Discriminative Few Shot Learning of Facial Dynamics in Interview Videos for Autism Trait Classification

Na Zhang, Mindi Ruan, Shuo Wang, Lynn Paul, and Xin Li, *Fellow, IEEE*

***Abstract*—Autism is a prevalent neurodevelopmental disorder characterized by impairments in social and communicative behaviors. Possible connections between autism and facial expression recognition have recently been studied in the literature. However, most works are based on facial images or short videos. Few works aim at Autism Diagnostic Observation Schedule (ADOS) videos due to their complexity (e.g., interaction between interviewer and interviewee) and length (e.g., usually last for hours). In this paper, we attempt to fill this gap by developing a novel discriminative few shot learning method to analyze hour-long video data and exploring the fusion of facial dynamics for the trait classification of ASD. Leveraging well-established computer vision tools from spatio-temporal feature extraction and marginal fisher analysis to few-shot learning and scene-level fusion, we have constructed a three-category system to classify an individual into Autism, Autism Spectrum, and Non-Spectrum. For the first time, we have shown that certain interview scenes carry more discriminative information for ASD trait classification than others. Experimental results are reported to demonstrate the potential of the proposed automatic ASD trait classification system (achieving 91.72% accuracy on the Caltech ADOS video dataset) and the benefits of few-shot learning and scene-level fusion strategy by extensive ablation studies.**

***Index Terms*—Autism Spectrum Disorder (ASD), Autism trait classification, facial dynamic features, marginal fisher analysis (MFA), few-shot learning (FSL), scene-level fusion**

I. Introduction

Autism Spectrum Disorder (ASD) is a developmental disorder that affects communication and behavior [1]. Individuals with ASD often have difficulty interpreting and regulating their own emotions, as well as understanding the emotions expressed by others [2]. Studies on facial expression/emotion recognition and ASD have mainly used static images with posed expressions in the literature (e.g., [3] and [4]). Despite the extension to dynamic video with posed facial expression [5], there is still no automated and comprehensive analysis of facial expression in autism, especially in *natural settings* [6]. To our knowledge, existing research on computer vision for the diagnosis of ASD has been limited to eye tracking data [7] and perspective photographs of the self [8]. This is mainly due to the lack of video data collected from realistic interviews that capture the facial expression of patients with ASD.

The motivation behind this work is two-fold. On the one hand, faces provide non-verbal information that is important for social communication among typically developing people. Studies have shown that more than half of non-verbal visual-based behaviors of people are around the facial region - e.g., facial expression changes, head movements, eye glances, eyebrow raising, etc., in human communication activities [9], [10]. Some behaviors related to facial dynamics, such as gaze patterns, have been explored in autism analysis and have been shown to be useful for autism detection. However, there are no video data that capture the facial expression of patients with ASD in a naturalistic setting. To fill in this gap, it is desirable to construct a video data set collected from realistic interviews, such as the autism diagnosis observation schedule (ADOS) [11]. Our collaboration with Caltech researchers has greatly facilitated the construction of this database based on ADOS interviews of nearly 50 patients from 2015 to 2017.

On the other hand, the rise of artificial intelligence (AI), including machine learning (ML) and computer vision (CV), has made impressive progress from face recognition and emotion analysis to action detection and speech recognition. Rapid advances in AI have also been exploited in the field of behavior imaging to understand human behaviors [12], [13] and the early diagnosis of autism [14]. ML technology has improved diagnosis and intervention research in behavioral sciences, such as depression diagnosis [15], [16] and stroke rehabilitation [17]. In recent years, CV-based approaches have presented a class of quantitative and objective diagnostic tools for ASD by focusing on gaze patterns (e.g., eye movement [18], [19], visual attention [7], [20] or eye tracking [21], [22]) or body movements (e.g., gesture analysis [23], motor skills [24], and repetitive behaviors [25]). Currently, there is still no CV-based study on the feasibility of ASD classification based on facial expression.

Patterns of non-verbal behavior of a person, such as mobility, complexity, and dynamic activation, can be quantified to provide clues for behavior analysis. Motivated by [15], we present a novel extension of previous studies to the classification of autism traits by facial expressions. More specifically, we will extract the nonverbal features of facial dynamics to automatically classify people with ASD of different severity

Na Zhang, Mindi Ruan, and Xin Li are with the Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown WV 26506-6109.

Shuo Wang is with Department of Radiology, Washington University in St. Louis, 4525 Scott Ave, St. Louis, MO 63110.

Lynn Paul is with Division of the Humanities and Social Sciences, Caltech, 1200 E California Blvd, Pasadena, CA 91125.

This research was supported by an NSF CAREER Award (BCS-1945230), Air Force Young Investigator Program Award (FA9550-21-l-0088), Dana Foundation Clinical Neuroscience Award, ORAU Ralph E. Powe Junior Faculty Enhancement Award (to SW), and an NSF grant (IIS-1908215 and IIS-2114644) and the WV Higher Education Policy Commission grant (HEPC.dsr.18.5; to XL).

from the raw video data of the interaction, called ADOS. As the gold standard for the research diagnosis of autism, each ADOS video contains 15 observation activities, such as telling stories and brushing teeth. During the interview, the examiner presents the participant, who is evaluated, with numerous opportunities to exhibit behaviors of interest in the diagnosis of autism through standard procedures for communication and social interaction. These videos are designed to capture abnormal behaviors in people with ASD and are rich in terms of behavior to analyze. All videos have been scored by ADOS-reliable clinical psychologists with consensus, and overall ratings are made at the end of the schedule. These ratings (i.e. autism, autism spectrum, and non-spectrum) can be used to formulate a diagnosis result through the use of a diagnostic algorithm. We aim to discover whether an AI-enabled method can be developed to automatically evaluate ASD traits from interview videos. To our knowledge, this work is the first to examine a computational approach in ADOS interview videos for autism analysis and to measure the severity of autism computationally.

The main contributions of this research are summarized in the following four aspects.

- Construct the first long interview video dataset for the diagnosis of ASD with manually labeled scores and data sheet. Unlike popular short videos (e.g., TikTok), efficient and reliable analysis of hour-long video has remained an underresearched field in computer vision.
- Develop an ADOS video classification system capable of ASD trait classification by integrating spatio-temporal feature extraction and K-SVD sparse coding with marginal Fisher analysis (MFA).
- Propose a few-shot learning (FSL) extension of the developed system for ASD classification based on distribution calibration and adaptive posterior learning. When combined with the fusion of the feature levels of each scene, our FSL system has reached an accuracy of 91.72% on the Caltech ADOS video data.
- Demonstrate the benefit of fusion at the scene level, as well as unequal distribution of diagnostic information on ASD between different scenes. Our scene-level analysis results support the hypothesis that ASD is a complex condition beyond the classification of three categories, implying the need for further study on ASD phenotyping.

The rest of this paper is organized as follows. Section II reviews related work on autism analysis and detection. Section III introduces the newly constructed ADOS interview video database. Section IV describes our computational classification of ASD traits based on facial dynamics extraction, few-shot learning, and scene-level fusion. Extensive experiments are carried out, including ablation studies, and their results are reported in Section V. Finally, Section VI includes the discussion and conclusions.

II. Related Works

Recently, machine learning methods have been applied to autism analysis and diagnosis with various modalities, such as atypical visual scanning patterns during face and emotion perception, abnormal hand gestures and body behaviors, strange speech traits (e.g., loud volume, limited vocal variation, abnormal speech speed), etc. In this section, major related works are discussed briefly.

A. ASD Assessment Through Gaze Pattern Analysis

People with ASD have atypical attention to visual stimuli, such as human faces [57]. Various studies on the gaze pattern have been conducted in autism and developmental disorders [58], [59], [60], [61], [62], [63], [64], [65], [66], [27], [30], [26]. ML-based methods have been widely used and achieved good performance in various applications of gaze pattern analysis. In [7], the authors concentrated on the analysis of differences in eye movement patterns between people with typical development (TD) and those with ASD using a Deep Neural Network (DNN), and a Fisher score-based image selection method was adopted to learn more discriminative features for an efficient and accurate diagnosis. In [21], [30] and [20] proposed saliency prediction models based on deep neural networks, e.g., the Generative Adversarial Network (GAN), the Convolutional Neural Network (CNN), to predict atypical visual attention of children with ASD. Most recently, SP-ASDNet [22] is a framework that uses both CNN and long-short-term memory (LSTM) networks to classify whether an observer is TD or has ASD based on the gaze's scan path.

Traditional ML methods, such as the support vector machine (SVM), have also been used for gaze pattern analysis [27], [19], [28], [29]. In [27], five gaze characteristics (standard deviation of gaze points, standard deviation of difference in gaze points, standard deviation between gaze and annotated object of interest, RMSE between gaze and annotated object of interest, and delay in looking at the object of interest) were calculated for binary classification using SVM. In [28], a method was proposed to automatically recognize children with ASD for raw video data by analyzing the trajectory of eye movement and using the SVM for classification. In [29], the fixation times of children with ASD for classification were investigated and demonstrated that a short video clip may provide enough information to distinguish ASD from children with TD. In [18], both electroencephalography (EEG) and eye movements were considered for the diagnosis of ASD. They have used several methods such as SVM, logistic regression, deep neural network, and Gaussian naive Bayes for classification. Recent work [26] used scan path trend analysis (STA) to identify the trending path of users on a web page based on their eye movements. [33] applied eye-tracking in children and adults to assess visual attention allocation in a dynamic social orientation paradigm, and found qualitative differences between ages in ASD.

B. Interpretation of Facial Expressions of ASD

The human faces are among the most important visual stimuli for social interactions. The failure to accurately interpret facial expressions (that is, happiness, surprise, fear, anger, disgust, sadness) [67], [68], [69], [70] and facial processing [71], [72], [73] is one of the key impairments in ASD. Recent work also indicates that observers with ASD have difficulty using facial motion patterns to judge identity or

TABLE I
THE ASD DIAGNOSIS METHODS USING ML TECHNIQUES. "-" MEANS N/A.

