapiro-Wilk tests show that the data collected from the participants exhibit possible departure from normality and may have extreme values or may be thought of as a mixture of feature characteristics. Due to time dependence associated with ordering of tasks, the data collected includes correlation and does not reflect complete independence. To further explore these observations about the data, we propose the deep neural network models and Boruta/r-Boruta models that are appropriate for data with correlation and do not assume independence or normality of the data.

Table 11. Participant characteristics (M: mean, SD: standard deviation)

Characteristic	**ASD Group**	**TD Group**
Number of participants	11	11
Gender (N Males, N Females)	8, 3	8, 3
%Male	72.7%	72.7%
Age in years (M, SD)	14.09 (4.44)	14.00 (4.05)
KBIT Full-scale IQ (M, SD)	100.09 (15.16)	115.45 (12.48)
KBIT Verbal IQ (M, SD)	98.27 (14.42)	110.54 (9.08)
KBIT Nonverbal IQ (M, SD)	101.54 (20.04)	115.72 (17.68)
Standardized Alexithymia Score (M, SD)	0.0103 (0.5315)	-0.3427 (0.9709)
REMT (M, SD)	0.6905 (0.1364)	0.7482 (0.0965)
CMTF (M, SD)	0.8063 (0.1568)	0.9146 (0.0576)
McGurk (M, SD)	0.3636 (0.3931)	0.4091 (0.3754)

5.3.2 DATA LOSS

Table 12 reports the percentage of missing samples by task and data modality (REC ET REC button clicks, MIM ET, and MIM VT). One participant in the ASD group (9.09%) and two participants in the NT group (18.18%) do not have any ET samples due to technological issues (e.g., incompatible hardware, low Internet bandwidth). Additional ET samples are lost due to track loss, e.g., due to participants moving out of range. The primary reason for loss of VT is participants leaning in too close to the camera (cutting off their lower face). The greatest data loss is seen for the CD condition, which occurs at the end of the experimental session, likely due to participants moving out of their original calibrated position. To address data loss, we impute missing values using the robust KNN imputation approach, which provides the best performance in the imputation study with much higher levels of simulated data loss (50%) than observed in the experimental data.

5.3.3 CONSTRUCT VALIDITY

A total of 220 out of 314 DVs have a valid construct. Table 13 summarizes the results for the tests of construct validity. A construct may be invalid either because the task does not elicit the expected response in the NT group and/or we are unable to measure the elicited response. We further validate these findings citing relevant literature in the discussion section.

5.3.4 IMPUTATION STUDY

Comparison of five imputation methods (mean imputation, median imputation, MICE using Bayesian ridge regression, MICE using random forest, and KNN imputation) is carried out with nine participants that have no missing data. Table 14 reports mean MSE, RMSE, and MAE metrics, averaged over all DVs in the missing set. The missing set consists of DVs that have one or more missing samples in the full data set, e.g., 20 or 21 samples instead of the full 22 samples. There are 158 DVs in the missing set. We note that 158 DVs may appear more inflated than actuality as loss of VT samples, e.g., due to a participant leaning in too close to the webcam, may affect multiple DVs. For example, VT for MIM of 'anger' during any stimulus condition (US, UD, CS, CD) is associated with 7 ACT DVs, 7 ASYM DVs, and 1 EXPR DV. Therefore, loss of one VT frame during MIM of 'anger', e.g., under the US condition, will add 15 DVs to the missing set, even if only 1 frame is missing out of the total possible 22. Also, our imputation study simulates higher levels of data loss (50% of samples missing for all DVs) than present in the full data set. The lowest MSE, RMSE, and MAE are achieved by KNN imputation. Therefore, we use KNN imputation to impute the missing values in the full data set.

5.3.5 GROUP DISCRIMINABILITY

Considering a Type I error rate of $\alpha = 0.05$, we identify one candidate biomarker with Boruta: (MIM-US 'Disgust', ET %Gaze Face). To understand the partitioning of groups using (MIM-US 'Disgust', ET %Gaze Face), we fit the CIT [108] shown in Figure 24. The CIT identifies a binary partitioning of samples based on (MIM-US 'Disgust', ET %Gaze Face) and tests the null hypothesis of independence between (MIM-US 'Disgust', ET %Gaze Face) and the group label. The null hypothesis is rejected with a p-value of 0.001. The CIT identifies the partition between groups as 0.53 with all participants in the ASD group, as well as three participants from the NT group, having a value of (MIM-US 'Disgust', ET %Gaze Face) that is greater than 0.53.

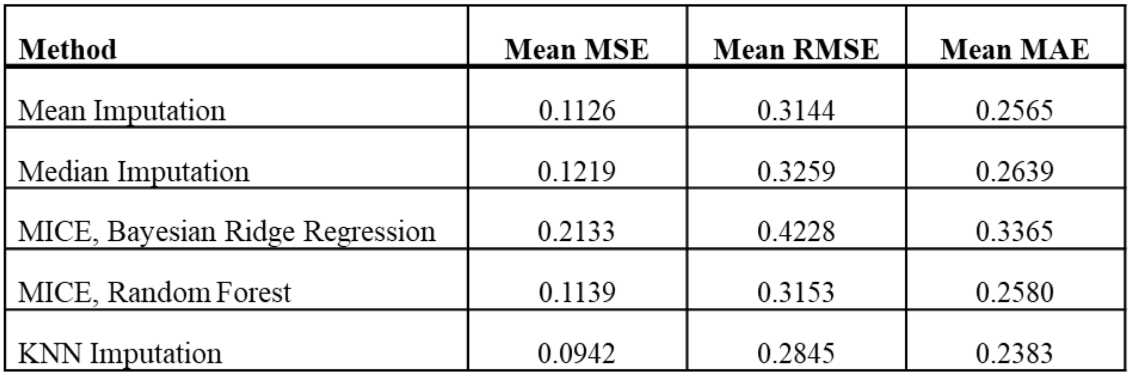
Table 12. Percentage of missing data by task, condition, and expression for ASD and NT groups

surprise	**CD**	18.18%	27.27%	0.00%	0.00%	36.36%	36.36%	27.27%	9.09%
	CS	9.09%	18.18%	0.00%	0.00%	9.09%	18.18%	9.09%	0.00%
	UD	9.09%	18.18%	0.00%	0.00%	9.09%	18.18%	18.18%	0.00%
	US	9.09%	18.18%	0.00%	0.00%	9.09%	18.18%	0.00%	0.00%
sad	**CD**	18.18%	27.27%	0.00%	0.00%	36.36%	36.36%	27.27%	0.00%
	CS	9.09%	18.18%	0.00%	0.00%	9.09%	18.18%	9.09%	0.00%
	UD	9.09%	18.18%	0.00%	0.00%	9.09%	18.18%	9.09%	0.00%
	US	9.09%	18.18%	0.00%	0.00%	9.09%	18.18%	0.00%	0.00%
happy	**CD**	18.18%	36.36%	0.00%	0.00%	36.36%	27.27%	27.27%	0.00%
	CS	9.09%	18.18%	0.00%	0.00%	18.18%	27.27%	9.09%	0.00%
	UD	9.09%	18.18%	0.00%	0.00%	9.09%	18.18%	18.18%	0.00%
	US	9.09%	18.18%	0.00%	0.00%	9.09%	18.18%	0.00%	0.00%
fear	**CD**	27.27%	36.36%	0.00%	0.00%	27.27%	36.36%	27.27%	0.00%
	CS	9.09%	18.18%	0.00%	0.00%	9.09%	18.18%	9.09%	0.00%
	UD	9.09%	18.18%	0.00%	0.00%	9.09%	18.18%	9.09%	0.00%
	US	9.09%	18.18%	0.00%	0.00%	9.09%	18.18%	0.00%	0.00%
disgust	**CD**	18.18%	36.36%	0.00%	0.00%	36.36%	27.27%	27.27%	9.09%
	CS	9.09%	27.27%	0.00%	0.00%	9.09%	18.18%	9.09%	0.00%
	UD	9.09%	18.18%	0.00%	0.00%	9.09%	18.18%	18.18%	0.00%
	US	9.09%	18.18%	0.00%	0.00%	9.09%	18.18%	0.00%	0.00%
anger	**CD**	27.27%	27.27%	0.00%	0.00%	36.36%	45.45%	27.27%	0.00%
	CS	9.09%	18.18%	0.00%	0.00%	9.09%	18.18%	18.18%	0.00%
	UD	9.09%	18.18%	0.00%	0.00%	9.09%	18.18%	9.09%	0.00%
	US	9.09%	18.18%	0.00%	0.00%	9.09%	18.18%	0.00%	0.00%
Task/Modality		**ASD**	**NT**	**ASD**	**NT**	**ASD**	**NT**	**ASD**	**NT**
Legend 0% ▭ 50%		**REC ET**		**REC Button Clicks**		**MIM ET**		**MIM VT**	

Table 13. Construct validity for DVs by task, condition, and expression ('check' denotes a valid construct; for AUs, a 'check' without '*' means both ACT and ASYM are valid while a 'check' with '*' means only ACT is valid)

surprise	CD	✓	✓	✓	✓	✓	✓	▨	✓	▨	▨	▨	▨	▨	▨	▨	▨	▨	✓	▨	✓
	CS	✓	✓	✓	✓	✓	✓	▨	✓	▨	▨	▨	▨	▨	▨	▨	▨	▨	✓	▨	✓
	UD	✓	✓	✓	✓	✓	✓	▨	✓	▨	▨	▨	▨	▨	▨	▨	▨	▨	✓	▨	✓
	US	✓	✓	✓	✓	✓	✓	▨	✓	▨	▨	▨	▨	▨	▨	▨	▨	▨	✓	▨	✓
sad	CD		✓	✓	✓	✓	▨	✓	▨	▨	▨	▨		▨	✓*	▨	▨	▨	▨	▨	▨
	CS	✓	✓	✓		✓	▨	✓	▨	▨	▨	▨		▨	✓	▨	▨	▨	▨	▨	▨
	UD	✓	✓	✓	✓	✓	▨	✓	▨	▨	▨	▨		▨	✓*	▨	▨	▨	▨	▨	▨
	US	✓	✓	✓		✓	▨	✓	▨	▨	▨	▨		▨	✓	▨	▨	▨	▨	▨	▨
happy	CD	✓	✓	✓		▨	▨	▨	▨	✓	▨	▨	▨	✓	▨	▨	▨	▨	▨	▨	▨
	CS	✓	✓	✓		▨	▨	▨	▨	✓	▨	▨	▨	✓	▨	▨	▨	▨	▨	▨	▨
	UD	✓	✓	✓	✓	▨	▨	▨	▨		▨	▨	▨	✓	▨	▨	▨	▨	▨	▨	▨
	US	✓	✓	✓		▨	▨	▨	▨	✓	▨	▨	▨	✓	▨	▨	▨	▨	▨	▨	▨
fear	CD	✓	✓	✓		✓	✓		✓	▨	▨	▨	▨	▨	▨	▨		▨	✓	▨	
	CS	✓	✓	✓		✓	✓	✓	✓	▨	▨	▨	▨	▨	▨	▨	✓	▨	✓	▨	
	UD	✓	✓	✓		✓	✓		✓	▨	▨	▨	▨	▨	▨	▨		▨	✓	▨	
	US	✓	✓	✓		✓	✓		✓	▨	▨	▨	▨	▨	▨	▨	✓	▨	✓	▨	
disgust	CD	✓	✓	✓	✓	▨	▨	▨	▨	▨	▨		▨	▨	▨	✓	▨	▨	▨	▨	▨
	CS	✓	✓	✓	✓	▨	▨	▨	▨	▨	▨		▨	▨	▨	✓	▨	▨	▨	▨	▨
	UD	✓	✓	✓	✓	▨	▨	▨	▨	▨	▨		▨	▨	▨	✓	▨	▨	▨	▨	▨
	US	✓	✓	✓	✓	▨	▨	▨	▨	▨	▨		▨	▨	▨	✓	▨	▨	▨	▨	▨
anger	CD		✓	✓	✓	▨	▨	✓		▨	✓		▨	▨	▨	▨	▨				▨
	CS	✓	✓	✓	✓	▨	▨	✓		▨	✓		▨	▨	▨	▨	▨		✓*		▨
	UD	✓	✓	✓	✓	▨	▨	✓		▨	✓		▨	▨	▨	▨	▨				▨
	US	✓	✓	✓	✓	▨	▨	✓		▨	✓		▨	▨	▨	▨	▨				▨
Task/ Measure		REC %Gaze Face	REC %Acc	MIM %Gaze Face	MIM EXPR	AU 1	AU 2	AU 4	AU 5	AU 6	AU 7	AU 10	AU 11	AU 12	AU 15	AU 17	AU 20	AU 23	AU 25	AU 26	AU 27

Table 14. Comparison of five different imputation methods



Method	Mean MSE	Mean RMSE	Mean MAE
Mean Imputation	0.1126	0.3144	0.2565
Median Imputation	0.1219	0.3259	0.2639
MICE, Bayesian Ridge Regression	0.2133	0.4228	0.3365
MICE, Random Forest	0.1139	0.3153	0.2580
KNN Imputation	0.0942	0.2845	0.2383

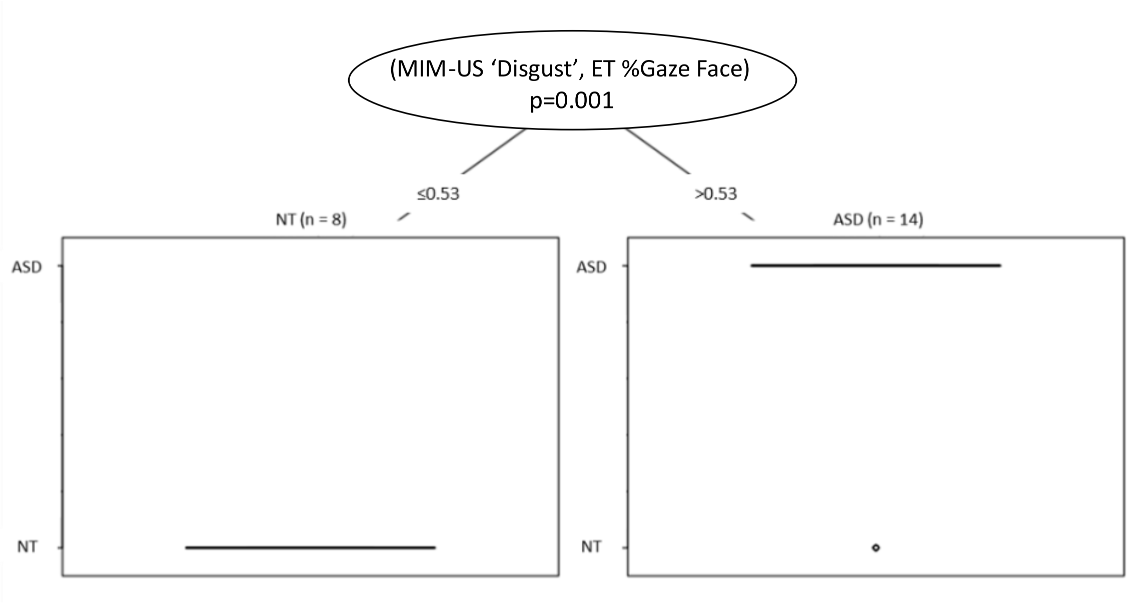


Figure 24. CIT of Group (ASD or NT) with Boruta-selected DV

For r-Boruta, we determine the value of $perc$ for our sample size to be 84.89%. Fifteen DVs are deemed important by r-Boruta, including (MIM-US 'Disgust', ET %Gaze Face) identified by Boruta. Table 15 reports these fifteen DVs along with their group means and standard deviations. Figure 25 shows the box plots by group for the fifteen DVs selected by Boruta and r-Boruta.