Methods	year	Target Group	Characteristic	Data	Participants	Algorithm/Model
Jiang&Zhao [7]	2017	adult	gaze pattern	image	20ASD+19TD	DNN
Eraslan et al. [26]	2020	adult	gaze pattern	image	15ASD+15TD	Scanpath Trend Analysis
Ahuja et al. [27]	2020	adult	gaze pattern	video	35ASD+25TD	Gaze features
Tao&Shyu [22]	2019	child	gaze pattern	image	14ASD+14TD	CNN+LSTM
Duan et al. [21]	2018	child	gaze pattern	image	13ASD	GAN
Dris et al. [19]	2019	child	gaze pattern	image	-	Region of Interests
Wei et al. [20]	2019	child	gaze pattern	image	-	CNN
Li et al. [28]	2018	child	gaze pattern	video	53ASD+136TD	Displacement Feature
Wan et al. [29]	2019	child	gaze pattern	image	37ASD+37TD	Areas of Interest
Fernández et al. [30]	2020	child	gaze pattern	image	8ASD+23TD	CNN
Liu et al. [31]	2016	child	gaze pattern on face	image	29ASD+2groups TD	K-means
Jiang et al. [32]	2019	-	gaze pattern on face	image	23ASD+35TD	DNN
Kaliukhovich et al. [33]	2021	child+adult	gaze pattern	images	94ASD+38TD	-
Leo et al. [34]	2019	child	facial expression	image	17ASD+10TD	CNN
Beary et al. [35]	2020	child	facial expression	image	1,507ASD+1,507TD	MobileNet
Akter et al. [36]	2021	child	facial expression	image	1,468ASD+1,468TD	MobileNet-V1
Lu& Perkowski [37]	2021	child	facial features	image	561ASD+561TD	VGG16
Kowallik et al. [38]	2021	adult	facial expression	image	55ASD	logistic regression
Lecciso et al. [39]	2021	child	facial expression	image	12ASD	-
Guo et al. [40]	2021	child	facial expression	image	30ASD+30TD	-
Elshoky et al. [41]	2021	child	facial features	image	2,936	A set of ML methods
Bangerter et al. [42]	2020	child+adult	facial expression	video	124ASD+41NT	Gaussian Mixture Model
Banire et al. [43]	2021	child	facial expression	video	20ASD+26TD	CNN
Zlibut et al. [44]	2021	adult	facial expression	video	27ASD+57NT	K-means
Alvari et al. [45]	2021	child	facial expression	video	18ASD+15TD	Openface
Zunino et al. [23]	2018	child	grasping a bottle	video	20ASD+20TD	LSTM
Crippa et al. [24]	2015	child	reach-to-drop task	video	15ASD+15TD	Kinematic Measures
Tian et al. [25]	2019	child	repetitive behavior	video	-	CNN
Oller et al. [46]	2010	child	speech	acoustic data	232	LDA
Ecker et al. [47]	2010	adult	brain anatomy	-	-	Volumetric&Geometric feature of cortical surface
Sherkatghanad et al. [48]	2020	adult	brain imaging	image	539ASD+573TD	CNN
Thabtah&Peebles [49]	2020	adult	questionnaires	text	189ASD+515TD	Rule-based architecture
Devika&Chinnaiyan [50]	2020	adult+toddler	questionnaires	text	-	A set of ML methods
Nasser et al. [51]	2019	all	questionnaires	text	1100 instances	ANN
Raj&Masood [52]	2019	all	questionnaires	text	1100 instances	A set of ML methods
Peral et al. [53]	2020	all	questionnaires	text	1100 instances	A set of ML methods
Hossain et al. [54]	2020	all	questionnaires	text	1100 instances	A set of ML methods
Küpper et al. [55]	2020	adult	behavior	ADOS codes	385ASD+288TD	SVM
Ruan et al. [8]	2021	adult	attention behavior	photos	16ASD+21TD	DNN
Subah et al. [56]	2021	child+adult	fMRI	image	402ASD+464TD	DNN
Ours	2022	adult	facial dynamics	ADOS video	33ASD (3 levels)	LPQ-TOP+K-SVD+MFA +few-shot learning

gender and may be less able to derive a global perception of motion [57], [74], [75], [76]. Machine learning methods have been used in this kind of analysis. [31] proposed a machine learning method to classify children with ASD and two groups of matched controls by analyzing gaze patterns in a face recognition task. Despite their prominent accuracy, face stimuli and the structured recognition task are highly dependent on existing knowledge about ASD, limiting their generalizability to other clinical populations or young children who may not understand or comply with task instruction. In [32], a dynamic affect recognition assessment (DARE) task was adopted to distinguish between ASD and TD. Participants were asked to recognize one of six emotions while watching a slow transitioning face video, and their response time and eye movements were recorded.

Furthermore, analyzing the facial expressions of the participants to distinguish the differences between ASD and TD is also a good point. In recent years, many works have been proposed. [36] proposed a transfer learning-based autism face recognition framework to identify children with ASD in the early stages of face images collected from the Kaggle data repository. In [35], a MobileNet-based deep learning model was introduced to classify children with ASD. [40] proposed a facial expression analysis system to evaluate differences in empathy ability between children with ASD and TD by analyzing real-time facial expressions of children. [37] designed a VGG16 transfer learning-based ASD screening solution to detect ASD using facial images on a unique ASD dataset of clinically diagnosed children. [41] used various machine learning methods (SVM, Random Forest, deep learning, etc.) to predict ASD in children using facial images. [39] designed several computer-based interventions to help children with autism spectrum disorders improve their emotional skills.

Furthermore, the video data can provide more informa-

tion for analysis. In [42], it used automated facial analysis software to investigate the differences between the ASD and TD groups of children and adults with short video clips. [43] designed two face-based attention recognition models to detect and classify children with ASD. One is based on geometric feature transformation and the other is based on the transformation of time-domain spatial features into 2D spatial images. [45] investigated the facial expressions of babies to analyze facial micro-motions in videos to extract the subtle dynamics of Social Smiles. [34] analyzed how children with ASD and TD produce facial expressions by monitoring facial muscle movements, and then the output is fused to model the individual ability to produce facial expressions. [44] applied facial expression coding and clustering approach to find differences between autistic and neurotypical adults. [38] applied a baseline intervention-retest design to investigate the impact of imitation on facial emotion recognition with six basic emotional expressions.

C. ASD Assessment Using Body Movements Analysis

The underlying rationale for using body movements for the detection of ASD comes from psychological and neuroscience studies, claiming that the execution of simple motor acts is different between pathological and TD subjects, and this can be used to discriminate between them. In this category, the behaviors of subjects are mainly recorded by video cameras [23], [25]. In [23], a simple gesture of grasping a bottle by patients and children with TD was recorded and processed by a Recurrent Neural Network (RNN) for classification. [25] introduced an end-to-end deep architecture, the one-look-ahead early ASD detection (O-GAD) network, for video-based early ASD detection by taking arbitrary length videos as input. The network can detect typical ASD actions and determine if repetitive behaviors appeared at one glance only. [24] developed a supervised ML method to determine whether a simple movement of the upper extremities (reach-to-drop task) could be useful for accurately classifying low-functioning children with ASD aged 2 to 4. This work provided insight into a possible motor signature of ASD that may be potentially useful in identifying a well-defined subset of patients, reducing clinical heterogeneity within the broad behavioral phenotype.

D. ASD Assessment by Speech Traits Analysis

For generations, the vocal study and its role in language have been carried out laboriously, with human transcribers and analysts coding and taking measurements from small recorded samples. Large-scale statistical analysis of strategically selected acoustic parameters on the development of infant control over infrastructural characteristics of speech is able not only to track children's development of acoustic parameters known to play key roles in speech, but also to differentiate vocalizations from typically developing children and children with autism or language delay. [46] adopted this analysis method to demonstrate the potential to fundamentally enhance research on vocal development and add a fully objective measure to detect speech-related disorders, such as autism, in early childhood.

E. Others

Other methods such as brain imaging are also used for the detection of ASD; for example, a multiparameter classification approach was developed in [47] to characterize the complex and subtle structural pattern of gray matter anatomy involved in adults with ASD and discriminate between ASD and TD by SVM. Resting-state functional magnetic resonance imaging (fMRI) data from a multisite data set named Autism Brain Imaging Exchange (ABIDE) were used in [48] for the detection of ASD. In [56] proposed, a deep neural network (DNN) classifier was proposed to detect ASD using functional connectivity features of resting state fMRI data, such as the ABIDE dataset. Most recently, photos taken by ASD have been shown to have different characteristics than controls in [8], [60]. A summary of existing autism diagnosis methods using ML techniques is shown in Table I. More survey work can be found in [77], [78], [79].

III. CalTech ADOS Video Dataset Construction

For ASD diagnosis, we utilize a video dataset of ASD evaluations performed using the Autism Diagnostic Observation Schedule (ADOS) [11], which is a structured but natural discussion between the interviewer and the participant. ADOS interviews capture the complicated and rich behaviors of ASD in adults. All participants provided their informed consent in writing using procedures approved by the Institutional Review Board (IRB) of West Virginia University (WVU) and the California Institute of Technology (CALTECH).

A. Autism Diagnostic Observation Schedule (ADOS)

The Autism Diagnostic Observation Schedule (ADOS) is considered the gold standard clinically for the diagnosis of ASD. It consists of standardized activities that allow the examiner to observe the occurrence or non-occurrence of behaviors that have been identified as important for the diagnosis of autism. Structured activities and materials, as well as less structured interactions, provide standardized contexts in which social and communicative behaviors are observed. The responses of the participants to each activity are recorded by highly trained interviewers, and the interactions between the interviewer and the participant are recorded. The interviewer provides a detailed score of multiple aspects of ASD after completing ADOS. These scores are used to formulate a diagnosis through the use of a diagnostic algorithm. In effect, ADOS interviews provide a one-hour observation period, during which an examiner presents the individual being evaluated with numerous opportunities to exhibit behaviors of interest in the diagnosis of autism by standard "presses" for communication and social interaction [80]. "Presses" consist of planned social interactions in which ADOS evaluators are likely to elicit specific behavioral responses to differentiate those with ASD.

B. ADOS Interview Participants and Video Acquisition

For the recruited individuals, 33 participants completed ADOS interviews. Participants (age range = $16 \sim 37$ years; 26 men, 25.00 years; 7 women, 22.86 years), were primarily

TABLE II
THE SCORING CATEGORIES FOR THE ADOS-2 MODULE 4 OBSERVATIONS.

	Category	Items	Description
A	Language and communication	$A1 \sim A10$	speech, gesture, etc.
B	Reciprocal social interaction	$B1 \sim B13$	eye contact, facial expression, speech, gesture, gaze, etc.
C	Imagination	C1	expressive language skills
D	Stereotyped behaviors and restricted interests	$D1 \sim D5$	hand or figure behaviors, etc.
E	Other abnormal behavior	$E1 \sim E3$	over activity, anxiety, tantrums, etc.

TABLE III
ADOS-2 MODULE 4 SCORE CALCULATION AND CLASSIFICATION.

Label	Communication (Comm) A4, A8, A9, A10	Social Interaction (RSI) B1, B2, B6, B8, B9, B11, B12	Comm + RSI	Operation
Autism	≥ 3	≥ 6	≥ 10	AND
Autism Spectrum	< 3	< 6	< 10	OR
	≥ 2	≥ 4	≥ 7	AND
Non-Spectrum	< 2	< 4	< 7	OR

dominant right-handed (n = 31). Nine participants were interviewed twice (first visit and return visit). The time interval between the two interviews is about a half-year. So, we obtained 42 videos in total. All participants had a diagnosis of ASD, informed by the Autism Diagnostic Observation Schedule, Second Edition (ADOS-2) [81], and confirmed by expert clinical judgment. They met the cut-off scores for ASD on the ADOS-2 revised scoring system for Module 4 [82]. ADOS-2 was scored according to the latest method, and we also derived calibrated severity scores (CSS) for exploratory correlation analysis. (1) social affect (SA): 8.29 $\pm$ 4.55 (mean $\pm$ SD); (2) restricted and repetitive behavior (RRB): 2.43 $\pm$ 1.50; (3) severity score for social affect (CSS-SA): 6.0 $\pm$ 2.52; (4) severity score for restricted and repetitive behavior (CSS-RRB): 5.95 $\pm$ 2.40; (5) severity score for social affect plus restricted and repetitive behavior (CSS-All): 5.64 $\pm$ 2.79. The ASD group had a full-scale IQ (FSIQ) of 96.83 $\pm$ 13.48 (from the Wechsler Abbreviated Scale of Intelligence-2), a mean age of 24.05 $\pm$ 5.14 years. The videos were acquired from the California Institute of Technology. The research diagnosis of ADOS is a laboratory routine for all participants with ASD. These videos were recorded separately for people with autism in a quiet room in the hospital using a video camera with a data rate of 9.1 Mbps, a frame rate of 30 frames per second, and an image resolution of 720×480. Both the examiner and the participant sit at a table. The camera is set nearby, where the behavior of the participant can be clearly captured. The captured information contains the participant's body behaviors, face emotions, hand gestures, eye contact, speech traits, and reciprocal social exchanges with the examiner.

ADOS videos include 15 interview sections (scenes) between a clinician and a person suspected of having ASD. The tasks included in ADOS-2 Module 4 include: (1) Construction task: the participant uses puzzle pieces to complete a diagram and is instructed to request more pieces when needed; (2) Telling a story from a book: since the book has few words, the participant interprets the story from visual cues including reading emotions on the faces of the people in the story; (3) Description of a Picture; the picture provides opportunities for interaction with the interviewer and to gauge spontaneous language; (4) Conversation and Reporting: based largely on the picture the participant saw in (3); (5) Current Work or School: a series of questions about these aspects of their life; (6) Social Difficulties and Annoyance: discussion about social interactions and how they perceive them; (7) Emotions: talking about the events/objects that elicit different emotions in the participant and ask them to describe their feelings; (8) Demonstration Task: the participant is asked to show and tell the interviewer how to do a typical procedure such as brushing their teeth; (9) Cartoons: a series of cards depicting cartoon characters that tell a story then the participant stands to retell the story and their use of the gestures, emotions, and reference to relationships is evaluated; (10) Break: the participant is given a few items (magazines, toys, color pens, papers) and the interviewer observes their behavior during this free time; (11) Daily Living: questions about their current living situation gauge their level of independence; (12) Friends, Relationships, and Marriage: gauge the participant's understanding of the nature of these types of relationships; (13) Loneliness: the participant's understanding of loneliness is evaluated; (14) Plans and Hopes: what does the participant anticipate in the future for them self; (15) Creating a Story: the participant uses their imagination to create a novel story using some objects.

The 15 ADOS sections were proceeded one by one in order. In each section, there are multiple standard questions/instructions. Usually, the interviewer poses questions and each participant gives his or her corresponding responses, such as answering questions verbally, performing some actions such as gesturing with his hands. At the same time, the interviewer takes notes about the participant's responses in real time from the ADOS evaluation booklet. The duration of each section depends on the response of the participant. Different scenes are designed for the analysis of different aspects, e.g., facial expressions, body action, and hand movements. For example, Sections 8 and 9 show more body actions, while Sections 11 and 12 show more verbal communications.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2022.3178946, IEEE Transactions on Affective Computing

TABLE IV
THE INFORMATION OF SELECTED ADOS SCENES.

Scene	Average Length (Seconds)	# Frame (In total)	# Frames (Average)
5	398	493,052	11,738
6	381	471,860	11,234
7	473	579,509	13,797
11	409	502,536	11,965
12	515	630,211	15,005
13	79	94,659	2,254
14	67	80,917	1,927

C. ADOS Interview Videos

ADOS is a semi-structured assessment of communication, social interaction, and play (or imaginative use of materials) for people suspected of having autism. In this study, all individuals with ASD were videotaped during their ADOS interviews and all videos were scored with consensus by ADOS-reliable clinical psychologists. Scores will serve as a basis for training the machine learning algorithm. In summary, forty-two videos totaling 3,165 minutes have been captured. There are two people (participant and interviewer) in each video. Most of the time, each participant was asked to sit in the chair facing the camera when talking to the interviewer, except the cartoon section (#9) requires the participant to stand up to perform body movements to tell the story in the cartoon. Each video consists of $50 \sim 170$ minutes. The average duration of the videos is about 75.36 minutes.

For further analysis, the raw video data is first preprocessed. For each video, we mark the starting and end points of each task and split the video into 15 separate sub-videos based on the 15 ADOS tasks. To study the feasibility of automatically analyzing videos, we carefully chose interview sections for tasks 5, 6, 7, 11, 12, 13, and 14, focusing more on facial dynamics for feature extraction. Since the participants were sitting in the chair most of the time during the interviews, it is appropriate to capture the dynamic features around their face regions in a short and continuous period. 292 subvideos with 2,852,744 frames were finally picked up with an average length of 334 seconds, as shown in Table IV. Scenes 5, 6, 7, 11, 12 take longer than 13 and 14.

D. ADOS Scores and Labeling

The scoring of ADOS-2 Module 4 videos (based on the entire video of all interview questions, not a single observation) by ADOS experts includes the following 5 broad categories with 32 items: (A) Language and Communication, (B) Reciprocal Social Interaction, (C) Imagination/Creativity, (D) Stereotyped Behaviors and Restricted Interests, and (E) Other Abnormal Behaviors. In each scoring section (A-E), there are several detailed questions. Table II shows a detailed list of items for each scoring category. Each item contains a few score levels: 0, 1, 2, 3, 7, and 8. The score $0 \sim 3$ indicates the severity level of the ASD behavior targeted in that question. 0 means that the participant's response was at the level expected for a person without ASD, while a score of 3 would be highly indicative of ASD. Few questions include a possible score of 7 or 8 and indicate behaviors (e.g., limited by physical disability) that do not contribute meaningfully to the ASD scoring and thus would be scored 0 in the total score.

According to the revised ADOS-2 scoring system for module 4 algorithm [82], the ADOS scores of the 32 items must be converted to module 4 algorithm scores by (1) transferring the assigned ratings of 0,1 and 2 directly into the algorithm form (do not convert), (2) converting the assigned ratings of 3 to algorithm scores of 2, and (3) converting assigned ratings of 7 or 8 to algorithm scores of 0. The Module 4 algorithm adopts the transferred scores of categories A (communication), B (reciprocal social interaction), and A + B, for the diagnostic classification of the autism spectrum (as shown in Table III). There are three diagnostic categories in total. (1) Autism: all three totals are greater than or equal to the three separate corresponding autism cut-offs (A: 3, B: 6, A+B: 10); (2) Autism Spectrum: all three totals are greater than or equal to the three separate corresponding autism spectrum cut-offs (A: 2, B: 4, A+B: 7), but at least one of them is less than its corresponding autism cut-off (A: 3, B: 6, A+B: 10); (3) Non-Spectrum: any one of the three totals is less than the autism spectrum cut-offs (A: 2, B: 4, A+B: 7). According to the ADOS-2 Module 4 classification, we have 17 videos with participants who have the diagnosis of autism, 10 videos with participants who have the diagnosis of Autism Spectrum, and 15 videos of individuals whose ADOS score resulted in the diagnosis of NonSpectrum, indicating that they do not meet the criteria for a reliable diagnosis of ASD.