Table 15. Group Means (M) and Standard Deviations (SD) for DVs selected by Boruta (italicized) and r-Boruta (all)

DV	ASD	NT
(MIM-US ‘Disgust’, ET %Gaze Face)	*M = 0.77, SD = 0.16*	*M = 0.49, SD = 0.13*
(MIM-UD ‘Anger’, VT ACT AU 7)	M = 0.12, SD = 0.17	M = 0.34, SD = 0.33
(MIM-CD ‘Fear’, VT ACT AU 5)	M = 0.37, SD = 0.29	M = 0.22, SD = 0.34
(MIM-CS ‘Fear’, VT ACT AU 5)	M = 0.35, SD = 0.34	M = 0.14, SD = 0.17
(MIM-US ‘Fear’, VT ACT AU 5)	M = 0.43, SD = 0.32	M = 0.25, SD = 0.20
(MIM-CD 'Happy’, VT ACT AU 6)	M = 0.09, SD = 0.09	M = 0.29, SD = 0.29
(MIM-US ‘Surprise’, VT ACT AU 2)	M = 0.72, SD = 0.21	M = 0.47, SD = 0.30
(MIM-CS ‘Fear’, VT ASYM AU 25)	M = 0.37, SD = 0.33	M = 0.12, SD = 0.18
(MIM-CD ‘Fear’, VT ASYM AU 5)	M = 0.28, SD = 0.28	M = 0.11, SD = 0.15
(MIM-CS ‘Fear’, VT ASYM AU 5)	M = 0.30, SD = 0.29	M = 0.13, SD = 0.19
(MIM-US ‘Surprise’, VT ASYM AU 5)	M = 0.44, SD = 0.29	M = 0.24, SD = 0.30
(REC-CD ‘Fear’, ET %Gaze Face)	M = 0.50, SD = 0.18	M = 0.34, SD = 0.08
(REC-CS ‘Fear’, ET %Gaze Face)	M = 0.61, SD = 0.19	M = 0.52, SD = 0.28
(MIM-CS ‘Sad’, ET %Gaze Face)	M = 0.40, SD = 0.18	M = 0.56, SD = 0.30
(MIM-UD ‘Sad’, ET %Gaze Face)	M = 0.60, SD = 0.20	M = 0.41, SD = 0.18

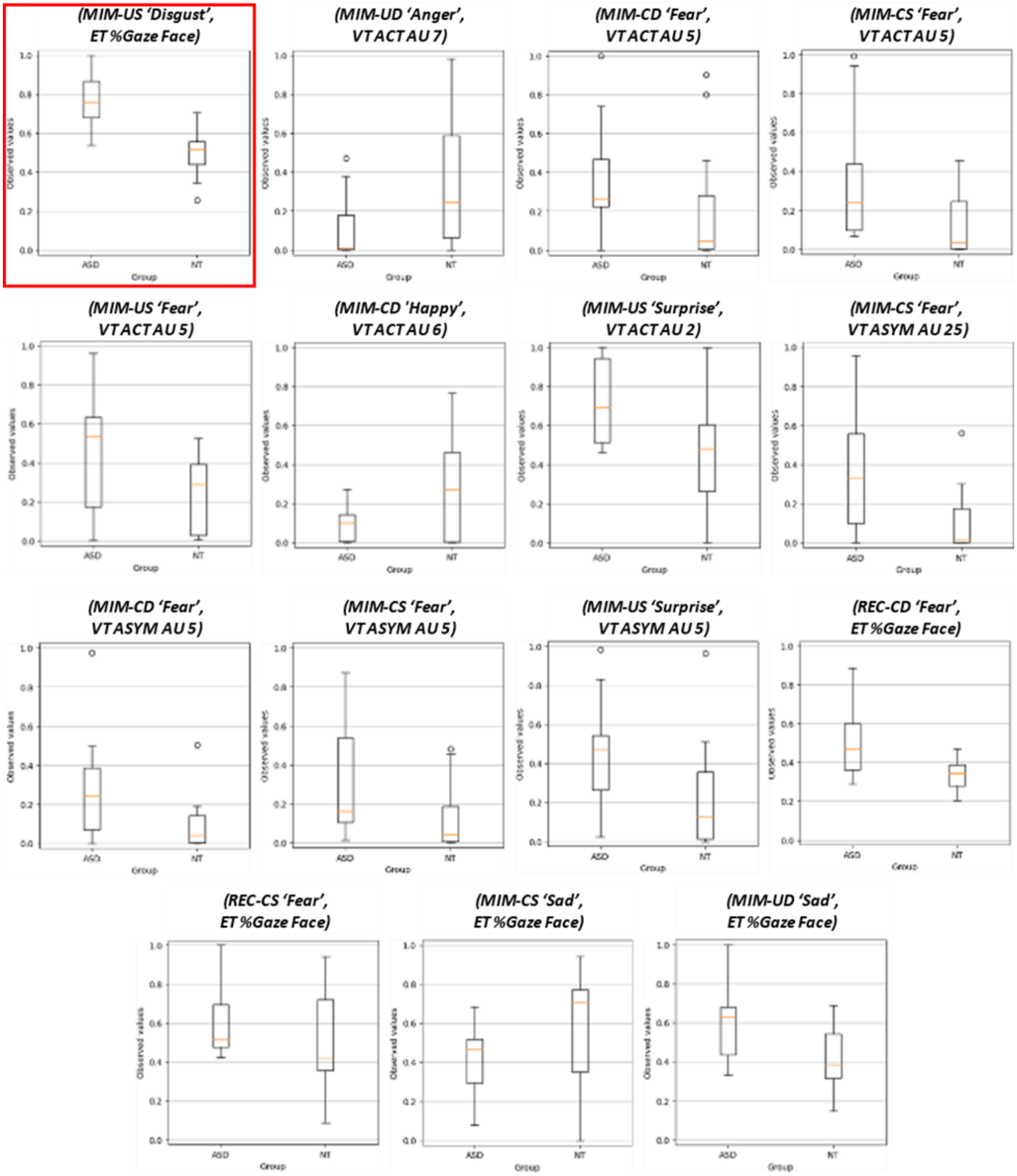


Figure 25. Box plots for fifteen DVs selected by Boruta (outlined in red) and r-Boruta (all)

5.3.6 POWER ANALYSIS

Figure 26 shows the PowerTools [213] output for estimated statistical power using a series of simulated samples of sizes 22, 44, 88, 176, 352, and 704 and the observed effect sizes for the fifteen DVs selected by Boruta and r-Boruta. The power analysis corroborates the choice of Boruta for candidate biomarker selection, showing that maximum power for (MIM-US 'Disgust', ET %Gaze Face) is achieved for our current sample size of 22.

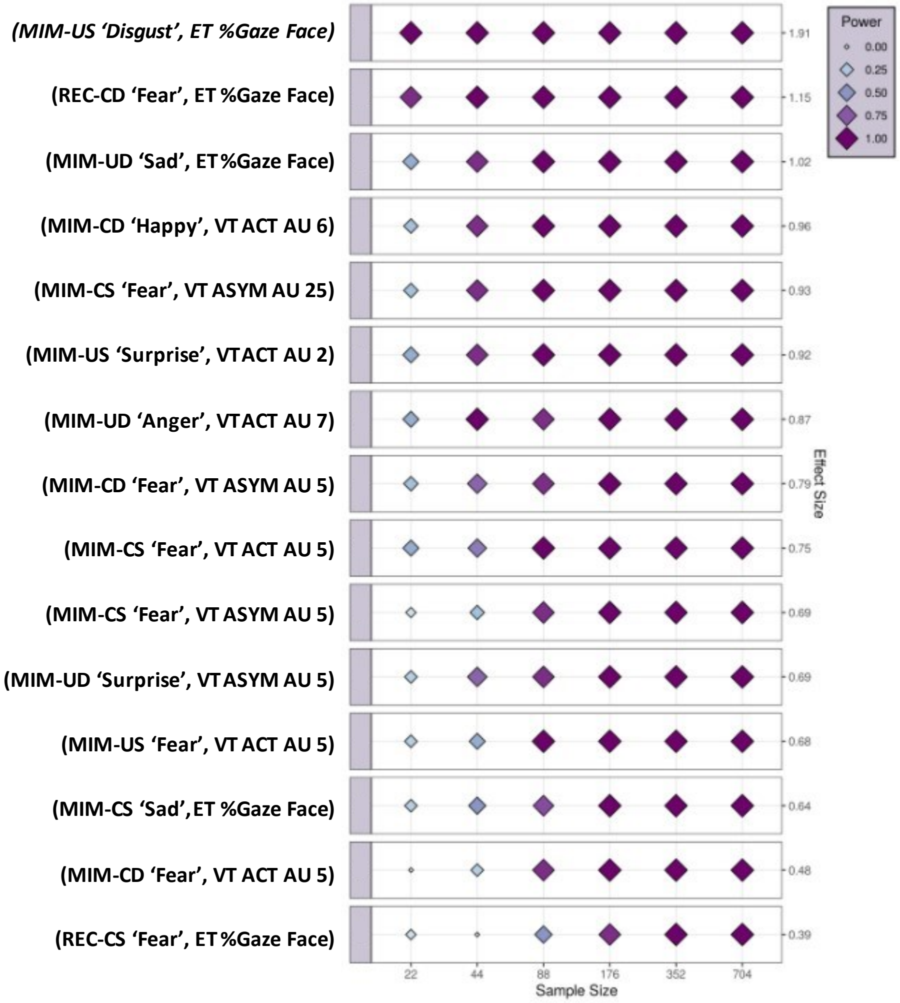


Figure 26. PowerTools visualization of observed effect sizes for selected DVs (candidate biomarker selected by Boruta in italics) and statistical power estimates for different sample sizes

5.4 DISCUSSION

Our findings demonstrate the feasibility of DVs related to facial expression perception and production in stratification biomarker discovery for ASD, considering the criteria of construct validity and group discriminability. With regard to construct validity (Table 13), invalid constructs may be due to limitations of the measurement tools (e.g., deep neural network models) or failure to elicit the expected response from NT participants (e.g., participants do not look at the face or do not produce the expression as expected).

For the tools, we use the stimuli and models for AU measurement developed in Chapter 4. Prior study of these tools report limitations for construct validity. In our prior study (Chapter 4) [235] of these tools, the following AUs are reported to have invalid constructs: AUs 5, 23, and 26 during ‘anger’; AU 10 during ‘disgust’; AU 20 during ‘fear’; AU 6 during ‘happy’; and AU 11 during ‘sad’. Therefore, it is an expected finding that some of our constructs involving these same AUs (5, 6, 10, 11, 20, 23, and 26) are invalid. For the DVs involving EXPR measurements, some of the constructs for ‘fear’ (US, UD, CS, CD), ‘happy’ (US, CS, CD), and ‘sad’ (US, CS) are invalid. This may be due to differences in the image characteristics of our data set (e.g., lighting, backgrounds) or in the imitated expressions from our NT group compared to the prototypical expressions of the model’s training data (Chapter 3) [87].

NT participant characteristics may also have an impact on construct validity. Studies of facial expression production with NT children have reported that negative expressions may be more difficult to elicit from children, including ‘anger’ [245], ‘fear’ [141], and ‘sad’ [245]. Furthermore, Grossard et al. [245] find that NT children produce significantly higher quality facial expressions on request versus during imitation of an avatar and that the imitated expressions may be less credible and less recognizable. Therefore, some of the AUs and expressions in our study may be difficult to elicit especially among younger members of our age group. This may explain why we see invalid constructs for AU 25 ‘lips part’ (ACT and ASYM for US, UD, CD; ASYM only for CS) during ‘anger’, and for AU 4 ‘brow lowerer’ (ACT and ASYM for US, UD, CD) and AU 27 ‘mouth stretch’ (ACT and ASYM for US, UD, CS, CD) during ‘fear’. Also, the only two ET DV constructs that are not valid, (REC-CD ‘Sad’, ET %Gaze Face) and (REC-CD ‘Anger’, ET %Gaze Face), both occur at the end (CD condition) of the experiment and may be invalid due to waning attention from the some of the participants.

Assessment of group discriminability criteria using Boruta has identified one candidate biomarker (MIM-US ‘Disgust’, ET %Gaze Face). (MIM-US ‘Disgust’, ET %Gaze Face) is greater for the ASD group, indicating that participants diagnosed with ASD spend more percentage gaze duration to the face while viewing the static ‘disgust’ expression presented on the uncustomized avatar compared to the NT group (Table 15). Our results are consistent with existing literature noting increased preference and engagement towards avatars among individuals diagnosed with ASD [226, 227], as well as Pino et al. [75]’s finding of higher gaze duration towards negative emotions. Using r-Boruta, we also identify fourteen additional DVs of interest, which we interpret with caution. No DVs involving %Acc or EXPR measurements are selected by either Boruta or r-Boruta. Like (MIM-US ‘Disgust’, ET %Gaze Face), the other ET DVs identified by r-Boruta ((REC-CD ‘Fear’, ET %Gaze Face), (REC-CS ‘Fear’, ET %Gaze Face), (MIM-CS ‘Sad’, ET %Gaze Face), and (MIM-UD ‘Sad’, ET %Gaze Face)) also involve negative emotions (‘fear’ or ‘sad’) and except for (MIM-CS ‘Sad’, ET %Gaze Face), report higher percentage gaze duration in the ASD group compared to the NT group (Table 15).

Six DVs involving ACT of AUs ((MIM-UD ‘Anger’, VT ACT AU 7), (MIM-CD ‘Fear’, VT ACT AU 5), (MIM-CS ‘Fear’, VT ACT AU 5), (MIM-US ‘Fear’, VT ACT AU 5), (MIM-

CD ‘Happy’, VT ACT AU 6), (MIM-US ‘Surprise’, VT ACT AU 2)) are selected by r-Boruta. Four of these DVs ((MIM-CD ‘Fear’, VT ACT AU 5), (MIM-CS ‘Fear’, VT ACT AU 5), (MIM-US ‘Fear’, VT ACT AU 5), (MIM-US ‘Surprise’, VT ACT AU 2)) show greater activation, on average, by the ASD group compared to the NT group (Table 15). These findings corroborate prior mimicry studies that report more intense ‘fear’ [218] and ‘surprise’ [219] expressions by the ASD group relative to the NT group. In Manfredonia et al. [23]’s study, group differences are also reported for ‘fear’ and ‘surprise’ expressions, specifically in AU 5 ‘upper lid raiser’. Manfredonia et al. [23]’s ASD group shows lower activation of AU 5 on average relative to the NT group. However, Manfredonia et al. [23] elicit facial responses using textual prompts rather than mimicry. On average, our ASD group has lower (MIM-UD ‘Anger’, VT ACT AU 7) and (MIM-CD ‘Happy’, VT ACT AU 6) compared to the NT group (Table 15). Prior studies have reported more intense [218] or no difference [23, 219] in the production of ‘anger’. However, these prior studies use either static expressions [218, 219] or a textual prompt [23] to elicit facial responses. Recently, Keating et al. [246] report significantly poorer recognition of dynamic ‘anger’ among their sample of participants diagnosed with ASD compared to an NT control group, even when controlling for alexithymia. It is possible that group differences in responses elicited by dynamic ‘anger’ stimuli may also extend to AU production and may be of interest for further study. Similar to Bangerter et al. [82], we also find less activation of AU 6 during production of ‘happy’ ((MIM-CD ‘Happy’, VT ACT AU 6)), on average, in the ASD group compared to the NT group. We do not observe [217]’s finding of lower standard deviations of AU activations in the NT group compared to the ASD group. Heterogenous findings across our study and others may support the presence of subgroups of autistic participants that are more and less expressive [85]. However, these are preliminary findings and follow up studies are required.

For all five DVs involving ASYM of AUs ((MIM-CS ‘Fear’, VT ASYM AU 25), (MIM-CD ‘Fear’, VT ASYM AU 5), (MIM-CS ‘Fear’, VT ASYM AU 5), (MIM-UD ‘Surprise’, VT ASYM AU 5)) identified by r-Boruta, the ASD group shows higher asymmetry on average than the NT group (Table 15). Prior studies [202, 220, 221] on facial expression asymmetry in ASD have also found that individuals diagnosed with ASD may have higher levels of asymmetry in their expressions, specifically in the activation of left and right levator anguli oris muscles of the lower face (associated with AU 13 ‘cheek puffer’, which is not present in the stimulus expressions) [202, 221]. In the lower face, we identify one DV that involves AU 25 ‘lips part’, which may be associated with multiple facial muscles including depressor labii inferioris, mentalis, and orbicularis oris [130]. While prior studies focus on overall left-right asymmetry [220] or a few muscles of the lower face [202, 221], we are able to capture muscle activations of the upper face as well. Three DVs related to asymmetry of AU 5 ‘upper lid raiser’ activations identified in our analysis suggest that asymmetry of movements of the upper face may also be of interest for future study.

Power analyses (Figure 26) show that maximum power is attained for (MIM-US ‘Disgust’, ET %Gaze Face) with our current sample size of 22, corroborating the use of Boruta for reliable selection of candidate biomarkers. For replication of our findings, future studies may consider sample size recommendations from Figure 26 based on the DVs of interest and desired statistical power.

5.5 LIMITATIONS

Although our sample size is relatively small, the variance and covariance patterns in our sample have been managed by matching participants on age and gender. Further examination of larger data may shed additional information on perception and production behaviors. Our intuition in selecting robust DVs guided us in the use of statistical tools (Boruta and r-Boruta) that do not require independence or normality assumptions and work well for correlated data. Another direction could have been to use nonparametric or more general distribution functions (such as copula types). However, with the nonparametric approach, we may lose the dependence structure of the data, and with a copula-based approach, the dependence structure will be transformed based on the cumulative distribution function at the cost of interpretability. Further larger studies that incorporate deep phenotyping of participants and replication samples will move findings toward a more comprehensive understanding of these DVs and how they relate to IQ and other phenotypic variables (e.g., severity of ASD-related symptoms, social communication scores, adaptive function, etc.).