E. Characterization of ASD Traits by ADOS Videos

Most existing databases for autism analysis and detection focus on capturing the gaze pattern [7], [22], [29] of individuals when they visually scan images of natural scenes or perform facial processing or facial emotion recognition [31], [32]. Some databases used in body behavior analysis just videotaped certain specific body movements, such as grasping a bottle [23] or a simple upper limb movement [24]. These gaze patterns, simple hand movements, and body behavior are conventional characteristics present in ASD, which only present the one-fold symptom of autism. In addition, the number of images or video clips in these databases is not large either.

Our database of ADOS interview videos is rich in terms of the variety of behaviors exhibited, including facial dynamics, gaze patterns, eye contact, hand movements, body behavior, speech traits, etc. It is specially designed to capture the more complicated and diverse behaviors of adults with ASD, not just one aspect of the developmental disorder. Additionally, each video contains 15 ADOS interview tasks, which also provide abundant information for analysis. Furthermore, the database is very large and includes 42 videos from 33 recruited individuals, totaling 3,165 minutes of video. After splitting each video into 15 separate sub-videos by the time point of each scene, 292 sub-videos were obtained with 2,852,744 frames with an average length of 334 seconds. Lastly, all videos have already been scored by ADOS-reliable clinical

psychologists with consensus, giving us the ground truth for ASD diagnostic classification.

IV. Computational Method of ASD Trait Classification

Our method focuses on the analysis of the facial dynamics of people with ASD when participating in ADOS interviews. Figure 1 illustrates an overview of the framework. We first extract key frames containing the subject of interest from ADOS scenes and crop face regions as a preprocessing step. Then we perform spatio-temporal feature extraction from the cropped video and apply sparse coding to generate discriminative features. Next, feature distribution calibration and adaptive posterior learning are performed for few-shot classification.

A. 3D Spatio-Temporal Facial Feature Extraction

Unlike image-based facial expression analysis [3], [4], both spatial and temporal information in ADOS videos is important for the classification of autism traits. On the one hand, spatial information relevant to ASD is embedded in the form of facial appearance, static expression, and eye movements of participants. However, temporal information related to ASD is characterized by facial movement in frames, conveying the facial dynamics of subjects, such as expression and microexpression changes, gaze patterns, and variations in head pose. To fully exploit the discriminative information in space and time, it is plausible to consider a method of 3D spatio-temporal feature extraction by encoding both appearance and dynamics from the given input video.

Video pre-processing. Before extracting dynamic facial features, a face detection method is applied to determine the facial region of the participant in video frames. Here, a face detector with multitask cascaded convolutional networks (MTCNN) [83] is adopted, which is a multitask-based deep cascaded face detector. All detected faces are cropped by a square bounding box, and 60 continuous face frames (about two seconds long) are integrated as a 3D volume for feature extraction. In some subvideo clips, the frames in which the face detector fails are simply skipped.

Extension of spatial temporal features. Local Phase Quantization in Three Orthogonal Planes (LPQ-TOP) [84] is a descriptor extended from the LPQ purely spatial representation for spatio-temporal analysis. It is obtained from small space-time video volumes. Histograms from all space-time video volumes are concatenated as a feature vector to represent the corresponding face image sequence. First, the basic features of LPQ, denoted XY-LPQ, XT-LPQ, and YT-LPQ, are independently extracted from three orthogonal planes. XY, XT, and YT, respectively, while considering the co-occurrence statistics in these three directions. The XY plane provides the spatial domain, while the XT and YT planes have the temporal information. Thus, by using this dynamic texture descriptor, both appearance and motion in three directions are considered.

B. Discriminative Representation Learning

For the task of classification of ASD traits, it is important to work with discriminative characteristics instead of the original data representations, such as the LPQ-TOP characteristics. The problem of learning discriminative representation [85] has been extensively studied in the literature on ML and CV. Generally speaking, there are two classes of strategies: dictionary learning through sparse coding (e.g., K-SVD [86]) and dimensionality reduction through factor analysis (e.g., Marginal Fisher Analysis [87]). In this work, we have considered a combination of both methods to generate composite discriminative characteristics for our video analysis of autism.

Sparse Coding through K-SVD. After Spatio-Temporal feature extraction, hundreds of LPQ-TOP features are generated from a single subvideo. Sparse coding is used to organize these feature descriptors together, which aims at obtaining a sparse representation. Sparsity means that only a small fraction of the entries are non-zero among all coefficients of base vectors. This kind of representation can be discriminative and concise, as it could select a subset of base vectors that express the concentrated input signal. A popular approach to signal modeling is the synthesis-based sparse representation model, where a signal $\mathbf{x} \in \mathcal{R}^d$ is assumed to be composed as a linear combination of a few atoms from a given dictionary $\mathbf{D} \in \mathcal{R}^{d \times n}$ [88], [89]. The main activity in studying this model was to estimate the representation of a corrupted signal and to learn the dictionary $\mathbf{D}$ from examples of signals. K-Singular Value Decomposition (K-SVD) [86] is one of the popular dictionary learning algorithms.

K-SVD is an unsupervised dictionary learning method and focuses on representational power [90]. Given a set of input signals of n dimensions $\mathbf{Y} = [y_1, ..., y_N] \in \mathcal{R}^{n \times N}$, a dictionary $\mathbf{D} = [d_1, ..., d_K] \in \mathcal{R}^{n \times K}$ can be learned by lowering the reconstruction error by sparse coding as follows:

$$arg\ min_{D,X}\ \|Y - DX\|_2^2\ \ s.t.\ \forall i,\ \|x_i\|_0 \leq T, \qquad (1)$$

where $\mathbf{X} = [x_1, ..., x_n] \in \mathcal{R}^{K \times N}$ consists of the sparse codes of the input signals $\mathbf{Y}$, and $\mathbf{T}$ is a positive integer that specifies the sparsity level. An LPQ-TOP feature can be treated as a sparse linear combination of all dictionary words plus a residual or sparse error. The values of the coefficient of the linear combination are generated as a sparse code. Finally, all these sparse codes on the whole sub-video are averaged as a descriptor of the visual-based nonverbal behavior manner for the subject. Each code can be considered a typical behavior pattern. The average sparse codes provide a better characterization towards a clarified behavior manner.

Reduced dimensionality using marginal Fisher analysis (MFA). To further enhance discriminative capacity, a supervised dimensionality reduction algorithm called Marginal Fisher Analysis (MFA) [87] is used to map the sparse feature in a new space with better discrimination. Compared to Linear Discriminant Analysis (LDA), there is no assumption about the data distribution; thus, it is more general for discriminant analysis. This method utilizes the graph embedding framework as a tool, designs two graphs that characterize the compactness of the infraclass and the separability between classes, respectively, and optimizes their corresponding criteria based on the graph embedding framework by obtaining the optimal projection vector $\hat{v}$ to satisfy the equation. (2):

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2022.3178946, IEEE Transactions on Affective Computing

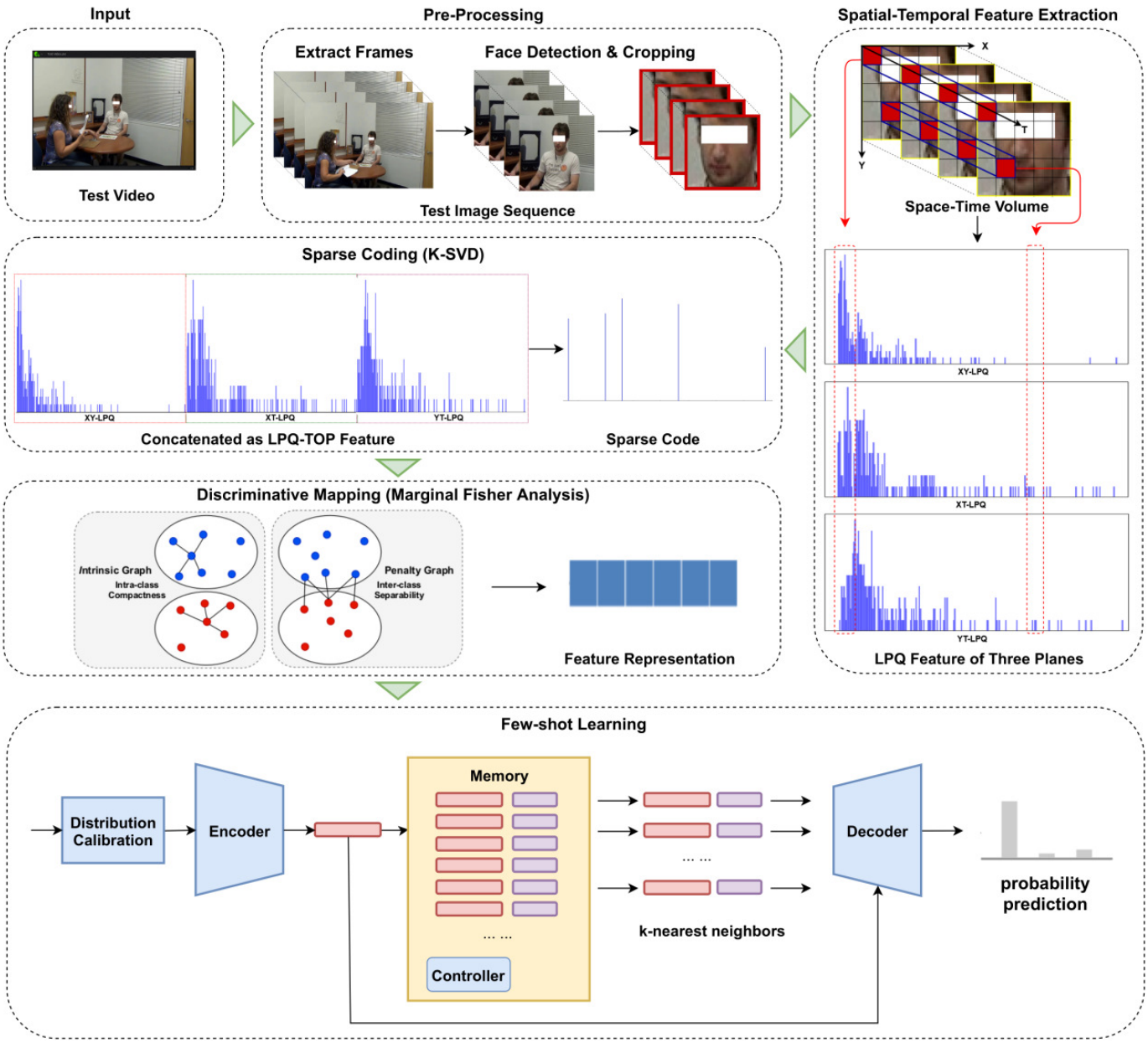


Fig. 1. The pipeline of video-based facial dynamics analysis for people with autism spectrum disorder (ASD) trait classification.

$$\hat{v} = arg\ min_{\mathbf{v}} \frac{v^T X L_{intra} X^T v}{v^T X L_{inter} X^T v} \tag{2}$$

where $\mathbf{X} = [x_1, ..., x_n]$ is the input data, $\mathbf{L}_{intra}$ is the Laplacian matrix within the class and $\mathbf{L}_{inter}$ is the Laplacian matrix between classes. $\mathbf{L}_{intra}$ is calculated by $\mathbf{D}_{intra} - \mathbf{S}_{intra}$, and $\mathbf{L}_{inter}$ is $\mathbf{D}_{inter} - \mathbf{S}_{inter}$. In them, $\mathbf{S}_{intra}$ is the affinity weight matrix where $s_{ij} = 1$ when x_i and x_j are $\mathbf{k}$ closest neighbors of each other in the same class, otherwise $s_{ij} = 0$. $\mathbf{S}_{inter}$ is the opposite. $\mathbf{D}$ is a diagonal matrix in which $\mathbf{D}_{i,i} = \sum_j s_{ji}$.

C. Few-shot Learning

Distribution calibration (DC). The model is easy to overfit if it is trained on data with a biased distribution containing only a limited number of samples, such as the ASD population. An effective strategy to combat these few-shot learning scenarios is to calibrate the distribution of these few-sample classes by transferring statistics from classes with sufficient examples [91]. Inspired by the success of distribution calibration [91], we assume that some examples can be sampled from the calibrated distribution to expand the input to the classifier. In the framework of autism trait classification, we further assume that every dimension of the extracted facial feature in the previous subsection follows a Gaussian distribution. Therefore, the mean (μ) and the variance (σ) of the distribution of each class in the target data can be borrowed from that of similar classes (base data) whose statistics are better estimated with a few samples.

Similarly to [91], the characteristic of the target data can be transformed using Tukey's ladder of power transformation, as described in the equation. (3) to reduce the skewness of the distribution:

TABLE V
ACCURACY (%) OF OUR METHOD (LPQ-TOP + K-SVD + MFA + APL + DC), LPQ-TOP, LPQ-TOP + K-SVD, LPQ-TOP + K-SVD + MFA AND LPQ-TOP + K-SVD + MFA + APL, IN INDIVIDUAL SCENES.

Scene No.	5	6	7	11	12	13	14
LPQ-TOP	61.40	61.35	52.29	55.11	51.29	47.98	44.72
LPQ-TOP+K-SVD	65.35	64.21	61.26	55.61	53.81	56.35	61.41
LPQ-TOP+K-SVD+MFA	81.45	81.51	73.49	69.58	72.09	76.50	72.00
LPQ-TOP+K-SVD+MFA+APL	88.67	87.35	79.17	79.17	83.33	87.25	83.33
LPQ-TOP+K-SVD+MFA+APL+DC (ours)	**89.64**	**89.55**	**86.83**	**87.12**	**88.33**	**88.67**	**88.49**

$$\tilde{x} = \begin{cases} x^{\lambda} & if\ \lambda \neq 0 \\ log(x) & if\ \lambda = 0 \end{cases} \quad (3)$$

where λ is a hyperparameter to adjust the distribution. For each class, Eq. (4) is used to calibrate the mean $\hat{\mu}$ and the covariance $\hat{\sigma}$ for each class using $\tilde{x}$, and then the generated characteristics are achieved from the calibrated distribution.

$$\hat{\mu} = \frac{\mu + \tilde{x}}{2}, \hat{\sigma} = \sigma + \alpha \quad (4)$$

where α is a hyperparameter that determines the degree of dispersion of spatiotemporal features after discriminative mapping.

Adaptive Posterior Learning (APL). Next, the calibrated features are fed into an adaptive posterior learning model (APL) [92] to perform a few shots of learning. The key idea behind APL is to approximate probability distributions by recalling the most surprising observations it has encountered. In the situation of ASD trait classification, past observations can be recalled from an external memory module and processed by a decoder network. The objective of FSL is achieved by combining information from different memory slots to generalize beyond direct recall. More specifically, our APL implementation consists of three parts: Encoder, Decoder, and Memory. The encoder encodes the input. It is implemented by a convolutional network which is made up of a single first convolution and 15 convolutional blocks. Each block has a Batch Normalization layer, a ReLU activation, and a convolutional layer with kernel size 3. For every three blocks, the convolution contains a stride of two to downsample the feature. Finally, the feature is flattened to a 1D vector and passed through a Layer Normalization function.

Memory stores the codes that the encoder has seen as a key value format. The key is the encoded embedding, and the value is the true label. A controller is designed to decide which embedding can be written, while at the same time trying to minimize the amount of written embedding. A quantity-metric surprise is defined to indicate that the probability model assigns the input to the true class correctly. The higher the probability, the less surprised it will be. If an embedding is surprising, it should be written in memory; otherwise, it should be discarded. When querying, the memory is scanned for the nearest k neighbors of the input. The concatenation of query embeddings, recalled neighbor embeddings with memory labels, and distances are fed to the decoder.

The decoder predicts the probability distribution on the targets. It is implemented by a relational feed-forward module with self-attention. It compares each neighbor individually with the query using a cross-element comparison with a self-attention module, and then reduces the activations with an attention vector calculated from neighbor distances. Self-attention blocks repeat five times in a residual manner. The resulting tensors are called activation tensors. Also, the distances between neighbors and query are passed through a softmax layer to generate an attention vector, which is summed with the activation tensor over the first axis to obtain the final logit result for classification.

TABLE VI
PERFORMANCE (%) OF OUR METHOD IN FEATURE-LEVEL FUSION (FEATURE CONCATENATION).

Features	Scenes	Accuracy	F1 Score
TOP 3	5,6,13	90.00	86.5
TOP 5	5,6,12,13,14	91.67	90.1
TOP 7	5,6,7,11,12,13,14	**91.72**	90.11

V. EXPERIMENTAL RESULTS

The experiment is carried out on the ADOS videos that we introduce in Section III. In this section, we first describe the implementation settings and the procedure in detail. Then, we show the performance of individual scenes and the fusion of selected scenes. We have also compared this work with several recently proposed image-based classification methods to demonstrate the superiority of the proposed approach. Significant improvements in the accuracy of ASD trait classification have been achieved with the proposed feature-level and scene-level fusion strategies. Finally, the results of our ablation study are reported; limitations of the proposed method and discussions about future research directions are presented.

A. Experimental Setup

Preprocessing. To extract a facial representation from the videos, the first step is to apply face detection and cropping to obtain the region of interest for each video frame. In our experiments, we chose MTCNN [83] as a face detector to crop the faces of the participants. All the cropped faces are resized to grayscale images with a size of 100 $\times$ 100. Finally, approximately 98.9% of frames containing faces are detected successfully.

Facial Feature Extraction. In our experiment, we adopt a three-dimensional face region subvolume as the basis for feature extraction. Two-second-long face frames are considered the basic unit. Frames in which faces are not detected are removed directly. After the extraction of 3D spatiotemporal features, a 768-dimensional feature vector is obtained for each 3D subvolume ($100 \times 100 \times 60$), in which the first 256 features are extracted from the XY plane, the second 256 from the XT plane, and the third 256 from the TY plane. To learn the sparse coding dictionary, only the LPQ-TOP features are used from the training data. We borrowed the settings from [15] in which the dictionary with a sparse level 3 and word size 250 for K-SVD, and for MFA, the number of nearest neighbors k is 4 for within-class definitions and 2000 for between-class definitions. The averaged sparse code after K-SVD is 250-dimensional for each video, and finally we got a feature vector with a fixed size of 40 for each video after MFA.

Few-shot Learning. For the calibration distribution, the extracted features of all selected scenes are treated as base data to estimate the mean and variance of each class for each scene. The mean of the feature vector is calculated as the mean of every single dimension of the vector. Covariance is used to better represent the variance between any pair of elements in the feature vector. Here, λ is 0.5 and α is 0.21. The calibrated features are then fed into the APL module to predict probabilities. The APL module applies a squared l_2 distance to compute the distance between queries and embeddings stored in memory, and the first three nearest neighbors are returned from memory. The threshold for the surprise measure metric is set to 0.75. We train 20,000 episodes using a cross-entropy loss and save the model per 100 episodes. The model with the highest accuracy is selected to evaluate the performance in the test set.