5.6 SUMMARY

This study demonstrates the feasibility of and provides a framework for ASD stratification biomarker discovery based on production and perception of facial expressions. We identify one candidate biomarker as well as fourteen additional DVs of interest and provide sample size recommendations for future studies. Our study has found evidence of both more and less intense expressions in our ASD group, on average, depending on the stimulus type. More research is required to confirm and understand the significance of possible subgroups of autistic individuals based on these initial findings. Furthermore, we find several important DVs related to asymmetry of facial movements among individuals on the spectrum that we hope will facilitate follow up studies. Furthermore, we hope that this study will provide a foundation for larger studies involving deep phenotyping of participants and replication samples.

CHAPTER 6

CONCLUSION AND FUTURE WORK

This dissertation investigates facial expression perception and production in behavioral imaging (VT) and eye gaze (ET) to identify candidate stratification biomarkers for children and young adults diagnosed with ASD. To accomplish this objective, we define three goals for the dissertation. The outcomes of the dissertation are summarized for each goal in

Table 16 and further discussed in Section 6.1 below, followed by directions for future work in Section 6.2.

6.1 CONCLUSION

In the first goal [87], we address the important challenge of age-invariant FEA by proposing novel deep domain adaptation and fusion of geometric landmark features to yield a deep learning model that is appropriate for FEA across adult and child facial expression domains. Accordingly, novel concurrent learning of adult and child facial expressions produces a domain-invariant latent feature representation for improved generalizability of facial expression classification across age groups. For the first time in the literature, we use the betaMix method to perform feature selection for deep learning. Using the betaMix method, we decompose landmark features based on their correlations with expression, domain, and identity factors to select and fuse useful and explainable features for expression classification that are invariant to domain and identity. Our proposed model performs competitively or better over comparison methods (baseline CNNs, transfer learning, and other domain adaptation approaches) across multiple benchmark data sets. Visualization of SHAP values provides explainability to corroborate the classification performance.

Table 16. Summary of dissertation outcomes

Contributions	**Results**	**Publications**
GOAL 1 (CHAPTER 3): Deep representation learning of adult and child facial expressions using domain adaptation fusing facial landmark features		
• Performed novel concurrent adult and child expression learning to yield domain-invariant facial expression classification. • Decomposed facial landmark features based on expression, domain, and identity correlations. • Proposed novel feature selection for deep learning using the betaMix statistical approach. • Fused facial landmark measurements with deep feature representations for robust expression learning across age groups. • Provided feature explainability using SHAP values	The proposed method shows competitive or better facial expression classification performance over comparison methods for multiple benchmark data sets. Explainability and visualization of SHAP values corroborates the facial expression classification performance.	Published **three conference** papers [43, 44, 88] and **one journal** paper [87]
GOAL 2 (CHAPTER 4): Customizable avatars with dynamic facial action coded expressions for improved user engagement		
• Developed six customizable avatars for improved user engagement, labeled with AUs by a FACS expert. • Trained and evaluated deep learning models for bilateral and unilateral AU detection. • Improved representation learning by fusing geometric landmark and deep learning-based texture features while jointly learning AUs and expressions. • Proposed novel beta-guided correlation loss to encourage feature correlation with AUs while discouraging correlation with subject identity. • Conducted a feasibility study with twenty heathy adults and assessed construct validity of proposed stimuli and measurements (AU activations and ET percentage gaze duration to the face).	The proposed AU detection model achieves state-of-the-art performance on the CK+ data set, our primary benchmark set. Assessment of construct validity reveals that all constructs are valid for ET based measurements in response to the stimuli and the majority of constructs are valid for the AU measurements.	**One journal** paper [89] (under review)
GOAL 3 (CHAPTER 5): Pilot study to discover candidate biomarkers for ASD based on perception and production of facial expressions		
• Conducted online pilot study of facial expression perception and production during recognition and mimicry tasks with participants diagnosed with ASD and matched NT peers. • Collected measurements of facial expressions in VT, activation and asymmetry of AUs in VT, and ET percentage gaze duration to the face while participants interact with static and dynamic facial expressions posed by customized and uncustomized avatars. • Evaluated DVs based on criteria of construct validity and group discriminability. • Proposed Boruta algorithm models for group discriminability to overcome assumptions of normality and independence.	One candidate biomarker (percentage gaze duration to the static 'disgust' facial expression shown by an uncustomized avatar) and fourteen additional DVs of possible interest for future study are identified. Based on power analysis, a sample size of at least 176 (to yield a power of at least 0.75 for DVs of potential interest) is suggested for future studies.	**One journal** paper [90] (under review)

The second goal [89] of this dissertation addresses another important challenge: FACS-labeled avatar-based stimuli for improved user engagement during elicitation of facial expression AUs and associated deep learning models for automated measurement of participant AU responses. To this end, we have worked with a certified FACS expert to develop customizable avatars with dynamic, FACS-labeled animations for six facial expressions (‘anger’, ‘disgust’, ‘fear’, ‘happy’, ‘sad’, and ‘surprise’). We have also developed deep learning models for multi-label AU detection, incorporating feature fusion, multi-task learning of AUs and expressions, and a novel beta-guided correlation loss to achieve state-of-the-art performance on the CK+ benchmark data set that is used as our primary benchmark data set. Construct validity of the proposed stimuli and associated measurements (ET percentage gaze duration to the face and activation of AUs) is evaluated based on data collected in an online feasibility study of twenty healthy adults. Assessment of construct validity reveals that all constructs are valid for ET based measurements in response to the stimuli and the majority of constructs are valid for the AU measurements.

Finally, in the third goal [90], we conduct an online pilot study where age- and gender-matched ASD and NT groups of participants interact with stimuli from Goal 2 while behavioral imaging (VT) and eye gaze (ET) of their facial expression production and perception are collected. Deep learning models for FEA developed in Goals 1 and 2 are used to process VT frames to yield quantitative measurements of participants’ facial expressions in response to the stimuli. Candidate stratification biomarker criteria based on construct validity and group discriminability are used to assess the DVs derived from ET and VT. For group discriminability, we build our approach on Boruta algorithm models to overcome assumptions on the independence and normality of the data. Our approach has resulted in the identification of one candidate biomarker for ASD based on percentage duration of ET gaze to a static ‘disgust’ expression posed by an uncustomized avatar. We also identify fourteen additional DVs of possible interest for future study, including DVs related to activations of AUs 2, 5, 6, and 7 during production of ‘anger’, ‘fear’, ‘happy’, and ‘surprise’ expressions; asymmetry of AU 5 and AU 25 activations while posing ‘fear’ and ‘surprise’; and percentage duration of gaze to the avatar’s face while viewing ‘fear’ and ‘sad’ expressions. Following a power analysis based on the observed effect sizes for these DVs, we recommend a sample size of at least 176 (or 88 in each group) for future studies to yield a power of at least 0.75 for all fourteen DVs.

6.2 FUTURE WORK

Future research directions may build on any of the aforementioned goals to facilitate the development of more sophisticated behavioral measurements, more complex and engaging stimuli, and improved experimental design for future studies of biomarker discovery or intervention. The measurements proposed in this dissertation do not capture the temporal features of participants’ behavioral responses. Rather, we perform our analysis for individual frames of VT and aggregate ET information over the duration of the task. Development of appropriate methods for analyzing facial expressions and AUs over a sequence of frames may yield interesting quantitative measurements of the temporal evolution of expressions. There are numerous exciting challenges including the increased complexity of deep learning models and (possibly unsupervised) domain adaptation methods that will require training and evaluation with limited available labeled data. Time-series features may also be extracted from ET data to build new DVs.

With regard to the stimuli, improvements may be made to the rigging and appearance (e.g., wrinkles) of the 3D avatar models for more accurate rendering of facial expressions and AUs. More complex expressions in addition to the basic six (‘anger’, ‘disgust’, ‘fear’, ‘happy’ ‘sad’, ‘surprise’) used in this dissertation may be created to elicit different combinations of AUs. The benefits of expanding the library of expressions for the avatars is twofold. First, it would enable study of perception and production of more complex expressions which may better differentiate among subgroups on the spectrum. Second, the extended library may be used to augment the limited training data for AU detection. While we only use the few basic avatar prototypes in this dissertation, the avatar generation software is able to create diverse avatars of many different ages, shapes, sizes, etc., at varying levels of realism with millions of possible combinations. Therefore, the limiting factor for the creation of a rich and diverse data set of avatar facial expressions is the process of FACS labeling. Development of a robust automated pipeline to assist with labeling may facilitate this process. Increasing the number and variety of customization options available to participants may also increase their engagement and self-identification with the avatars.

Future biomarker discovery and biomarker qualification efforts considering facial expression perception and production in ASD would benefit from larger sample sizes, which may allow for clustering of subgroups on the spectrum. Additionally, having participants undergo diagnostic confirmation (e.g., via the Autism Diagnostic Observation Schedule, Second Edition (ADOS-2)) and deep clinical phenotyping (e.g., communication scores, adaptive behavior scores, etc.) as a part of the research will enable investigation of relationships between candidate biomarkers and clinical scores. Test-retest reliability of candidate biomarkers may be investigated by having participants take part in multiple visits over time (e.g., after six weeks).

Furthermore, recruitment of a replication sample may be done to assess the reliability of the research findings for independent cohorts of participants. Integration of FEA models (expression classification and AU detection) in the loop with the tasks and stimuli may enable new interventions for facial expression mimicry that may provide specific feedback on participants' facial movements. Near real-time ET feedback may be used to prompt and guide participants' gaze towards specific areas of the face. Large studies such as ABC-CT involve large steering committees that weigh in on research directions for maximum benefit. Partnering with groups such as the Autistic Self Advocacy Network (ASAN) at the design stage of future research studies may help steer research in directions most valued by the autistic community.

REFERENCES

[1] W. Z. Maenner MJ, Williams AR, et al., *Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years — Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2020*, vol. 72, Centers for Disease Control, 11 Sites, United States, 2023.

[2] CDC Autism and Developmental Disabilities Monitoring Network, "Identified Prevalence of Autism Spectrum Disorder: ADDM Network 2000-2020 Combining Data from All Sites," Centers for Disease Control, 2023.

[3] Foundation for the National Institutes of Health. "Biomarkers Consortium - The Autism Biomarkers Consortium for Clinical Trials (ABC-CT)," July 1, 2022; https://fnih.org/our-programs/biomarkers-consortium/autism-biomarkers.

[4] H. Hodges, C. Fealko, and N. Soares, "Autism spectrum disorder: definition, epidemiology, causes, and clinical evaluation," *Transl Pediatr,* vol. 9, no. Suppl 1, pp. S55-s65, Feb, 2020.

[5] J. W. Harrington, and K. Allen, "The clinician's guide to autism," *Pediatr Rev,* vol. 35, no. 2, pp. 62-78; quiz 78, Feb, 2014.

[6] C. Bridgemohan, N. S. Bauer, B. A. Nielsen, A. DeBattista, H. S. Ruch-Ross, L. B. Paul, and N. Roizen, "A Workforce Survey on Developmental-Behavioral Pediatrics," *Pediatrics*, pp. e20172164, 2018.

[7] R. A. J. de Belen, T. Bednarz, A. Sowmya, and D. Del Favero, "Computer vision in autism spectrum disorder research: a systematic review of published studies from 2009 to 2019," *Translational Psychiatry,* vol. 10, no. 1, pp. 333, 2020/09/30, 2020.

[8] H. L. Egger, G. Dawson, J. Hashemi, K. L. Carpenter, S. Espinosa, K. Campbell, S. Brotkin, J. Schaich-Borg, Q. Qiu, and M. Tepper, "Automatic emotion and attention analysis of young children at home: a ResearchKit autism feasibility study," *NPJ digital medicine,* vol. 1, no. 1, pp. 20, 2018.

[9] F. Marino, P. Chilà, C. Failla, I. Crimi, R. Minutoli, A. Puglisi, A. A. Arnao, G. Tartarisco, L. Ruta, D. Vagni, and G. Pioggia, "Tele-Assisted Behavioral Intervention for Families with Children with Autism Spectrum Disorders: A Randomized Control Trial," *Brain Sciences,* vol. 10, no. 9, pp. 649, 2020.

[10] A. Narzisi, "Phase 2 and Later of COVID-19 Lockdown: Is it Possible to Perform Remote Diagnosis and Intervention for Autism Spectrum Disorder? An Online-Mediated Approach," *Journal of Clinical Medicine,* vol. 9, no. 6, pp. 1850, 2020.

[11] P. Washington, N. Park, P. Srivastava, C. Voss, A. Kline, M. Varma, Q. Tariq, H. Kalantarian, J. Schwartz, R. Patnaik, B. Chrisman, N. Stockham, K. Paskov, N. Haber, and D. P. Wall, "Data-Driven Diagnostics and the Potential of Mobile Artificial Intelligence for Digital Therapeutic Phenotyping in Computational Psychiatry," *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging,* vol. 5, no. 8, pp. 759-769, 2020/08/01/, 2020.

[12] *Diagnostic and statistical manual of mental disorders : DSM-5*, Arlington, VA: American Psychiatric Association, 2013.

[13] C. T. Keating, and J. L. Cook, "Facial expression production and recognition in autism spectrum disorders: A shifting landscape," *Child and Adolescent Psychiatric Clinics,* vol. 29, no. 3, pp. 557-571, 2020.

[14] F. Happé, and U. Frith, “Annual Research Review: Looking back to look forward–changes in the concept of autism and implications for future research,” *Journal of Child Psychology and Psychiatry,* vol. 61, no. 3, pp. 218-232, 2020.

[15] E. Loth, and D. W. Evans, “Converting tests of fundamental social, cognitive, and affective processes into clinically useful bio-behavioral markers for neurodevelopmental conditions,” *Wiley Interdisciplinary Reviews: Cognitive Science,* vol. 10, no. 5, pp. e1499, 2019.

[16] E. Loth, T. Charman, L. Mason, J. Tillmann, E. J. Jones, C. Wooldridge, J. Ahmad, B. Auyeung, C. Brogna, and S. Ambrosino, “The EU-AIMS Longitudinal European Autism Project (LEAP): design and methodologies to identify and validate stratification biomarkers for autism spectrum disorders,” *Molecular autism,* vol. 8, no. 1, pp. 1-19, 2017.

[17] S. Ness, N. V. Manyakov, A. Bangerter, D. Lewin, S. Jagannatha, M. Boice, A. Skalkin, G. Dawson, M. S. Goodwin, and R. L. Hendren, “1.32 THE JANSSEN AUTISM KNOWLEDGE ENGINE (JAKE™): A SET OF TOOLS AND TECHNOLOGIES TO ASSESS POTENTIAL BIOMARKERS FOR AUTISM SPECTRUM DISORDERS,” *Journal of the American Academy of Child & Adolescent Psychiatry,* vol. 10, no. 55, pp. S110, 2016.

[18] J. C. McPartland, R. A. Bernier, S. S. Jeste, G. Dawson, C. A. Nelson, K. Chawarska, R. Earl, S. Faja, S. P. Johnson, and L. Sikich, “The autism biomarkers consortium for clinical trials (ABC-CT): scientific context, study design, and progress toward biomarker qualification,” *Frontiers in integrative neuroscience,* vol. 14, pp. 16, 2020.

[19] U.S. Food & Drug Administration. "Biomarker Qualification Submissions," 01/31/2024, 2024; https://www.fda.gov/drugs/biomarker-qualification-program/biomarker-qualification-submissions.

[20] U. M. Schaller, M. Biscaldi, A. Burkhardt, C. Fleischhaker, M. Herbert, A. Isringhausen, L. Tebartz van Elst, and R. Rauh, “ADOS-Eye-Tracking: The Archimedean Point of View and Its Absence in Autism Spectrum Conditions,” *Frontiers in Psychology,* vol. 12, 2021-March-18, 2021.

[21] M. H. Black, N. T. M. Chen, K. K. Iyer, O. V. Lipp, S. Bölte, M. Falkmer, T. Tan, and S. Girdler, “Mechanisms of facial emotion recognition in autism spectrum disorders: Insights from eye tracking and electroencephalography,” *Neuroscience & Biobehavioral Reviews,* vol. 80, pp. 488-515, 2017/09/01/, 2017.

[22] E. A. Papagiannopoulou, K. M. Chitty, D. F. Hermens, I. B. Hickie, and J. Lagopoulos, “A systematic review and meta-analysis of eye-tracking studies in children with autism spectrum disorders,” *Social Neuroscience,* vol. 9, no. 6, pp. 610-632, 2014/11/02, 2014.