Performance Evaluation. In our experiment, we perform a three-class classification. Due to the small number of subjects in the database, we adopt a ten-fold cross-validation for each experiment. All 42 videos were partitioned into ten subsets. In them, two subsets contain five videos each and the other eight contain four videos each. Since some participants have two videos, these two videos are grouped into the same subset. During evaluation, one subset is retained as validation data to test the model, and the remaining nine subsets are used for training. The cross-validation process is then repeated ten times, and each of the ten subsets is used exactly once as validation data. The ten accuracy values are averaged to produce the final accuracy. We quantitatively measure the performance of the method in terms of the accuracy of each scene and the fusion of multiple scenes. For the fusion strategy, the concatenation of features is applied to selected scenes, and then the fused feature is fed into a few-shot learning module for classification.

B. Performance of Individual Scenes

Seven selected scenes (5, 6, 7, 11, 12, 13, 14) are experimented with ten-fold cross-validation, respectively. To facilitate visual inspection, we have shown the fusion results in both Table V and Fig. 3. One can clearly observe the improvement in accuracy as our computational model becomes more sophisticated (a significant gain is achieved by adding MFA and APL to the model).

From the result shown in Table V (last row), we can see that Scenes 5 and 6 gained pretty high accuracies (89.64% and 89.55%), the following Scenes 11, 12, 13 also obtained high accuracies (above 88%), and Scenes 7, 11 had the lowest accuracies (86.83% and 87.12%). It can be seen that the lengths of different scenes vary, as does their discriminative power for ASD trait classification. It seems that talking about topics that can cause mental or emotional stress in participants could reveal more explicit facial behaviors that show autism spectrum disorder. For example, as shown in Fig. 3, Scene 5 discusses work or school things to assess their realistic understanding of the possibilities for future employment, training, or experience necessary for future employment, and Scene 6 talks about social difficulties and annoyance, which contains problems or troubles getting along with other people, such as irritation, tease, or bullying.

C. Performance of Scene-level Feature Fusion

We have also compared the classification performance by combining several scenes at the feature level (that is, concatenating the features). Fusion experiments are conducted for the top three (5, 6, 13), the top five (5, 6, 12-14), and the top seven (5-7, 11-14), respectively. The average accuracy and average F1 scores of the 10-fold results are shown in Table VI. From the table, one can see that the top seven fusions achieve the best performance with an accuracy of **91.72%** by fusing the seven selected scenes that are comparable to the standardized diagnostic scales, with the advantages of efficiency and objectiveness. By contrast, scene-level fusion achieves 91.67% and 90.00% for the top 5 and top 3 settings. The comparison result shows the effectiveness of combining multiple scenes to improve performance.

To our knowledge, few video-based facial analysis methods for autism have been published in the open literature. For autism research, there are only a few video-based methods with special data constraints (e.g. [43]); others have only done some preliminary analysis work on extracted frames [45]. Taking [43] as an example. It uses face-based attention recognition models to detect and classify children with ASD in videos that record the facial behaviors of participants when they perform attention tasks. However, the class labels should be annotated as with and without attention according to the attention behavioral rules. For comparison, in our experiment, three image-based methods are chosen for facial analysis of autism. Since the codes are not released publicly, we try our best to reimplement the mentioned methods according to the description of the papers. Most methods are designed using transfer learning based on pre-trained deep models, such as VGG16, MobileNet, etc. We obtain the output of the last layer before the final classification layer of the model as a facial feature of the input frame. To obtain a proper feature that can efficiently represent the subvideo (scene), we first extract all frames of the subvideo, and then extract one deep facial feature from one frame for every 60 continuous face frames. Similarly

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2022.3178946, IEEE Transactions on Affective Computing

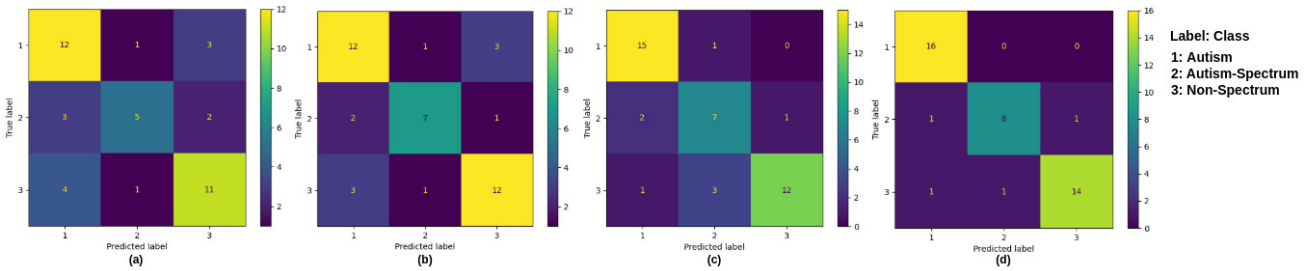


Fig. 2. Confusion matrix of (d) our method and other competing methods: (a) Beary et al. [35], (b) Akter et al. [36], and (c) Lu and Perkowski [37]. The classification results of all 42 videos from 10-fold experiments are shown in scene-level fusion with all 7 scenes. **Best view in color.**

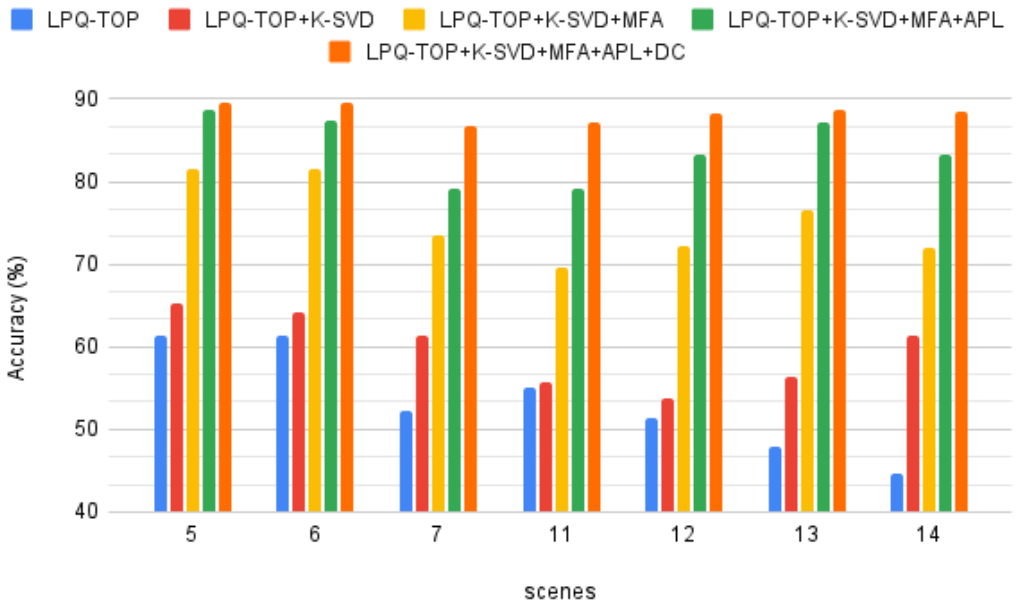


Fig. 3. Histogram of accuracy (%) of LPQ-TOP + K-SVD + MFA + APL + DC (our method), LPQ-TOP, LPQ-TOP + K-SVD, LPQ-TOP + K-SVD + MFA and LPQ-TOP + K-SVD + MFA + APL. **Best view in color.**

TABLE VII
PERFORMANCE (%) COMPARISON WITH OTHER COMPETING METHODS.

Methods	Beary et al.[35]		Akter et al.[36]		Lu& Perkowski [37]		Ours	
	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
Scene 5	66.07	63.41	73.57	74.92	75.43	74.15	89.64	86.66
Scene 6	64.29	62.26	61.07	59.92	70.71	74.16	89.55	87.5
Scene 7	55.36	59.77	64.64	63.41	57.86	61.73	86.83	86.79
Scene 11	62.5	60.85	64.64	66.24	56.07	53.41	87.12	84.36
Scene 12	62.5	63.41	57.5	61.73	60.86	57.44	88.33	85.17
Scene 13	67.86	71.26	68.21	66.22	66.79	60.77	88.33	85
Scene 14	66.07	68.24	70.57	68.73	54.29	55.06	88.49	86.83
Overall	68.46	68.53	75.9	74.92	78.96	76.26	**91.72**	90.11

to our experimental settings, all extracted facial features of the subvideo are used in K-SVD to generate sparse codes. All codes are averaged as the final descriptor of the subvideo. SVM is used for a three-class classification. As shown in Table VII and Fig. 2, our method performs better than others due to the fact that we consider temporal information of the entire 2-second clip, but comparing methods only use a single frame of each 2-second clip.Also, few-shot learning improves performance largely. The accuracy and F1 score are means of 10-fold results. For the confusion matrix, the classification results of all 42 videos from 10-fold experiments are shown.

D. Ablation Study

During feature extraction, we had several strategies to derive more discriminative characteristics, for example, K-SVD, MFA. For better classification, DC and APL are designed to further improve performance. Here, we validate the contribution of each component by performing an ablation

study. The results of our ablation study can be seen in Fig. 3. In the LPQ-TOP component experiment, the LPQ-TOP feature of each 3D space-time volume is extracted directly as a training/test sample for classification. In the remaining experiments, the feature of each subvideo is treated as a single data sample. From the result, one can see that most accuracies are lower than those of the complete method. Figure 3 gives a visual illustration of the contribution of different components. All components are useful for the classification algorithm to improve the recognition accuracy, especially the addition of MFA and APL. Since there is no assumption on the data distribution, MFA is more general for discriminant analysis, which explains its significant role in improving accuracy. The interclass margin can better characterize the separability of different classes. The deep learning-based APL module can better predict the probability distributions of input by recalling past observations and combining information from different memory slots to generalize beyond direct recall.

E. Limitations and Discussions

Although our model performs well on ADOS interview videos, it still has limitations in realistic operations. First, our experiment adopts a scene-level feature fusion strategy, which requires manually splitting entire hour-long videos into 15 separate scenes by time markers and extracting facial-dynamics features of each scene. Second, to reduce the bias of the data distribution, a distribution calibration strategy is adopted in the few-shot learning module. It needs base data with sufficient examples to estimate the mean and variance of each class, and it transfers the statistics to calibrate the distribution of our data that contain only a limited number of samples. In our experiment, the extracted features of all selected scenes are treated as base data. Although it works, the volume of base data is not enough. It is better to obtain more base data for calibration.

In future work, we will extend our experiment to all scenes and consider converting the classification problem to a regression task. In our experiment, we selected 7 scenes out of the 15 to analyze. Our original reason for not selecting the rest scenes was to try to analyze facial behaviors by extracting features of faces that are nearly frontal, with low pose angles, and low motion blurring. We believe that high-quality faces can provide more information. However, it is important to understand how the proposed approach performs in the other scenes to get a fair evaluation of the approach. For a fairly complete evaluation, we will consider adding these scenes in the following work. Also, it is also a good point to evaluate our approach on the entire hour-long videos. It is an open issue to determine the optimal dimension for extracting LPQ-TOP features. In our experiment, we chose 2-second-long clips, which can be extended to video clips of different sizes. The data also include the severity scores for each video. It would also be interesting to evaluate the approach to predicting severity scores.

VI. Conclusions

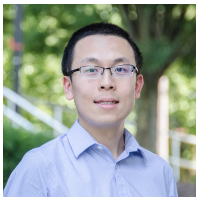
We have studied the feasibility of developing a method for autism analysis using ADOS interview video data, based on dynamic facial characteristics. The model first extracts the spatio-temporal features of the video and uses the combination of K-SVD with MFA to get more discriminative representations. A few-shot learning module is designed to further improve classification performance. The experimental results have shown that the proposed approach has a reasonably good result. The study is significant. First, an effective method is proposed to analyze human facial behaviors from ADOS interview videos. Second, objective measurement and analysis of facial dynamics in humans provide an objective characterization of atypical behaviors in autism. Although a large literature documents abnormal social communicative behavior in autism, essentially all have focused on an extremely narrow aspect, typically performing facial expression on images or videos shown on a screen, without real social interactions as ADOS data. Third, using an ML method to characterize behavior and/or score ADOS videos will make it much faster and more efficient. The algorithms can eventually serve as a screening tool to facilitate the analysis of hour-long ADOS videos. Our future work will focus on further improving classification accuracy by considering all scenes and extending to a video-based ASD severity score prediction task.

Acknowledgments

Thanks to Drs. Ralph Adolphs and Umit Keles for providing the Caltech dataset of ADOS interview videos.

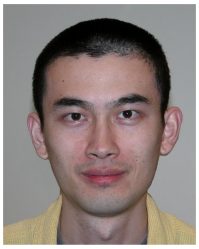
Na Zhang received her bachelor's degree in computer science from Beijing Information Science and Technology University, Beijing, China, in 2009 and her master's degree in computer science from Beihang University, Beijing, China, in 2012. She worked in State Grid Corporation of China (SGCC) from 2012 to 2015 as a software engineer. She is currently pursuing a Ph.D. degree in computer science at West Virginia University. Her main research interests include deep learning, computer vision, face analysis, with applications to face morphing and detection, and autism diagnosis.

Mindi Ruan Mindi Ruan received his B.S. degree in computer science from Heilongjiang University, Harbin, China, in 2017 and his M.S. degree in computer science from West Virginia University, Morgantown, WV, in 2020, where he is currently pursuing his Ph.D. degree with computer science. His research interests include autism diagnosis via deep learning, action recognition, behavior modeling, and video understanding.



Shuo Wang Dr. Shuo Wang received his Ph.D. in Computation and Neural Systems from the California Institute of Technology in 2014 and did his postdoctoral research at the California Institute of Technology and Princeton University before joining the faculty at West Virginia University in 2017. His laboratory moved to Washington University in St. Louis in 2021. His research interests revolve around visual saliency, attention, and facial processing, and he also has a strong interest in autism research. He uses an array of neuroscience methods, including psychophysics, eye tracking, intracranial electrophysiology, fMRI, EEG, and computational modeling.

Lynn Paul Dr. Lynn K. Paul is a Senior Research Scientist at the California Institute of Technology, where she studies how variations in brain structure impact neural organization and higher cognitive skills. She has ongoing neuroimaging and behavioral studies with several unique populations, notably adults who had hemispherectomy during childhood, people with congenital malformations of the corpus callosum, and adults with autism spectrum disorders. Dr. Paul is the founding president of the International Research Consortium for Corpus Callosum and Cerebral Connectivity (IRC5).



Xin Li received the B.S. degree with the highest honors in electronic engineering and information science from University of Science and Technology of China, Hefei, in 1996, and the Ph.D. degree in electrical engineering from Princeton University, Princeton, NJ, in 2000. He was a Member of Technical Staff with Sharp Laboratories of America from August 2000 to December 2002. since January 2003, he has been a faculty member in the Lane Department of Computer Science and Electrical Engineering. His research interests include image and video processing, computer vision, and data mining. Dr. Li was elected a Fellow of IEEE in 2017.

REFERENCES

[1] I. A. Rosenthal, C. A. Hutcherson, R. Adolphs, and D. A. Stanley, "Deconstructing theory-of-mind impairment in high-functioning adults with autism," *Current Biology*, vol. 29, no. 3, pp. 513–519, 2019.

[2] M. B. Harms, A. Martin, and G. L. Wallace, "Facial emotion recognition in autism spectrum disorders: a review of behavioral and neuroimaging studies," *Neuropsychology review*, vol. 20, no. 3, pp. 290–322, 2010.

[3] K. A. Pelphrey, J. P. Morris, G. McCarthy, and K. S. LaBar, "Perception of dynamic changes in facial affect and identity in autism," *Social cognitive and affective neuroscience*, vol. 2, no. 2, pp. 140–149, 2007.

[4] C. S. Monk, S.-J. Weng, J. L. Wiggins, N. Kurapati, H. M. Louro, M. Carrasco, J. Maslowsky, S. Risi, and C. Lord, "Neural circuitry of emotional face processing in autism spectrum disorders," *Journal of psychiatry & neuroscience: JPN*, vol. 35, no. 2, p. 105, 2010.

[5] O. Golan, Y. Sinai-Gavrilov, and S. Baron-Cohen, "The cambridge mindreading face-voice battery for children (cam-c): complex emotion recognition in children with and without autism spectrum conditions," *Molecular autism*, vol. 6, no. 1, pp. 1–9, 2015.

[6] P. J. Webster, S. Wang, and X. Li, "Review: Posed vs. genuine facial emotion recognition and expression in autism and implications for intervention," *Frontiers in Psychology*, vol. 12, 2021.

[7] M. Jiang and Q. Zhao, "Learning visual attention to identify people with autism spectrum disorder," in *Proceedings of the ieee international conference on computer vision*, 2017, pp. 3267–3276.

[8] M. Ruan, P. J. Webster, X. Li, and S. Wang, "Deep neural network reveals the world of autism from a first-person perspective," *Autism Research*, vol. 14, no. 2, pp. 333–342, 2021.

[9] R. L. Birdwhistell, "Toward analyzing american movement," *Nonverbal communication*, pp. 134–143, 1974.

[10] P. Ekman and W. V. Friesen, "Nonverbal behavior and psychopathology," *The psychology of depression: Contemporary theory and research*, pp. 3–31, 1974.

[11] C. Lord, M. Rutter, S. Goode, J. Heemsbergen, H. Jordan, L. Mawhood, and E. Schopler, "Austism diagnostic observation schedule: A standardized observation of communicative and social behavior," *Journal of autism and developmental disorders*, vol. 19, no. 2, pp. 185–212, 1989.

[12] J. M. Rehg, "Behavior imaging: Using computer vision to study autism." *MVA*, vol. 11, pp. 14–21, 2011.

[13] J. M. Rehg, A. Rozga, G. D. Abowd, and M. S. Goodwin, "Behavioral imaging and autism," *IEEE Pervasive Computing*, vol. 13, no. 2, pp. 84–87, 2014.

[14] K. L. Carpenter, J. Hahemi, K. Campbell, S. J. Lippmann, J. P. Baker, H. L. Egger, S. Espinosa, S. Vermeer, G. Sapiro, and G. Dawson, "Digital behavioral phenotyping detects atypical pattern of facial expression in toddlers with autism," *Autism Research*, vol. 14, no. 3, pp. 488–499, 2021.

[15] L. Wen, X. Li, G. Guo, and Y. Zhu, "Automated depression diagnosis based on facial dynamic analysis and sparse coding," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 7, pp. 1432–1441, 2015.

[16] Y. Zhu, Y. Shang, Z. Shao, and G. Guo, "Automated depression diagnosis based on deep networks to encode facial appearance and dynamics," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 578–584, 2017.

[17] G. Yang, J. Deng, G. Pang, H. Zhang, J. Li, B. Deng, Z. Pang, J. Xu, M. Jiang, P. Liljeberg *et al.*, "An iot-enabled stroke rehabilitation system based on smart wearable armband and machine learning," *IEEE journal of translational engineering in health and medicine*, vol. 6, pp. 1–10, 2018.

[18] S. Thapaliya, "Evaluation of eeg and eye movement with machine learning for the classification of autism spectrum disorder," Ph.D. dissertation, 2018.

[19] A. B. Dris, A. Alsalman, A. Al-Wabil, and M. Aldosari, "Intelligent gaze-based screening system for autism," in *2019 2nd International Conference on Computer Applications & Information Security (ICCAIS)*. IEEE, 2019, pp. 1–5.

[20] W. Wei, Z. Liu, L. Huang, A. Nebout, and O. Le Meur, "Saliency prediction via multi-level features and deep supervision for children with autism spectrum disorder," 2019.