[23] J. Manfredonia, A. Bangerter, N. V. Manyakov, S. Ness, D. Lewin, A. Skalkin, M. Boice, M. S. Goodwin, G. Dawson, R. Hendren, B. Leventhal, F. Shic, and G. Pandina, “Automatic Recognition of Posed Facial Expression of Emotion in Individuals with Autism Spectrum Disorder,” *Journal of Autism and Developmental Disorders,* vol. 49, no. 1, pp. 279-293, 2019/01/01, 2019.

[24] M. Leo, P. Carcagnì, C. Distante, P. L. Mazzeo, P. Spagnolo, A. Levante, S. Petrocchi, and F. Lecciso, “Computational Analysis of Deep Visual Data for Quantifying Facial Expression Production,” *Applied Sciences,* vol. 9, no. 21, pp. 4542, 2019.

[25] M. Leo, P. Carcagnì, C. Distante, P. Spagnolo, P. L. Mazzeo, A. C. Rosato, S. Petrocchi, C. Pellegrino, A. Levante, F. De Lumè, and F. Lecciso, “Computational Assessment of Facial Expression Production in ASD Children,” *Sensors,* vol. 18, no. 11, pp. 3993, 2018.

[26] T. Guha, Z. Yang, R. B. Grossman, and S. S. Narayanan, “A Computational Study of Expressive Facial Dynamics in Children with Autism,” *IEEE Transactions on Affective Computing,* vol. 9, no. 1, pp. 14-20, 2018.

[27] M. D. Coco, M. Leo, P. Carcagnì, P. Spagnolo, P. L. Mazzeo, M. Bernava, F. Marino, G. Pioggia, and C. Distante, “A Computer Vision Based Approach for Understanding Emotional Involvements in Children with Autism Spectrum Disorders,” in 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), 2017, pp. 1401-1407.

[28] K. Owada, M. Kojima, W. Yassin, M. Kuroda, Y. Kawakubo, H. Kuwabara, Y. Kano, and H. Yamasue, “Computer-analyzed facial expression as a surrogate marker for autism spectrum social core symptoms,” *PLOS ONE,* vol. 13, no. 1, pp. e0190442, 2018.

[29] H. Drimalla, N. Landwehr, I. Baskow, B. Behnia, S. Roepke, I. Dziobek, and T. Scheffer, “Detecting Autism by Analyzing a Simulated Social Interaction,” Cham, 2019, pp. 193-208.

[30] B. Li, S. Mehta, D. Aneja, C. Foster, P. Ventola, F. Shic, and L. Shapiro, “A Facial Affect Analysis System for Autism Spectrum Disorder,” in 2019 IEEE International Conference on Image Processing (ICIP), 2019, pp. 4549-4553.

[31] M. D. Samad, N. Diawara, J. L. Bobzien, J. W. Harrington, M. A. Witherow, and K. M. Iftekharuddin, “A Feasibility Study of Autism Behavioral Markers in Spontaneous Facial, Visual, and Hand Movement Response Data,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering,* vol. 26, no. 2, pp. 353-361, 2018.

[32] M. D. Samad, N. Diawara, J. L. Bobzien, C. M. Taylor, J. W. Harrington, and K. M. Iftekharuddin, “A pilot study to identify autism related traits in spontaneous facial actions using computer vision,” *Research in Autism Spectrum Disorders,* vol. 65, pp. 14-24, 2019.

[33] C. Grossard, A. Dapogny, D. Cohen, S. Bernheim, E. Juillet, F. Hamel, S. Hun, J. Bourgeois, H. Pellerin, S. Serret, K. Bailly, and L. Chaby, “Children with autism spectrum disorder produce more ambiguous and less socially meaningful facial expressions: an experimental study using random forest classifiers,” *Molecular Autism,* vol. 11, no. 1, pp. 5, 2020/01/13, 2020.

[34] D. A. Trevisan, M. Hoskyn, and E. Birmingham, “Facial Expression Production in Autism: A Meta-Analysis,” *Autism Res,* vol. 11, no. 12, pp. 1586-1601, Dec, 2018.

[35] T. Baltrušaitis, P. Robinson, and L.-P. Morency, “Openface: an open source facial behavior analysis toolkit,” in 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), 2016, pp. 1-10.

[36] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, “Openface 2.0: Facial behavior analysis toolkit,” in 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018), 2018, pp. 59-66.

[37] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett, “The computer expression recognition toolbox (CERT),” in 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG), 2011, pp. 298-305.

[38] Noldus Information Technology bv. "FaceReader," 07/27/2022, 2022; https://www.noldus.com/facereader.

[39] iMotions A/S. "Facial Expession Analysis," 07/27/2022, 2022; https://imotions.com/biosensor/fea-facial-expression-analysis/.

[40] P. Burke, and C. Hughes-Lawson, “The growth and development of the soft tissues of the human face,” *Journal of anatomy,* vol. 158, pp. 115, 1988.

[41] C. Grossard, L. Chaby, S. Hun, H. Pellerin, J. Bourgeois, A. Dapogny, H. Ding, S. Serret, P. Foulon, and M. Chetouani, “Children facial expression production: influence of age, gender, emotion subtype, elicitation condition and culture,” *Frontiers in psychology*, pp. 446, 2018.

[42] A. Dapogny, C. Grossard, S. Hun, S. Serret, O. Grynszpan, S. Dubuisson, D. Cohen, and K. Bailly, “On Automatically Assessing Children's Facial Expressions Quality: A Study, Database, and Protocol,” *Frontiers in Computer Science,* vol. 1, 2019-October-11, 2019.

[43] M. Witherow, M. Samad, and K. Iftekharuddin, “Transfer learning approach to multiclass classification of child facial expressions,” in SPIE Optical Engineering + Applications, 2019.

[44] M. Witherow, W. Shields, M. Samad, and K. Iftekharuddin, “Learning latent expression labels of child facial expression images through data-limited domain adaptation and transfer learning,” in SPIE Optical Engineering + Applications, 2020.

[45] Z. Zheng, X. Li, J. Barnes, C.-H. Park, and M. Jeon, “Facial Expression Recognition for Children: Can Existing Methods Tuned for Adults Be Adopted for Children?,” in International Conference on Human-Computer Interaction, Cham, 2019, pp. 201-211.

[46] A. Amodia-Bidakowska, C. Laverty, and P. G. Ramchandani, “Father-child play: A systematic review of its frequency, characteristics and potential impact on children’s development,” *Developmental Review,* vol. 57, pp. 100924, 2020.

[47] H. Chen, H. W. Park, and C. Breazeal, “Teaching and learning with children: Impact of reciprocal peer learning with a social robot on children’s learning and emotive engagement,” *Computers & Education,* vol. 150, pp. 103836, 2020.

[48] M. M. Terwogt, and H. Stegge, "Children's perspective on the emotional process," *The social child*, pp. 249-269: Psychology Press, 2021.

[49] P. Goldberg, Ö. Sümer, K. Stürmer, W. Wagner, R. Göllner, P. Gerjets, E. Kasneci, and U. Trautwein, “Attentive or not? Toward a machine learning approach to assessing students’ visible engagement in classroom instruction,” *Educational Psychology Review,* vol. 33, pp. 27-49, 2021.

[50] S. K. Gupta, T. Ashwin, and R. M. R. Guddeti, “Students’ affective content analysis in smart classroom environment using deep learning techniques,” *Multimedia Tools and Applications,* vol. 78, pp. 25321-25348, 2019.

[51] Ö. Sümer, P. Goldberg, S. D'Mello, P. Gerjets, U. Trautwein, and E. Kasneci, “Multimodal engagement analysis from facial videos in the classroom,” *IEEE Transactions on Affective Computing*, 2021.

[52] T. Hassan, D. Seuß, J. Wollenberg, K. Weitz, M. Kunz, S. Lautenbacher, J.-U. Garbas, and U. Schmid, “Automatic detection of pain from facial expressions: a survey,” *IEEE transactions on pattern analysis and machine intelligence,* vol. 43, no. 6, pp. 1815-1831, 2019.

[53] G. Zamzmi, R. Paul, D. Goldgof, R. Kasturi, and Y. Sun, “Pain assessment from facial expression: Neonatal convolutional neural network (N-CNN),” in 2019 International Joint Conference on Neural Networks (IJCNN), 2019, pp. 1-7.

[54] Z. Fei, E. Yang, D. D.-U. Li, S. Butler, W. Ijomah, X. Li, and H. Zhou, “Deep convolution network based emotion analysis towards mental health care,” *Neurocomputing,* vol. 388, pp. 212-227, 2020.

[55] C. Su, Z. Xu, J. Pathak, and F. Wang, “Deep learning in mental health outcome research: a scoping review,” *Translational Psychiatry,* vol. 10, no. 1, pp. 116, 2020.

[56] M. T. Akbar, M. N. Ilmi, I. V. Rumayar, J. Moniaga, T.-K. Chen, and A. Chowanda, “Enhancing game experience with facial expression recognition as dynamic balancing,” *Procedia Computer Science,* vol. 157, pp. 388-395, 2019.

[57] P. M. Blom, S. Bakkes, and P. Spronck, “Modeling and adjusting in-game difficulty based on facial expression analysis,” *Entertainment Computing,* vol. 31, pp. 100307, 2019.

[58] G. Guo, R. Guo, and X. Li, “Facial Expression Recognition Influenced by Human Aging,” *IEEE Transactions on Affective Computing,* vol. 4, no. 3, pp. 291-298, 2013.

[59] T. G. Rebanowako, A. R. Yadav, and R. Joshi, “Age-Invariant Facial Expression Classification Method Using Deep Learning,” in Proceedings of 6th International Conference on Recent Trends in Computing, Singapore, 2021, pp. 571-579.

[60] M. V. Birk, and R. L. Mandryk, “Improving the Efficacy of Cognitive Training for Digital Mental Health Interventions Through Avatar Customization: Crowdsourced Quasi-Experimental Study,” *J Med Internet Res,* vol. 21, no. 1, pp. e10133, 2019.

[61] R. S. Calabrò, A. Cerasa, I. Ciancarelli, L. Pignolo, P. Tonin, M. Iosa, and G. Morone, “The Arrival of the Metaverse in Neurorehabilitation: Fact, Fake or Vision?,” *Biomedicines,* vol. 10, no. 10, pp. 2602, 2022.

[62] J. Hao, H. Xie, K. Harp, Z. Chen, and K.-C. Siu, “Effects of virtual reality intervention on neural plasticity in stroke rehabilitation: a systematic review,” *Archives of Physical Medicine and Rehabilitation,* vol. 103, no. 3, pp. 523-541, 2022.

[63] N. Garcia-Hernandez, M. Guzman-Alvarado, and V. Parra-Vega, “Virtual body representation for rehabilitation influences on motor performance of cerebral palsy children,” *Virtual Reality,* vol. 25, pp. 669-680, 2021.

[64] A. R. Alashram, G. Annino, E. Padua, C. Romagnoli, and N. B. Mercuri, “Cognitive rehabilitation post traumatic brain injury: A systematic review for emerging use of virtual reality technology,” *Journal of Clinical Neuroscience,* vol. 66, pp. 209-219, 2019.

[65] I. Cikajlo, and K. Peterlin Potisk, “Advantages of using 3D virtual reality based training in persons with Parkinson’s disease: A parallel study,” *Journal of neuroengineering and rehabilitation,* vol. 16, no. 1, pp. 1-14, 2019.

[66] C. Frasson, and H. B. Abdessalem, “Contribution of Virtual Reality Environments and Artificial Intelligence for Alzheimer,” *Medical Research Archives,* vol. 10, no. 9, 2022.

[67] S. Mezrar, and F. Bendella, “A Systematic Review of Serious Games Relating to Cognitive Impairment and Dementia,” *Journal of Digital Information Management,* vol. 20, no. 1, pp. 1-9, 2022.

[68] P. M. G. Emmelkamp, K. Meyerbröker, and N. Morina, “Virtual Reality Therapy in Social Anxiety Disorder,” *Current Psychiatry Reports,* vol. 22, no. 7, pp. 32, 2020/05/13, 2020.

[69] A. Takemoto, “Depression detection using virtual avatar communication and eye tracking system,” *Journal of Eye Movement Research,* vol. 16, no. 2, 08/06, 2023.

[70] A. Takemoto, I. Aispuriete, L. Niedra, and L. F. Dreimane, “Differentiating depression using facial expressions in a virtual avatar communication system,” *Frontiers in Digital Health,* vol. 5, 2023-March-10, 2023.

[71] N. I. Muros, A. S. García, C. Forner, P. López-Arcas, G. Lahera, R. Rodriguez-Jimenez, K. N. Nieto, J. M. Latorre, A. Fernández-Caballero, and P. Fernández-Sotos, “Facial Affect Recognition by Patients with Schizophrenia Using Human Avatars,” *Journal of Clinical Medicine,* vol. 10, no. 9, pp. 1904, 2021.

[72] M. D. Samad, N. Diawara, J. L. Bobzien, J. W. Harrington, M. A. Witherow, and K. M. Iftekharuddin, “A feasibility study of autism behavioral markers in spontaneous facial, visual, and hand movement response data,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering,* vol. 26, no. 2, pp. 353-361, 2017.

[73] M. A. Mosher, A. C. Carreon, S. L. Craig, and L. C. Ruhter, “Immersive Technology to Teach Social Skills to Students with Autism Spectrum Disorder: a Literature Review,” *Review Journal of Autism and Developmental Disorders,* vol. 9, no. 3, pp. 334-350, 2022/09/01, 2022.

[74] M. S. Jaliaawala, and R. A. Khan, “Can autism be catered with artificial intelligence-assisted intervention technology? A comprehensive survey,” *Artificial Intelligence Review,* vol. 53, no. 2, pp. 1039-1069, 2020/02/01, 2020.

[75] M. C. Pino, R. Vagnetti, M. Valenti, and M. Mazza, “Comparing virtual vs real faces expressing emotions in children with autism: An eye-tracking study,” *Education and Information Technologies,* vol. 26, no. 5, pp. 5717-5732, 2021/09/01, 2021.

[76] M. Hotton, E. Huggons, C. Hamlet, D. Shore, D. Johnson, J. H. Norris, S. Kilcoyne, and L. Dalton, “The psychosocial impact of facial palsy: A systematic review,” *British Journal of Health Psychology,* vol. 25, no. 3, pp. 695-727, 2020.

[77] M. Shin, S. J. Kim, and F. Biocca, “The uncanny valley: No need for any further judgments when an avatar looks eerie,” *Computers in Human Behavior,* vol. 94, pp. 100-109, 2019/05/01/, 2019.

[78] H.-W. Lee, K. Chang, J.-P. Uhm, and E. Owiro, “How Avatar Identification Affects Enjoyment in the Metaverse: The Roles of Avatar Customization and Social Engagement,” *Cyberpsychology, Behavior, and Social Networking,* vol. 26, no. 4, pp. 255-262, 2023/04/01, 2023.

[79] P. Ekman, W. Friesen, and J. Hager, “Facial action coding system [E-book],” *Salt Lake City, UT: Research Nexus*, 2002.

[80] S. J. Webb, F. Shic, M. Murias, C. A. Sugar, A. J. Naples, E. Barney, H. Borland, G. Hellemann, S. Johnson, M. Kim, A. R. Levin, M. Sabatos-DeVito, M. Santhosh, D. Senturk, J. Dziura, R. A. Bernier, K. Chawarska, G. Dawson, S. Faja, S. Jeste, J. McPartland, t. A. B. C. f. C. T. , A. Atyabi, M. Aubertine, C. Carlos, S.-A. A. Chang, S. Compton, K. Dommer, A. Gateman, S. Hasselmo, B. Heit, T. Howell, A. Harris, K. Hutchins, J. Holub, B. Li, S. Major, S. Marsan, T. McAllister, A. S. M. Leal, L. Nanamaker, C. A. Nelson, H. Seow, D. Stahl, and A. Yuan, “Biomarker Acquisition and Quality Control for Multi-Site Studies: The Autism Biomarkers Consortium for Clinical Trials,” *Frontiers in Integrative Neuroscience,* vol. 13, 2020-February-07, 2020.

[81] F. Shic, A. J. Naples, E. C. Barney, S. A. Chang, B. Li, T. McAllister, M. Kim, K. J. Dommer, S. Hasselmo, A. Atyabi, Q. Wang, G. Helleman, A. R. Levin, H. Seow, R. Bernier, K. Charwaska, G. Dawson, J. Dziura, S. Faja, S. S. Jeste, S. P. Johnson, M. Murias, C. A. Nelson, M. Sabatos-DeVito, D. Senturk, C. A. Sugar, S. J. Webb, and J. C. McPartland, "The Autism Biomarkers Consortium for Clinical Trials: evaluation of a battery of candidate eye-tracking biomarkers for use in autism clinical trials," *Molecular Autism,* vol. 13, no. 1, pp. 15, 2022/03/21, 2022.

[82] A. Bangerter, M. Chatterjee, J. Manfredonia, N. V. Manyakov, S. Ness, M. A. Boice, A. Skalkin, M. S. Goodwin, G. Dawson, R. Hendren, B. Leventhal, F. Shic, and G. Pandina, "Automated recognition of spontaneous facial expression in individuals with autism spectrum disorder: parsing response variability," *Molecular Autism,* vol. 11, no. 1, pp. 31, 2020/05/11, 2020.

[83] E. Loth, L. Garrido, J. Ahmad, E. Watson, A. Duff, and B. Duchaine, "Facial expression recognition as a candidate marker for autism spectrum disorder: how frequent and severe are deficits?," *Molecular autism,* vol. 9, no. 1, pp. 1-11, 2018.

[84] H. Meyer-Lindenberg, C. Moessnang, B. Oakley, J. Ahmad, L. Mason, E. J. H. Jones, H. L. Hayward, J. Cooke, D. Crawley, R. Holt, J. Tillmann, T. Charman, S. Baron-Cohen, T. Banaschewski, C. Beckmann, H. Tost, A. Meyer-Lindenberg, J. K. Buitelaar, D. G. Murphy, M. J. Brammer, and E. Loth, "Facial expression recognition is linked to clinical and neurofunctional differences in autism," *Molecular Autism,* vol. 13, no. 1, pp. 43, 2022/11/10, 2022.

[85] J. Quinde-Zlibut, A. Munshi, G. Biswas, and C. J. Cascio, "Identifying and describing subtypes of spontaneous empathic facial expression production in autistic adults," *Journal of Neurodevelopmental Disorders,* vol. 14, no. 1, pp. 43, 2022/08/01, 2022.

[86] M. B. Kursa, A. Jankowski, and W. R. Rudnicki, "Boruta – A System for Feature Selection," *Fundamenta Informaticae,* vol. 101, pp. 271-285, 2010.

[87] M. A. Witherow, M. D. Samad, N. Diawara, H. Y. Bar, and K. M. Iftekharuddin, "Deep Adaptation of Adult-Child Facial Expressions by Fusing Landmark Features," *IEEE Transactions on Affective Computing*, pp. 1-12, 2023.

[88] M. A. Witherow, M. D. Samad, N. Diawara, and K. M. Iftekharuddin, "Facial landmark feature fusion in transfer learning of child facial expressions," in Proc.SPIE, 2022, pp. 122270P.

[89] M. A. Witherow, C. Butler, W. J. Shields, F. Ilgin, N. Diawara, J. Keener, J. W. Harrington, and K. M. Iftekharuddin, "Customizable Avatars with Dynamic Facial Action Coded Expressions (CADyFACE) for Improved User Engagement," *ArXiv preprint*, 2024.

[90] M. A. Witherow, N. Diawara, J. Keener, J. W. Harrington, and K. M. Iftekharuddin, "Pilot Study to Discover Candidate Biomarkers for Autism based on Perception and Production of Facial Expressions," *arXiv preprint*, 2024.

[91] C. M. Bishop, *Pattern Recognition and Machine Learning*, New York: Springer Science+Business Media, LLC, 2006.

[92] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2 ed., New York, NY: Springer, 2017.

[93] M. Espadoto, R. M. Martins, A. Kerren, N. S. T. Hirata, and A. C. Telea, “Toward a Quantitative Survey of Dimension Reduction Techniques,” *IEEE Transactions on Visualization and Computer Graphics,* vol. 27, no. 3, pp. 2153-2173, 2021.

[94] D. Xu, and Y. Tian, “A Comprehensive Survey of Clustering Algorithms,” *Annals of Data Science,* vol. 2, no. 2, pp. 165-193, 2015/06/01, 2015.

[95] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *Journal of Computational and Applied Mathematics,* vol. 20, pp. 53-65, 1987/11/01/, 1987.

[96] T. Caliński, and H. Ja, “A Dendrite Method for Cluster Analysis,” *Communications in Statistics - Theory and Methods,* vol. 3, pp. 1-27, 01/01, 1974.

[97] D. L. Davies, and D. W. Bouldin, “A Cluster Separation Measure,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. PAMI-1, no. 2, pp. 224-227, 1979.

[98] G. James, D. Witten, T. Hastie, and R. Tibshirani, *A Introduction to Statistical Learning with Applications in R*, 2 ed., New York: Springer, 2023.

[99] T. O. Hodson, “Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not,” *Geosci. Model Dev.,* vol. 15, no. 14, pp. 5481-5487, 2022.

[100] S. Shalev-Shwartz, and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, Cambridge: Cambrige University Press, 2014.

[101] A. P. Bradley, “The use of the area under the ROC curve in the evaluation of machine learning algorithms,” *Pattern Recognition,* vol. 30, no. 7, pp. 1145-1159, 1997/07/01/, 1997.

[102] R. A. Berk, *Statistical Learning from a Regression Perspective*, 3 ed., Cham, Switzerland: Springer Nature Switzerland, 2020.

[103] T. Miller, *Introduction to Reinforcement Learning*, Brisbane/Meaanjin, Australia: The University of Queensland, 2022.

[104] F. Chollet, *Deep Learning with Python*, Shelter Island, New York: Manning Publication Co., 2018.

[105] J. R. Quinlan, “Induction of decision trees,” *Machine Learning,* vol. 1, no. 1, pp. 81-106, 1986/03/01, 1986.

[106] L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, 1 ed., Routledge: Chapman and Hall/CRC, 1984.

[107] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, “Scikit-learn: Machine Learning in Python,” *J. Mach. Learn. Res.,* vol. 12, no. null, pp. 2825–2830, 2011.

[108] T. Hothorn, K. Hornik, and A. Zeileis, “ctree: Conditional inference trees,” *The comprehensive R archive network,* vol. 8, 2015.

[109] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, Cambridge, Massachusetts: MIT Press, 2017.

[110] S. R. Dubey, S. K. Singh, and B. B. Chaudhuri, “Activation functions in deep learning: A comprehensive survey and benchmark,” *Neurocomputing,* vol. 503, pp. 92-108, 2022.

[111] Q. Wang, Y. Ma, K. Zhao, and Y. Tian, “A comprehensive survey of loss functions in machine learning,” *Annals of Data Science*, pp. 1-26, 2020.

[112] R. Grosse, "Lecture 4: Training a Classifier," *CSC 321: Intro to Neural Networks and Machine Learning*,

https://www.cs.toronto.edu/~rgrosse/courses/csc321_2017/readings/L04%20Training%20a%20Classifier.pdf, 2017].

[113] D. P. Kingma, and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[114] G. Csurka, "A Comprehensive Survey on Domain Adaptation for Visual Applications," *Domain Adaptation in Computer Vision Applications*, G. Csurka, ed., pp. 1-35, Cham: Springer International Publishing, 2017.

[115] K. Zhang, M. Gong, P. Stojanov, B. Huang, Q. Liu, and C. Glymour, “Domain adaptation as a problem of inference on graphical models,” *Advances in neural information processing systems,* vol. 33, pp. 4965-4976, 2020.

[116] P. Viola, and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, 2001, pp. I-I.

[117] S. Minaee, P. Luo, Z. Lin, and K. Bowyer, “Going deeper into face detection: A survey,” *arXiv preprint arXiv:2103.14983*, 2021.

[118] S. A. Rizwan, A. Jalal, and K. Kim, “An Accurate Facial Expression Detector using Multi-Landmarks Selection and Local Transform Features,” in 2020 3rd International Conference on Advancements in Computational Sciences (ICACS), 2020, pp. 1-6.

[119] M. Murtaza, M. Sharif, M. AbdullahYasmin, and T. Ahmad, “Facial expression detection using Six Facial Expressions Hexagon (SFEH) model,” in 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC), 2019, pp. 0190-0195.

[120] A. Barman, and P. Dutta, “Influence of shape and texture features on facial expression recognition,” *IET Image Processing,* vol. 13, no. 8, pp. 1349-1363, 2019.

[121] K. X. Beh, and K. M. Goh, “Micro-Expression Spotting Using Facial Landmarks,” in 2019 IEEE 15th International Colloquium on Signal Processing & Its Applications (CSPA), 2019, pp. 192-197.

[122] V. Kazemi, and J. Sullivan, “One millisecond face alignment with an ensemble of regression trees,” in 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1867-1874.

[123] K. Baruni, N. Mokoena, M. Veeraragoo, and R. Holder, “Age Invariant Face Recognition Methods: A Review,” in 2021 International Conference on Computational Science and Computational Intelligence (CSCI), 2021, pp. 1657-1662.

[124] M. M. Sawant, and K. M. Bhurchandi, “Age invariant face recognition: a survey on facial aging databases, techniques and effect of aging,” *Artificial Intelligence Review,* vol. 52, no. 2, pp. 981-1008, 2019/08/01, 2019.

[125] P. Punyani, R. Gupta, and A. Kumar, “Neural networks for facial age estimation: a survey on recent advances,” *Artificial Intelligence Review,* vol. 53, no. 5, pp. 3299-3347, 2020/06/01, 2020.

[126] D. Kollias, “ABAW: Valence-Arousal Estimation, Expression Recognition, Action Unit Detection &amp; Multi-Task Learning Challenges,” in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2022, pp. 2327-2335.

[127] D. Kollias, and S. Zafeiriou, “Affect Analysis in-the-wild: Valence-Arousal, Expressions, Action Units and a Unified Framework,” *ArXiv,* vol. abs/2103.15792, 2021.

[128] S. Zafeiriou, D. Kollias, M. A. Nicolaou, A. Papaioannou, G. Zhao, and I. Kotsia, “Aff-wild: valence and arousal'In-the-Wild'challenge,” in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2017, pp. 34-41.

[129] P. Ekman, “An argument for basic emotions,” *Cognition and Emotion,* vol. 6, no. 3-4, pp. 169-200, 1992/05/01, 1992.

[130] P. Ekman, W. V. Friesen, and J. C. Hager, *Facial Action Coding System: Manual and Investigator's Guide*, Salt Lake City, UT, USA: Reseach Nexus, 2002.

[131] D. Kollias, V. Sharmanska, and S. Zafeiriou, “Face Behavior à la carte: Expressions, Affect and Action Units in a Single Network,” *ArXiv,* vol. abs/1910.11111, 2019.

[132] S.-J. Wang, B. Lin, Y. Wang, T. Yi, B. Zou, and X.-w. Lyu, “Action Units recognition based on Deep Spatial-Convolutional and Multi-label Residual network,” *Neurocomputing,* vol. 359, pp. 130-138, 2019/09/24/, 2019.

[133] A. Schall, and J. Romano Bergstrom, "1 - Introduction to Eye Tracking," *Eye Tracking in User Experience Design*, J. Romano Bergstrom and A. J. Schall, eds., pp. 3-26, Boston: Morgan Kaufmann, 2014.

[134] S. Bhattacharya, and M. Gupta, “A survey on: Facial emotion recognition invariant to pose, illumination and age,” in 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP), 2019, pp. 1-6.

[135] C. Dalvi, M. Rathod, S. Patil, S. Gite, and K. Kotecha, “A Survey of AI-Based Facial Emotion Recognition: Features, ML & DL Techniques, Age-Wise Datasets and Future Directions,” *IEEE Access,* vol. 9, pp. 165806-165840, 2021.

[136] T. Kanade, J. F. Cohn, and Y. Tian, “Comprehensive database for facial expression analysis,” in Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580), 2000, pp. 46-53.

[137] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, “The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression,” in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, 2010, pp. 94-101.

[138] V. LoBue, and C. Thrasher, "The Child Affective Facial Expression (CAFE) set. Databrary. 2014," 2014.

[139] V. LoBue, and C. Thrasher, “The Child Affective Facial Expression (CAFE) set: validity and reliability from untrained adults,” *Frontiers in Psychology,* vol. 5, 2015-January-06, 2015.

[140] J. G. Negrão, A. A. C. Osorio, R. F. Siciliano, V. R. G. Lederman, E. H. Kozasa, M. E. F. D'Antino, A. Tamborim, V. Santos, D. L. B. de Leucas, P. S. Camargo, D. C. Mograbi, T. P. Mecca, and J. S. Schwartzman, “The Child Emotion Facial Expression Set: A Database for Emotion Recognition in Children,” *Frontiers in Psychology,* vol. 12, 2021-April-29, 2021.

[141] R. A. Khan, A. Crenn, A. Meyer, and S. Bouakaz, “A novel database of children's spontaneous facial expressions (LIRIS-CSE),” *Image and Vision Computing,* vol. 83-84, pp. 61-69, 2019/03/01/, 2019.

[142] R. Angulu, J. R. Tapamo, and A. O. Adewumi, “Age estimation via face images: a survey,” *EURASIP Journal on Image and Video Processing,* vol. 2018, no. 1, pp. 42, 2018/06/06, 2018.

[143] R. Angulu, J. R. Tapamo, and A. O. Adewumi, "Age-Group Estimation Using Feature and Decision Level Fusion," *The Computer Journal,* vol. 62, no. 3, pp. 346-358, 2018.

[144] Z. Lou, F. Alnajar, J. M. Alvarez, N. Hu, and T. Gevers, "Expression-Invariant Age Estimation Using Structured Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 40, no. 2, pp. 365-375, 2018.

[145] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto, "Unified deep supervised domain adaptation and generalization," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 5715-5725.

[146] A. S. Osman Ali, V. Sagayan, A. M. Saeed, H. Ameen, and A. Aziz, "Age-invariant face recognition system using combined shape and texture features," *IET Biometrics,* vol. 4, no. 2, pp. 98-115, 2015.

[147] A. Juhong, and C. Pintavirooj, "Face recognition based on facial landmark detection," in 2017 10th Biomedical Engineering International Conference (BMEiCON), 2017, pp. 1-4.

[148] A. Chinnnaswamy, P. Kumar, and S. Aravind, "Age Group Estimation using Facial Features," *International Journal of Emerging Technologies in Computational and Applied Sciences*, 01/01, 2014.

[149] A. Srivastava, "Estimation of Age Groups based on Facial Features," *International Journal of Engineering and Technical Research,* vol. 7, pp. 115-121, 07/01, 2018.

[150] D. Gong, Z. Li, D. Lin, J. Liu, and X. Tang, "Hidden Factor Analysis for Age Invariant Face Recognition," in 2013 IEEE International Conference on Computer Vision, 2013, pp. 2872-2879.

[151] H. Li, H. Zou, and H. Hu, "Modified Hidden Factor Analysis for Cross-Age Face Recognition," *IEEE Signal Processing Letters,* vol. 24, no. 4, pp. 465-469, 2017.

[152] H. Bar, and M. T. Wells, "On Graphical Models and Convex Geometry," *Comput Stat Data Anal,* vol. 187, Nov, 2023.

[153] D. Kollias, "ABAW: Learning from Synthetic Data & Multi-task Learning Challenges," Cham, 2023, pp. 157-172.

[154] D. Kollias, A. Schulc, E. Hajiyev, and S. Zafeiriou, "Analysing Affective Behavior in the First ABAW 2020 Competition," in 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), 2020, pp. 637-643.

[155] D. Kollias, V. Sharmanska, and S. Zafeiriou, "Distribution Matching for Heterogeneous Multi-Task Learning: a Large-scale Face Study," *ArXiv,* vol. abs/2105.03790, 2021.

[156] D. Kollias, and S. Zafeiriou, "Expression, Affect, Action Unit Recognition: Aff-Wild2, Multi-Task Learning and ArcFace," *ArXiv,* vol. abs/1910.04855, 2019.

[157] D. Kollias, and S. Zafeiriou, "Analysing Affective Behavior in the second ABAW2 Competition," in 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2021, pp. 3645-3653.

[158] C. Grossard, S. p. Hun, A. Dapogny, E. Juillet, F. Hamel, H. Jean-Marie, J. r. m. Bourgeois, H. Pellerin, P. Foulon, S. Serret, O. Grynszpan, K. Bailly, and D. Cohen, "Teaching Facial Expression Production in Autism: The Serious Game JEMImE," *Creative Education,* vol. Vol.10No.11, pp. 20, 2019.

[159] H. Kumazaki, T. Muramatsu, Y. Yoshikawa, B. A. Corbett, Y. Matsumoto, H. Higashida, T. Yuhi, H. Ishiguro, M. Mimura, and M. Kikuchi, "Job interview training targeting nonverbal communication using an android robot for individuals with autism spectrum disorder," *Autism,* vol. 23, no. 6, pp. 1586-1595, 2019.

[160] W.-T. Tsai, I.-J. Lee, and C.-H. Chen, “Inclusion of third-person perspective in CAVE-like immersive 3D virtual reality role-playing games for social reciprocity training of children with an autism spectrum disorder,” *Universal Access in the Information Society,* vol. 20, pp. 375-389, 2021.

[161] E. M. Medica, “Give me a kiss! An integrative rehabilitative training program with motor imagery and mirror therapy for recovery of facial palsy,” *European journal of physical and rehabilitation medicine*, pp. 1-38, 2019.

[162] R. Okamoto, K. Adachi, and K. Mizukami, “[Effects of facial rehabilitation exercise on the mood, facial expressions, and facial muscle activities in patients with Parkinson's disease],” *Nihon Ronen Igakkai zasshi. Japanese journal of geriatrics,* vol. 56, no. 4, pp. 478-486, 2019, 2019.

[163] D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, and S. Zafeiriou, “Deep Affect Prediction in-the-Wild: Aff-Wild Database and Challenge, Deep Architectures, and Beyond,” *International Journal of Computer Vision,* vol. 127, pp. 907-929, 2018.

[164] T. Bendokat, R. Zimmermann, and P. A. Absil, “A Grassmann manifold handbook: basic geometry and computational aspects,” *Advances in Computational Mathematics,* vol. 50, no. 1, pp. 6, 2024/01/05, 2024.

[165] J. N. Kundu, A. R. Kulkarni, S. Bhambri, D. Mehta, S. A. Kulkarni, V. Jampani, and V. B. Radhakrishnan, "Balancing discriminability and transferability for source-free domain adaptation." pp. 11710-11728.

[166] S. Pei, J. Sun, S. Xiang, and G. Meng, “Domain Decorrelation with Potential Energy Ranking,” *arXiv preprint arXiv:2207.12194*, 2022.

[167] L. N. Smith, “Cyclical learning rates for training neural networks,” in 2017 IEEE winter conference on applications of computer vision (WACV), 2017, pp. 464-472.

[168] D. Kollias, “ABAW: learning from synthetic data & multi-task learning challenges,” in Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI, 2023, pp. 157-172.

[169] S. M. Lundberg, and S.-I. Lee, “A unified approach to interpreting model predictions,” in Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, 2017, pp. 4768–4777.

[170] G. Erion, J. D. Janizek, P. Sturmfels, S. M. Lundberg, and S.-I. Lee, “Improving performance of deep learning models with axiomatic attribution priors and expected gradients,” *Nature Machine Intelligence,* vol. 3, no. 7, pp. 620-631, 2021/07/01, 2021.

[171] K. Zhao, W. S. Chu, and H. Zhang, “Deep Region and Multi-label Learning for Facial Action Unit Detection,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3391-3399.

[172] Z. Wang, Y. Li, S. Wang, and Q. Ji, “Capturing Global Semantic Relationships for Facial Action Unit Recognition,” in 2013 IEEE International Conference on Computer Vision, 2013, pp. 3304-3311.

[173] K. Zhao, W. S. Chu, F. D. l. Torre, J. F. Cohn, and H. Zhang, “Joint Patch and Multi-label Learning for Facial Action Unit and Holistic Expression Recognition,” *IEEE Transactions on Image Processing,* vol. 25, no. 8, pp. 3931-3946, 2016.

[174] Y. Song, D. McDuff, D. Vasisht, and A. Kapoor, “Exploiting sparsity and co-occurrence structure for action unit recognition,” in 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2015, pp. 1-8.

[175] C. Ma, L. Chen, and J. Yong, “AU R-CNN: Encoding expert prior knowledge into R-CNN for action unit detection,” *Neurocomputing,* vol. 355, pp. 35-47, 2019/08/25/, 2019.

[176] Z. Shao, Z. Liu, J. Cai, and L. Ma, “JÂA-Net: Joint Facial Action Unit Detection and Face Alignment Via Adaptive Attention,” *International Journal of Computer Vision,* vol. 129, no. 2, pp. 321-340, 2021/02/01, 2021.

[177] J. Chen, C. Wang, K. Wang, and M. Liu, “Lightweight network architecture using difference saliency maps for facial action unit detection,” *Applied Intelligence,* vol. 52, no. 6, pp. 6354-6375, 2022/04/01, 2022.

[178] J. C. McPartland, R. A. Bernier, S. S. Jeste, G. Dawson, C. A. Nelson, K. Chawarska, R. Earl, S. Faja, S. P. Johnson, L. Sikich, C. A. Brandt, J. D. Dziura, L. Rozenblit, G. Hellemann, A. R. Levin, M. Murias, A. J. Naples, M. L. Platt, M. Sabatos-DeVito, F. Shic, D. Senturk, C. A. Sugar, S. J. Webb, and t. A. B. C. f. C. T. , “The Autism Biomarkers Consortium for Clinical Trials (ABC-CT): Scientific Context, Study Design, and Progress Toward Biomarker Qualification,” *Frontiers in Integrative Neuroscience,* vol. 14, 2020-April-09, 2020.

[179] M. J. Dechant, M. V. Birk, Y. Shiban, K. Schnell, and R. L. Mandryk, “How Avatar Customization Affects Fear in a Game-based Digital Exposure Task for Social Anxiety,” *Proc. ACM Hum.-Comput. Interact.,* vol. 5, no. CHI PLAY, pp. Article 248, 2021.

[180] R. Zhu, and C. Yi, “Avatar design in Metaverse: the effect of avatar-user similarity in procedural and creative tasks,” *Internet Research,* vol. ahead-of-print, no. ahead-of-print, 2023.

[181] R. Cuthbert, S. Turkay, and R. Brown, “The Effects of Customisation on Player Experiences and Motivation in a Virtual Reality Game,” in Proceedings of the 31st Australian Conference on Human-Computer-Interaction, Fremantle, WA, Australia, 2020, pp. 221–232.

[182] J. Koulouris, Z. Jeffery, J. Best, E. O'neill, and C. Lutteroth, “Me vs. Super (wo) man: Effects of Customization and Identification in a VR Exergame,” in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020, pp. 1-17.

[183] P. Salehi, S. Z. Hassan, S. S. Sabet, G. A. Baugerud, M. S. Johnson, P. Halvorsen, and M. A. Riegler, “Is More Realistic Better? A Comparison of Game Engine and GAN-based Avatars for Investigative Interviews of Children,” in Proceedings of the 3rd ACM Workshop on Intelligent Cross-Data Analysis and Retrieval, Newark, NJ, USA, 2022, pp. 41–49.

[184] M. K. Young, J. J. Rieser, and B. Bodenheimer, “Dyadic interactions with avatars in immersive virtual environments: high fiving,” in Proceedings of the ACM SIGGRAPH Symposium on Applied Perception, Tübingen, Germany, 2015, pp. 119–126.

[185] J. A. Caine, B. Klein, and S. L. Edwards, “The Impact of a Novel Mimicry Task for Increasing Emotion Recognition in Adults with Autism Spectrum Disorder and Alexithymia: Protocol for a Randomized Controlled Trial,” *JMIR Res Protoc,* vol. 10, no. 6, pp. e24543, Jun 25, 2021.

[186] M. Olszanowski, G. Pochwatko, K. Kuklinski, M. Scibor-Rylski, P. Lewinski, and R. K. Ohme, “Warsaw set of emotional facial expression pictures: a validation study of facial display photographs,” *Front Psychol,* vol. 5, pp. 1516, 2014.

[187] S. Y. Yao, R. Bull, K. H. Khng, and A. Rahim, “Psychometric properties of the NEPSY-II affect recognition subtest in a preschool sample: a Rasch modeling approach,” *Clin Neuropsychol,* vol. 32, no. 1, pp. 63-80, Jan, 2018.

[188] M. M. Vandewouw, E. J. Choi, C. Hammill, J. P. Lerch, E. Anagnostou, and M. J. Taylor, “Changing Faces: Dynamic Emotional Face Processing in Autism Spectrum Disorder Across Childhood and Adulthood,” *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging,* vol. 6, no. 8, pp. 825-836, 2021/08/01/, 2021.

[189] A. J. d. L. Bomfim, R. A. d. S. Ribeiro, and M. H. N. Chagas, “Recognition of dynamic and static facial expressions of emotion among older adults with major depression,” *Trends in psychiatry and psychotherapy,* vol. 41, pp. 159-166, 2019.

[190] E. De Stefani, M. Ardizzi, Y. Nicolini, M. Belluardo, A. Barbot, C. Bertolini, G. Garofalo, B. Bianchi, G. Coudé, L. Murray, and P. F. Ferrari, “Children with facial paralysis due to Moebius syndrome exhibit reduced autonomic modulation during emotion processing,” *Journal of Neurodevelopmental Disorders,* vol. 11, no. 1, pp. 12, 2019/07/10, 2019.

[191] G. Kim, S. Park, and S. H. Lee, “Video Synthesis Method for Virtual Avatar Using FACS based GAN,” in Proceedings of the Korea Information Processing Society Conference, 2021, pp. 340-342.

[192] S. v. d. Struijk, H.-H. Huang, M. S. Mirzaei, and T. Nishida, “FACSvatar: An Open Source Modular Framework for Real-Time FACS based Facial Animation,” in Proceedings of the 18th International Conference on Intelligent Virtual Agents, Sydney, NSW, Australia, 2018, pp. 159–164.

[193] C. Butler, S. Michalowicz, L. Subramanian, and W. Burleson, “More than a Feeling: The MiFace Framework for Defining Facial Communication Mappings,” in Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology, Québec City, QC, Canada, 2017, pp. 773–786.

[194] R. Amini, C. Lisetti, and G. Ruiz, “HapFACS 3.0: FACS-Based Facial Expression Generator for 3D Speaking Virtual Characters,” *IEEE Transactions on Affective Computing,* vol. 6, no. 4, pp. 348-360, 2015.

[195] E. B. Roesch, L. Tamarit, L. Reveret, D. Grandjean, D. Sander, and K. R. Scherer, “FACSGen: A Tool to Synthesize Emotional Facial Expressions Through Systematic Manipulation of Facial Action Units,” *Journal of Nonverbal Behavior,* vol. 35, no. 1, pp. 1-16, 2011/03/01, 2011.

[196] M. Gilbert, S. Demarchi, and I. Urdapilleta, “FACSHuman, a software program for creating experimental material by modeling 3D facial expressions,” *Behavior Research Methods,* vol. 53, no. 5, pp. 2252-2272, 2021/10/01, 2021.

[197] A. S. García, P. Fernández-Sotos, M. A. Vicente-Querol, G. Lahera, R. Rodriguez-Jimenez, and A. Fernández-Caballero, “Design of reliable virtual human facial expressions and validation by healthy people,” *Integrated Computer-Aided Engineering,* vol. 27, pp. 287-299, 2020.

[198] D. Kollias, “ABAW: Learning from Synthetic Data & Multi-task Learning Challenges,” in Computer Vision – ECCV 2022 Workshops, Cham, 2023, pp. 157-172.

[199] M. Bishay, and I. Patras, "Fusing Multilabel Deep Networks for Facial Action Unit Detection," in 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), 2017, pp. 681-688.

[200] G. B. Dell'Olio, and M. Sra, "FaraPy: An Augmented Reality Feedback System for Facial Paralysis using Action Unit Intensity Estimation," in The 34th Annual ACM Symposium on User Interface Software and Technology, Virtual Event, USA, 2021, pp. 1027–1038.

[201] T. Guha, Z. Yang, A. Ramakrishna, R. B. Grossman, D. Hedley, S. Lee, and S. S. Narayanan, "On quantifying facial expression-related atypicality of children with Autism Spectrum Disorder," in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 803-807.

[202] M. D. Samad, J. L. Bobzien, J. W. Harrington, and K. M. Iftekharuddin, "Analysis of facial muscle activation in children with autism using 3D imaging," in 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2015, pp. 337-342.

[203] M. D. Samad, J. L. Bobzien, J. W. Harrington, and K. M. Iftekharuddin, "[INVITED] Non-intrusive optical imaging of face to probe physiological traits in Autism Spectrum Disorder," *Optics & Laser Technology,* vol. 77, pp. 221-228, 2016/03/01/, 2016.

[204] P. Frankl, and H. Maehara, "Some geometric applications of the beta distribution," *Annals of the Institute of Statistical Mathematics,* vol. 42, no. 3, pp. 463-474, 1990/09/01, 1990.

[205] T. Kanade, J. F. Cohn, and T. Yingli, "Comprehensive database for facial expression analysis," in Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580), 2000, pp. 46-53.

[206] M. Mavadati, P. Sanger, and M. H. Mahoor, "Extended disfa dataset: Investigating posed and spontaneous facial expressions," in proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2016, pp. 1-8.

[207] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "Disfa: A spontaneous facial action intensity database," *IEEE Transactions on Affective Computing,* vol. 4, no. 2, pp. 151-160, 2013.

[208] A. Iliev, N. Kyurkchiev, and S. Markov, "On the approximation of the step function by some sigmoid functions," *Mathematics and Computers in Simulation,* vol. 133, pp. 223-234, 2017/03/01/, 2017.

[209] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," *Advances in neural information processing systems,* vol. 12, 1999.

[210] M. R. Rezaei-Dastjerdehei, A. Mijani, and E. Fatemizadeh, "Addressing Imbalance in Multi-Label Classification Using Weighted Cross Entropy Loss Function," in 2020 27th National and 5th International Iranian Conference on Biomedical Engineering (ICBME), 2020, pp. 333-338.

[211] A. H. Mostafa, H. Abdel-Galil, and M. Belal, "Ensemble Model-based Weighted Categorical Cross-entropy Loss for Facial Expression Recognition," in 2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS), 2021, pp. 165-171.

[212] A. Papoutsaki, P. Sangkloy, J. Laskey, N. Daskalova, J. Huang, and J. Hays, "Webgazer: scalable webcam eye tracking using user interactions," in Proceedings of the Twenty-

Fifth International Joint Conference on Artificial Intelligence, New York, New York, USA, 2016, pp. 3839–3845.

[213] A. Acharjee, J. Larkman, Y. Xu, V. R. Cardoso, and G. V. Gkoutos, "A random forest based biomarker discovery and power analysis framework for diagnostics research," *BMC Medical Genomics,* vol. 13, no. 1, pp. 178, 2020/11/23, 2020.

[214] F. Hamidi, N. Gilani, R. Arabi Belaghi, H. Yaghoobi, E. Babaei, P. Sarbakhsh, and J. Malakouti, "Identifying potential circulating miRNA biomarkers for the diagnosis and prediction of ovarian cancer using machine-learning approach: application of Boruta," *Frontiers in Digital Health,* vol. 5, 2023-August-09, 2023.

[215] F. Shic, E. C. Barney, A. J. Naples, K. J. Dommer, S. A. Chang, B. Li, T. McAllister, A. Atyabi, Q. Wang, and R. Bernier, "The Selective Social Attention task in children with autism spectrum disorder: Results from the Autism Biomarkers Consortium for Clinical Trials (ABC-CT) feasibility study," *Autism Research,* vol. 16, no. 11, pp. 2150-2159, 2023.

[216] P. Feliciano, A. M. Daniels, L. G. Snyder, A. Beaumont, A. Camba, A. Esler, A. G. Gulsrud, A. Mason, A. Gutierrez, and A. Nicholson, "SPARK: A US cohort of 50,000 families to accelerate autism research," *Neuron,* vol. 97, no. 3, pp. 488-493, 2018.

[217] H. Drimalla, I. Baskow, B. Behnia, S. Roepke, and I. Dziobek, "Imitation and recognition of facial emotions in autism: a computer vision approach," *Molecular Autism,* vol. 12, no. 1, pp. 27, 2021/04/06, 2021.

[218] D. J. Faso, N. J. Sasson, and A. E. Pinkham, "Evaluating Posed and Evoked Facial Expressions of Emotion from Adults with Autism Spectrum Disorder," *Journal of Autism and Developmental Disorders,* vol. 45, no. 1, pp. 75-89, 2015/01/01, 2015.

[219] M. A. Volker, C. Lopata, D. A. Smith, and M. L. Thomeer, "Facial encoding of children with high-functioning autism spectrum disorders," *Focus on Autism and Other Developmental Disabilities,* vol. 24, no. 4, pp. 195-204, 2009.

[220] T. Guha, Z. Yang, A. Ramakrishna, R. B. Grossman, H. Darren, S. Lee, and S. S. Narayanan, "On Quantifying Facial Expression-Related Atypicality of Children with Autism Spectrum Disorder," *Proc IEEE Int Conf Acoust Speech Signal Process,* vol. 2015, pp. 803-807, Apr, 2015.

[221] M. D. Samad, J. L. Bobzien, J. W. Harrington, and K. M. Iftekharuddin, "Non-intrusive optical imaging of face to probe physiological traits in autism spectrum disorder," *Optics & Laser Technology,* vol. 77, pp. 221-228, 2016.

[222] H. C. Cuve, Y. Gao, and A. Fuse, "Is it avoidance or hypoarousal? A systematic review of emotion recognition, eye-tracking, and psychophysiological studies in young adults with autism spectrum conditions," *Research in Autism Spectrum Disorders,* vol. 55, pp. 1-13, 2018/11/01/, 2018.

[223] Q. Su, F. Chen, H. Li, N. Yan, and L. Wang, "Multimodal Emotion Perception in Children with Autism Spectrum Disorder by Eye Tracking Study," in 2018 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES), 2018, pp. 382-387.

[224] V. Tsang, "Eye-tracking study on facial emotion recognition tasks in individuals with high-functioning autism spectrum disorders," *Autism,* vol. 22, no. 2, pp. 161-170, 2018.

[225] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. P. Morency, "OpenFace 2.0: Facial Behavior Analysis Toolkit," in 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), 2018, pp. 59-66.

[226] R. O. Kellems, C. T. Charlton, B. Black, H. Bussey, R. Ferguson, B. F. Gonçalves, M. Jensen, and S. Vallejo, "Social Engagement of Elementary-Aged Children With Autism Live Animation Avatar Versus Human Interaction," *Journal of Special Education Technology,* vol. 38, no. 3, pp. 327-339, 2023.

[227] C. Putnam, C. Hanschke, J. Todd, J. Gemmell, and M. Kollia, "Interactive Technologies Designed for Children with Autism: Reports of Use and Desires from Parents, Teachers, and Therapists," *ACM Trans. Access. Comput.,* vol. 12, no. 3, pp. Article 12, 2019.

[228] L. Bozgeyikli, A. Raij, S. Katkoori, and R. Alqasemi, "A Survey on Virtual Reality for Individuals with Autism Spectrum Disorder: Design Considerations," *IEEE Transactions on Learning Technologies,* vol. 11, no. 2, pp. 133-151, 2018.

[229] E. Kinnaird, C. Stewart, and K. Tchanturia, "Investigating alexithymia in autism: A systematic review and meta-analysis," *Eur Psychiatry,* vol. 55, pp. 80-89, Jan, 2019.

[230] K. S. Goerlich, "The Multifaceted Nature of Alexithymia – A Neuroscientific Perspective," *Frontiers in Psychology,* vol. 9, 2018-August-29, 2018.

[231] Simons Foundation. "About SPARK," 03/24/2024, 2024; https://sparkforautism.org/portal/page/about-spark/.

[232] E. Pellicano, D. Adams, L. Crane, C. Hollingue, C. Allen, K. Almendinger, M. Botha, T. Haar, S. K. Kapp, and E. Wheeley, "Letter to the Editor: A possible threat to data integrity for online qualitative autism research," *Autism,* vol. 28, no. 3, pp. 786-792, 2024.

[233] E. Fombonne, L. Coppola, S. Mastel, and B. J. O'Roak, "Validation of Autism Diagnosis and Clinical Data in the SPARK Cohort," *Journal of Autism and Developmental Disorders,* vol. 52, no. 8, pp. 3383-3398, 2022/08/01, 2022.

[234] V. Gibbs, R. Y. Cai, F. Aldridge, and M. Wong, "Autism assessment via telehealth during the Covid 19 pandemic: Experiences and perspectives of autistic adults, parents/carers and clinicians," *Research in Autism Spectrum Disorders,* vol. 88, pp. 101859, 2021/10/01/, 2021.

[235] M. A. Witherow, C. Butler, A. Stedman, W. Shields, F. Ilgin, N. Diawara, J. Keener, J. W. Harrington, and K. M. Iftekharuddin, "Customizable Avatars with Dynamic Facial Action Coded Expressions (CADyFACE) for Improved User Engagement," *ArXiv*, 2024.

[236] M. A. Witherow, W. J. Shields, M. D. Samad, and K. M. Iftekharuddin, "Learning latent expression labels of child facial expression images through data-limited domain adaptation and transfer learning," in Applications of Machine Learning 2020, 2020, pp. 67-75.

[237] Y. Ganin, and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in International conference on machine learning, Lille, France, 2015, pp. 1180-1189.

[238] H. Kaneko, "Examining variable selection methods for the predictive performance of regression models and the proportion of selected variables and selected random variables," *Heliyon,* vol. 7, no. 6, pp. e07356, 2021/06/01/, 2021.

[239] S. van Buuren, and K. Groothuis-Oudshoorn, "mice: Multivariate Imputation by Chained Equations in R," *Journal of Statistical Software,* vol. 45, no. 3, pp. 1 - 67, 12/12, 2011.

[240] E. Slade, and M. G. Naylor, "A fair comparison of tree-based and parametric methods in multiple imputation by chained equations," *Statistics in Medicine,* vol. 39, no. 8, pp. 1156-1166, 2020.

[241] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, “Missing value estimation methods for DNA microarrays,” *Bioinformatics,* vol. 17, no. 6, pp. 520-525, 2001.

[242] M. Toure, N. Klutse, M. Sarr, A. Kenne, M. A. E. Bhuiyan, O. Ndiaye, B. Daouda, W. Thiaw, I. Sy, C. Mbow, M. Sall, and A. T. Gaye, “A New Multiple Imputation Approach Using Machine Learning to Enhance Climate Databases in Senegal,” *ResearchGate,* 2023.

[243] I. F. Way, B. Applegate, X. Cai *, L. K. Franck, C. Black-Pond, P. Yelsma, E. Roberts, Y. Hyter, and M. Muliett, “Children's Alexithymia Measure (CAM): A New Instrument for Screening Difficulties with Emotional Expression,” *Journal of Child & Adolescent Trauma,* vol. 3, no. 4, pp. 303-318, 2010/11/16, 2010.

[244] H. C. Vorst, and B. Bermond, “Validity and reliability of the Bermond–Vorst alexithymia questionnaire,” *Personality and individual differences,* vol. 30, no. 3, pp. 413-434, 2001.

[245] C. Grossard, L. Chaby, S. Hun, H. Pellerin, J. Bourgeois, A. Dapogny, H. Ding, S. Serret, P. Foulon, M. Chetouani, L. Chen, K. Bailly, O. Grynszpan, and D. Cohen, “Children Facial Expression Production: Influence of Age, Gender, Emotion Subtype, Elicitation Condition and Culture,” *Frontiers in Psychology,* vol. 9, 2018-April-04, 2018.

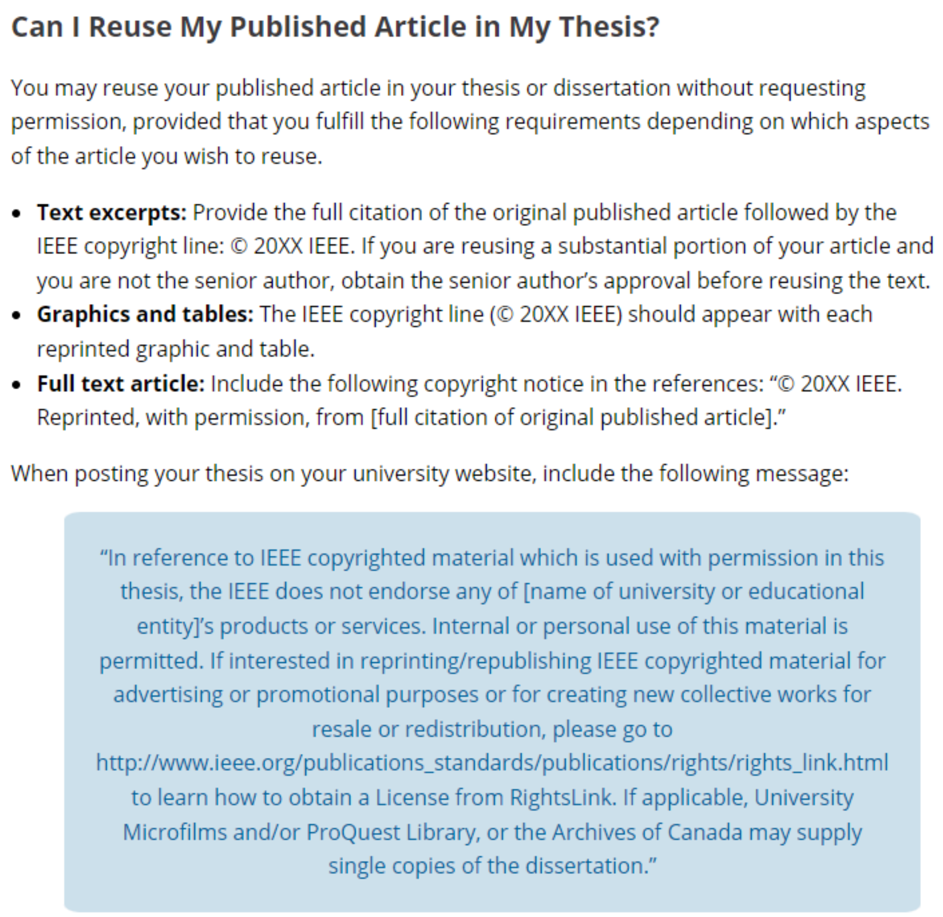
[246] C. T. Keating, D. S. Fraser, S. Sowden, and J. L. Cook, “Differences Between Autistic and Non-Autistic Adults in the Recognition of Anger from Facial Motion Remain after Controlling for Alexithymia,” *Journal of Autism and Developmental Disorders,* vol. 52, no. 4, pp. 1855-1871, 2022/04/01, 2022.

APPENDICES

APPENDIX A. IEEE COPYRIGHT NOTICE

IEEE Author Center Guidance on Reuse of Material from Published Articles

(https://journals.ieeeauthorcenter.ieee.org/choose-a-publishing-agreement/avoid-infringement-upon-ieee-copyright/)



Can I Reuse My Published Article in My Thesis?

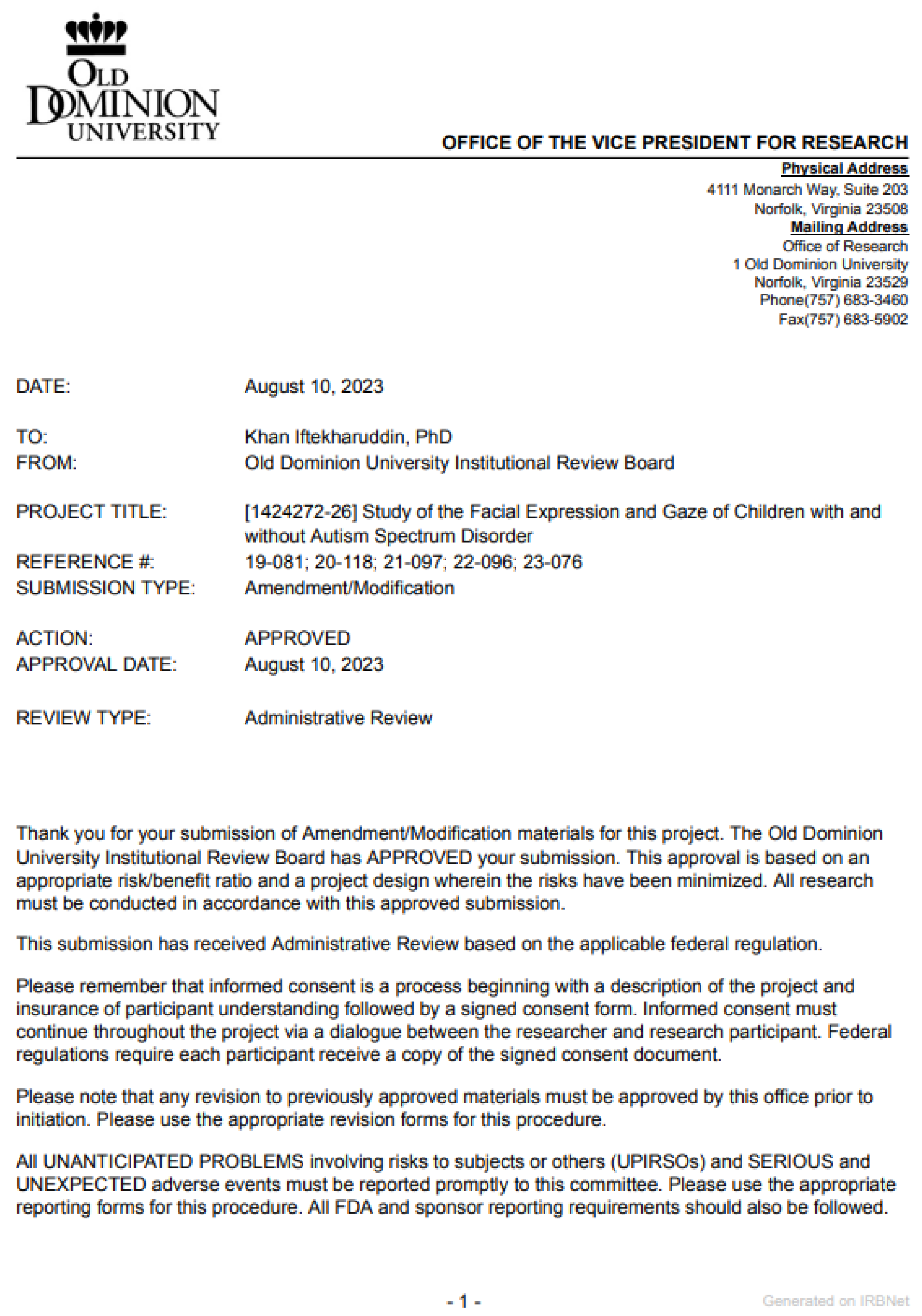
You may reuse your published article in your thesis or dissertation without requesting permission, provided that you fulfill the following requirements depending on which aspects of the article you wish to reuse.

- **Text excerpts:** Provide the full citation of the original published article followed by the IEEE copyright line: © 20XX IEEE. If you are reusing a substantial portion of your article and you are not the senior author, obtain the senior author's approval before reusing the text.
- **Graphics and tables:** The IEEE copyright line (© 20XX IEEE) should appear with each reprinted graphic and table.
- **Full text article:** Include the following copyright notice in the references: "© 20XX IEEE. Reprinted, with permission, from [full citation of original published article]."

When posting your thesis on your university website, include the following message:

> "In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [name of university or educational entity]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink. If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation."

APPENDIX B. ODU IRB APPROVAL LETTER



OLD DOMINION UNIVERSITY

OFFICE OF THE VICE PRESIDENT FOR RESEARCH

Physical Address
4111 Monarch Way, Suite 203
Norfolk, Virginia 23508
Mailing Address
Office of Research
1 Old Dominion University
Norfolk, Virginia 23529
Phone(757) 683-3460
Fax(757) 683-5902

DATE:	August 10, 2023
TO:	Khan Iftekharuddin, PhD
FROM:	Old Dominion University Institutional Review Board
PROJECT TITLE:	[1424272-26] Study of the Facial Expression and Gaze of Children with and without Autism Spectrum Disorder
REFERENCE #:	19-081; 20-118; 21-097; 22-096; 23-076
SUBMISSION TYPE:	Amendment/Modification
ACTION:	APPROVED
APPROVAL DATE:	August 10, 2023
REVIEW TYPE:	Administrative Review

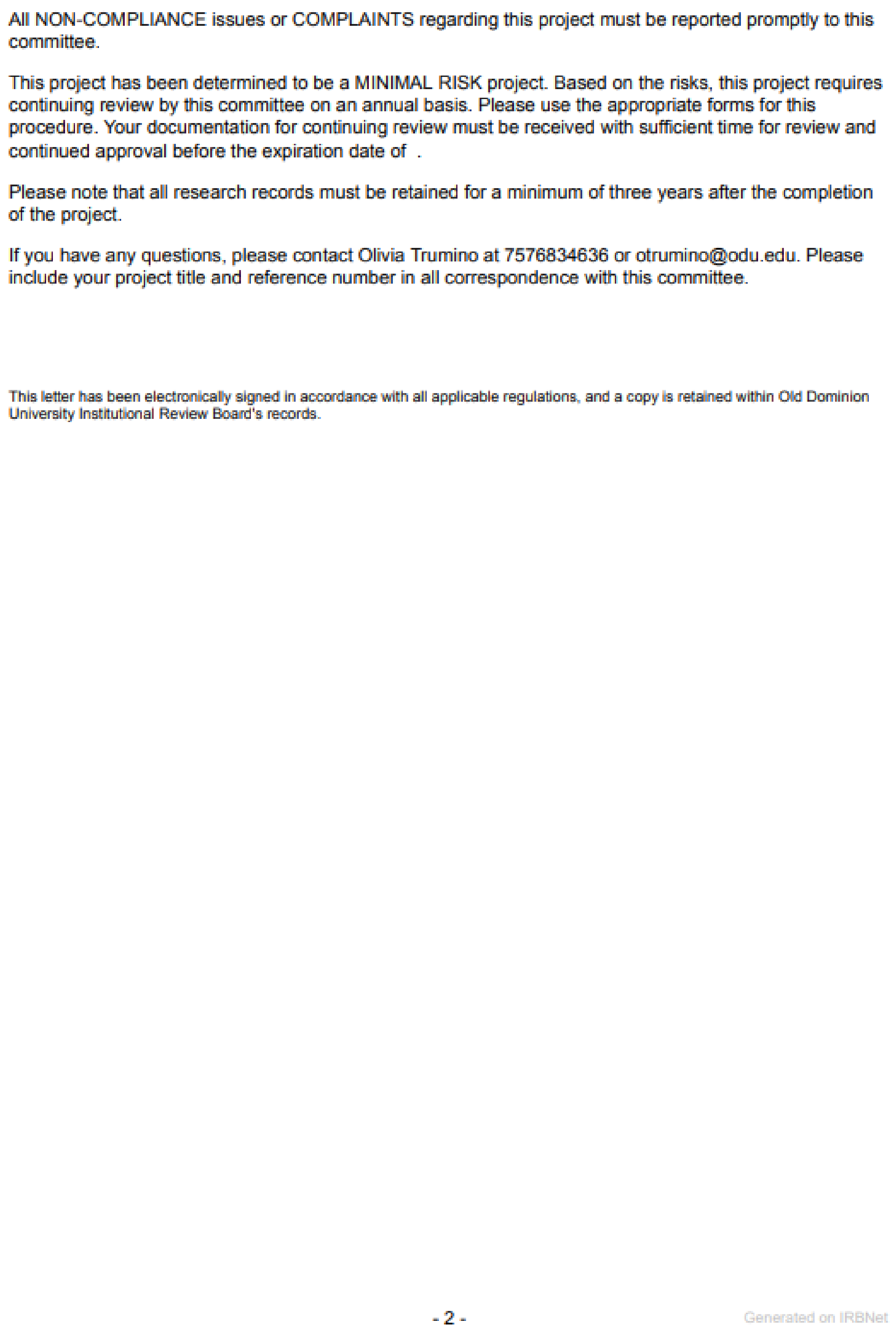
Thank you for your submission of Amendment/Modification materials for this project. The Old Dominion University Institutional Review Board has APPROVED your submission. This approval is based on an appropriate risk/benefit ratio and a project design wherein the risks have been minimized. All research must be conducted in accordance with this approved submission.

This submission has received Administrative Review based on the applicable federal regulation.

Please remember that informed consent is a process beginning with a description of the project and insurance of participant understanding followed by a signed consent form. Informed consent must continue throughout the project via a dialogue between the researcher and research participant. Federal regulations require each participant receive a copy of the signed consent document.

Please note that any revision to previously approved materials must be approved by this office prior to initiation. Please use the appropriate revision forms for this procedure.

All UNANTICIPATED PROBLEMS involving risks to subjects or others (UPIRSOs) and SERIOUS and UNEXPECTED adverse events must be reported promptly to this committee. Please use the appropriate reporting forms for this procedure. All FDA and sponsor reporting requirements should also be followed.



All NON-COMPLIANCE issues or COMPLAINTS regarding this project must be reported promptly to this committee.

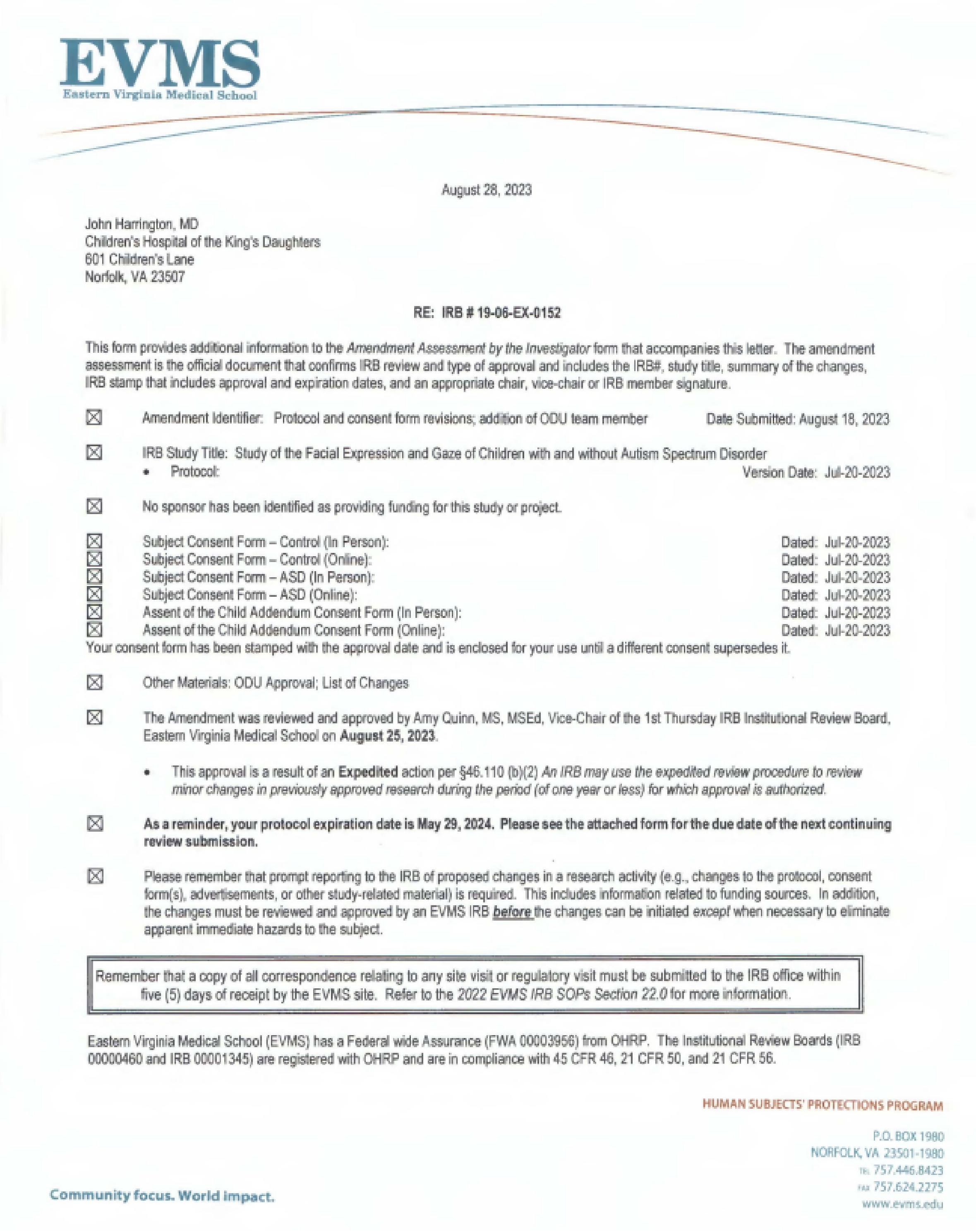
This project has been determined to be a MINIMAL RISK project. Based on the risks, this project requires continuing review by this committee on an annual basis. Please use the appropriate forms for this procedure. Your documentation for continuing review must be received with sufficient time for review and continued approval before the expiration date of .

Please note that all research records must be retained for a minimum of three years after the completion of the project.

If you have any questions, please contact Olivia Trumino at 7576834636 or otrumino@odu.edu. Please include your project title and reference number in all correspondence with this committee.

This letter has been electronically signed in accordance with all applicable regulations, and a copy is retained within Old Dominion University Institutional Review Board's records.

APPENDIX C. EVMS IRB APPROVAL LETTER



EVMS
Eastern Virginia Medical School

August 28, 2023

John Harrington, MD
Children's Hospital of the King's Daughters
601 Children's Lane
Norfolk, VA 23507

RE: IRB # 19-06-EX-0152

This form provides additional information to the *Amendment Assessment by the Investigator* form that accompanies this letter. The amendment assessment is the official document that confirms IRB review and type of approval and includes the IRB#, study title, summary of the changes, IRB stamp that includes approval and expiration dates, and an appropriate chair, vice-chair or IRB member signature.

☒ Amendment Identifier: Protocol and consent form revisions; addition of ODU team member Date Submitted: August 18, 2023

☒ IRB Study Title: Study of the Facial Expression and Gaze of Children with and without Autism Spectrum Disorder

- Protocol: Version Date: Jul-20-2023

☒ No sponsor has been identified as providing funding for this study or project.

☒	Subject Consent Form – Control (In Person):	Dated: Jul-20-2023
☒	Subject Consent Form – Control (Online):	Dated: Jul-20-2023
☒	Subject Consent Form – ASD (In Person):	Dated: Jul-20-2023
☒	Subject Consent Form – ASD (Online):	Dated: Jul-20-2023
☒	Assent of the Child Addendum Consent Form (In Person):	Dated: Jul-20-2023
☒	Assent of the Child Addendum Consent Form (Online):	Dated: Jul-20-2023

Your consent form has been stamped with the approval date and is enclosed for your use until a different consent supersedes it.

☒ Other Materials: ODU Approval; List of Changes

☒ The Amendment was reviewed and approved by Amy Quinn, MS, MSEd, Vice-Chair of the 1st Thursday IRB Institutional Review Board, Eastern Virginia Medical School on **August 25, 2023**.

- This approval is a result of an **Expedited** action per §46.110 (b)(2) *An IRB may use the expedited review procedure to review minor changes in previously approved research during the period (of one year or less) for which approval is authorized.*

☒ **As a reminder, your protocol expiration date is May 29, 2024. Please see the attached form for the due date of the next continuing review submission.**

☒ Please remember that prompt reporting to the IRB of proposed changes in a research activity (e.g., changes to the protocol, consent form(s), advertisements, or other study-related material) is required. This includes information related to funding sources. In addition, the changes must be reviewed and approved by an EVMS IRB ***before*** the changes can be initiated *except* when necessary to eliminate apparent immediate hazards to the subject.

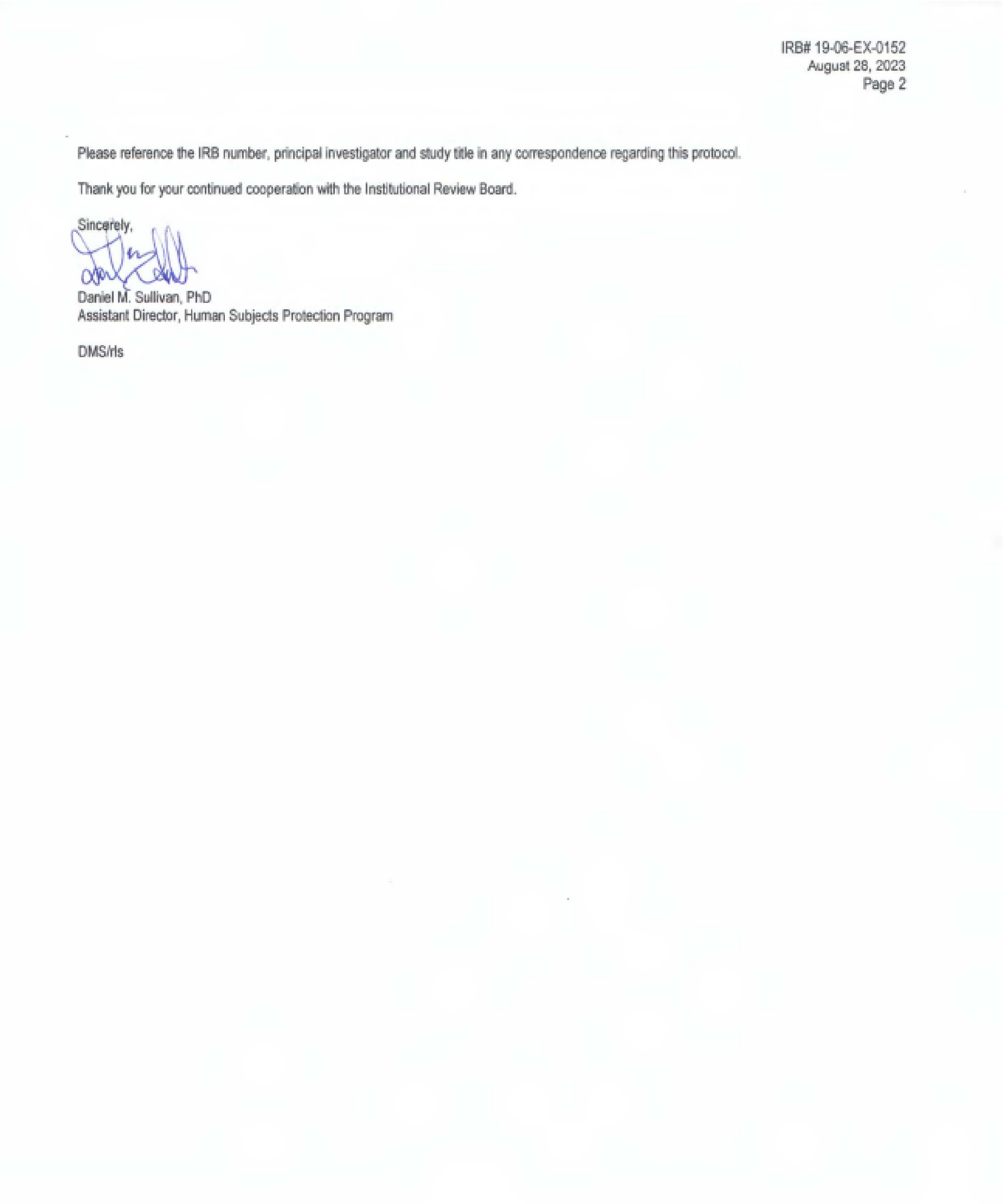
Remember that a copy of all correspondence relating to any site visit or regulatory visit must be submitted to the IRB office within five (5) days of receipt by the EVMS site. Refer to the *2022 EVMS IRB SOPs Section 22.0* for more information.

Eastern Virginia Medical School (EVMS) has a Federal wide Assurance (FWA 00003956) from OHRP. The Institutional Review Boards (IRB 00000460 and IRB 00001345) are registered with OHRP and are in compliance with 45 CFR 46, 21 CFR 50, and 21 CFR 56.

HUMAN SUBJECTS' PROTECTIONS PROGRAM

P.O. BOX 1980
NORFOLK, VA 23501-1980
TEL 757.446.8423
FAX 757.624.2275
www.evms.edu

Community focus. World impact.



IRB# 19-06-EX-0152
August 28, 2023
Page 2

Please reference the IRB number, principal investigator and study title in any correspondence regarding this protocol.

Thank you for your continued cooperation with the Institutional Review Board.

Sincerely,

Daniel M. Sullivan, PhD
Assistant Director, Human Subjects Protection Program

DMS/rls

VITA

Megan Anita Witherow
Department of Electrical and Computer Engineering
Old Dominion University
Norfolk, VA 23529

Megan A. Witherow received her B.S. degree in computer engineering from Old Dominion University (ODU), Norfolk, VA, USA in Spring 2018. She joined the Vision Lab, Dept. of Electrical and Computer Engineering, ODU in 2015 as an undergraduate student and began her PhD program in the Vision Lab in Fall 2018. She is a 2020 National Science Foundation Graduate Research Fellow. Her research interests include computer vision, machine learning, deep learning, human-computer interaction, and affective computing.

SELECTED PUBLICATIONS

JOURNAL

M. A. Witherow, N. Diawara, J. Keener, J. W. Harrington, and K. M. Iftekharuddin, "Pilot Study to Discover Candidate Biomarkers for Autism based on Perception and Production of Facial Expressions," *ArXiv preprint*, 2024.

M. A. Witherow, C. Butler, F. Ilgin, N. Diawara, J. Keener, J. W. Harrington, and K. M. Iftekharuddin, "Customizable Avatars with Dynamic Facial Action Coded Expressions (CADyFACE) for Improved User Engagement," *ArXiv preprint*, 2024.

M. A. Witherow, M. D. Samad, N. Diawara, H. Y. Bar, and K. M. Iftekharuddin, "Deep Adaptation of Adult-Child Facial Expressions by Fusing Landmark Features," *IEEE Transactions on Affective Computing*, pp. 1-12, 2023.

CONFERENCE

M. A. Witherow, M. D. Samad, N. Diawara, and K. M. Iftekharuddin, "Facial landmark feature fusion in transfer learning of child facial expressions," in Proc.SPIE, 2022.

M. Witherow, W. Shields, M. Samad, and K. Iftekharuddin, "Learning latent expression labels of child facial expression images through data-limited domain adaptation and transfer learning," in SPIE Optical Engineering + Applications, 2020.

M. Witherow, M. Samad, and K. Iftekharuddin, "Transfer learning approach to multiclass classification of child facial expressions," in SPIE Optical Engineering + Applications, 2019.