[21] H. Duan, G. Zhai, X. Min, Y. Fang, Z. Che, X. Yang, C. Zhi, H. Yang, and N. Liu, "Learning to predict where the children with asd look," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 704–708.

[22] Y. Tao and M.-L. Shyu, "Sp-asdnet: Cnn-lstm based asd classification model using observer scanpaths," in *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2019, pp. 641–646.

[23] A. Zunino, P. Morerio, A. Cavallo, C. Ansuini, J. Podda, F. Battaglia, E. Veneselli, C. Becchio, and V. Murino, "Video gesture analysis for autism spectrum disorder detection," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 3421–3426.

[24] A. Crippa, C. Salvatore, P. Perego, S. Forti, M. Nobile, M. Molteni, and I. Castiglioni, "Use of machine learning to identify children with autism and their motor abnormalities," *Journal of autism and developmental disorders*, vol. 45, no. 7, pp. 2146–2156, 2015.

[25] Y. Tian, X. Min, G. Zhai, and Z. Gao, "Video-based early asd detection via temporal pyramid networks," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2019, pp. 272–277.

[26] S. Eraslan, Y. Yesilada, V. Yaneva, and S. Harper, "Autism detection based on eye movement sequences on the web: a scanpath trend analysis approach," in *Proceedings of the 17th International Web for All Conference*, 2020, pp. 1–10.

[27] K. Ahuja, A. Bose, M. Jain, K. Dey, A. Joshi, K. Achary, B. Varkey, C. Harrison, and M. Goel, "Gaze-based screening of autistic traits for adolescents and young adults using prosaic videos," in *Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies*, 2020, pp. 324–324.

[28] J. Li, Y. Zhong, and G. Ouyang, "Identification of asd children based on video data," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 367–372.

[29] G. Wan, X. Kong, B. Sun, S. Yu, Y. Tu, J. Park, C. Lang, M. Koh, Z. Wei, Z. Feng *et al.*, "Applying eye tracking to identify autism spectrum disorder in children," *Journal of autism and developmental disorders*, vol. 49, no. 1, pp. 209–215, 2019.

[30] D. N. Fernández, F. B. Porras, R. H. Gilman, M. V. Mondonedo, P. Sheen, and M. Zimic, "A convolutional neural network for gaze preference detection: A potential tool for diagnostics of autism spectrum disorder in children," *arXiv preprint arXiv:2007.14432*, 2020.

[31] W. Liu, M. Li, and L. Yi, "Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework," *Autism Research*, vol. 9, no. 8, pp. 888–898, 2016.

[32] M. Jiang, S. Francis, D. Srishyla, C. Conelea, Q. Zhao, and S. Jacob, "Classifying individuals with asd through facial emotion recognition and eye-tracking," 07 2019.

[33] D. A. Kaliukhovich, N. V. Manyakov, A. Bangerter, and G. Pandina, "Context modulates attention to faces in dynamic social scenes in children and adults with autism spectrum disorder," *Journal of Autism and Developmental Disorders*, pp. 1–14, 2021.

[34] 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, p. 4542, 2019.

[35] M. Beary, A. Hadsell, R. Messersmith, and M.-P. Hosseini, "Diagnosis of autism in children using facial analysis and deep learning," *arXiv preprint arXiv:2008.02890*, 2020.

[36] T. Akter, M. H. Ali, M. Khan, M. Satu, M. Uddin, S. A. Alyami, S. Ali, A. Azad, M. A. Moni *et al.*, "Improved transfer-learning-based facial recognition framework to detect autistic children at an early stage," *Brain Sciences*, vol. 11, no. 6, p. 734, 2021.

[37] A. Lu and M. Perkowski, "Deep learning approach for screening autism spectrum disorder in children with facial images and analysis of ethnoracial factors in model development and application," *Brain Sciences*, vol. 11, no. 11, p. 1446, 2021.

[38] A. E. Kowallik, M. Pohl, and S. R. Schweinberger, "Facial imitation improves emotion recognition in adults with different levels of subclinical autistic traits," *Journal of Intelligence*, vol. 9, no. 1, p. 4, 2021.

[39] F. Lecciso, A. Levante, R. A. Fabio, T. Caprì, M. Leo, P. Carcagnì, C. Distante, P. L. Mazzeo, P. Spagnolo, and S. Petrocchi, "Emotional expression in children with asd: A pre-study on a two-group pre-post-test design comparing robot-based and computer-based training," *Frontiers in Psychology*, p. 2826, 2021.

[40] C. Guo, K. Zhang, J. Chen, R. Xu, and L. Gao, "Design and application of facial expression analysis system in empathy ability of children with autism spectrum disorder," in *2021 16th Conference on Computer Science and Intelligence Systems (FedCSIS)*. IEEE, 2021, pp. 319–325.

[41] B. Elshoky, O. A. S. Ibrahim, A. A. Ali, and E. M. Younis, "Comparing automated and non-automated machine learning for autism spectrum disorders classification using facial images," 2021.

[42] A. Bangerter, M. Chatterjee, J. Manfredonia, N. V. Manyakov, S. Ness, M. A. Boice, A. Skalkin, M. S. Goodwin, G. Dawson, R. Hendren *et al.*, "Automated recognition of spontaneous facial expression in individuals with autism spectrum disorder: parsing response variability," *Molecular autism*, vol. 11, no. 1, pp. 1–15, 2020.

[43] B. Banire, D. Al Thani, M. Qaraqe, and B. Mansoor, "Face-based attention recognition model for children with autism spectrum disorder," *Journal of Healthcare Informatics Research*, vol. 5, no. 4, pp. 420–445, 2021.

[44] J. Q. Zlibut, A. Munshi, G. Biswas, and C. Cascio, "Identifying and describing subtypes of spontaneous empathic facial expression production in autistic adults," 2021.

[45] G. Alvari, C. Furlanello, and P. Venuti, "Is smiling the key? machine learning analytics detect subtle patterns in micro-expressions of infants with asd," *Journal of clinical medicine*, vol. 10, no. 8, p. 1776, 2021.

[46] D. K. Oller, P. Niyogi, S. Gray, J. Richards, J. Gilkerson, D. Xu, U. Yapanel, and S. Warren, "Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development," *Proceedings of the National Academy of Sciences*, vol. 107, no. 30, pp. 13 354–13 359, 2010.

[47] C. Ecker, A. Marquand, J. Mourão-Miranda, P. Johnston, E. M. Daly, M. J. Brammer, S. Maltezos, C. M. Murphy, D. Robertson, S. C. Williams *et al.*, "Describing the brain in autism in five dimensions—magnetic resonance imaging-assisted diagnosis of autism spectrum disorder using a multiparameter classification approach," *Journal of Neuroscience*, vol. 30, no. 32, pp. 10 612–10 623, 2010.

[48] Z. Sherkatghanad, M. Akhondzadeh, S. Salari, M. Zomorodi-Moghadam, M. Abdar, U. R. Acharya, R. Khosrowabadi, and V. Salari, "Automated detection of autism spectrum disorder using a convolutional neural network," *Frontiers in neuroscience*, vol. 13, p. 1325, 2020.

[49] F. Thabtah and D. Peebles, "A new machine learning model based on induction of rules for autism detection," *Health informatics journal*, vol. 26, no. 1, pp. 264–286, 2020.

[50] G. Devika Varshini and R. Chinnaiyan, "Optimized machine learning classification approaches for prediction of autism spectrum disorder," *Ann Autism Dev Disord. 2020; 1 (1)*, vol. 1001.

[51] I. M. Nasser, M. Al-Shawwa, and S. S. Abu-Naser, "Artificial neural network for diagnose autism spectrum disorder," 2019.

[52] S. Raj and S. Masood, "Analysis and detection of autism spectrum disorder using machine learning techniques," *Procedia Computer Science*, vol. 167, pp. 994–1004, 2020.

[53] J. Peral, D. Gil, S. Rotbei, S. Amador, M. Guerrero, and H. Moradi, "A machine learning and integration based architecture for cognitive disorder detection used for early autism screening," *Electronics*, vol. 9, no. 3, p. 516, 2020.

[54] M. D. Hossain, M. A. Kabir, A. Anwar, and M. Z. Islam, "Detecting autism spectrum disorder using machine learning," *arXiv preprint arXiv:2009.14499*, 2020.

[55] C. Küpper, S. Stroth, N. Wolff, F. Hauck, N. Kliewer, T. Schad-Hansjosten, I. Kamp-Becker, L. Poustka, V. Roessner, K. Schultebraucks *et al.*, "identifying predictive features of autism spectrum disorders in a clinical sample of adolescents and adults using machine learning," *Scientific reports*, vol. 10, no. 1, pp. 1–11, 2020.

[56] F. Z. Subah, K. Deb, P. K. Dhar, and T. Koshiba, "A deep learning approach to predict autism spectrum disorder using multisite resting-state fmri," *Applied Sciences*, vol. 11, no. 8, p. 3636, 2021.

[57] K. A. Pelphrey, N. J. Sasson, J. S. Reznick, G. Paul, B. D. Goldman, and J. Piven, "Visual scanning of faces in autism," *Journal of autism and developmental disorders*, vol. 32, no. 4, pp. 249–261, 2002.

[58] A. Klin, W. Jones, R. Schultz, F. Volkmar, and D. Cohen, "Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism," *Archives of general psychiatry*, vol. 59, no. 9, pp. 809–816, 2002.

[59] M. Freeth, T. Foulsham, and P. Chapman, "The influence of visual saliency on fixation patterns in individuals with autism spectrum disorders," *Neuropsychologia*, vol. 49, no. 1, pp. 156–160, 2011.

[60] S. Wang, M. Jiang, X. M. Duchesne, E. A. Laugeson, D. P. Kennedy, R. Adolphs, and Q. Zhao, "Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking," *Neuron*, vol. 88, no. 3, pp. 604–616, 2015.

[61] S. N. Rigby, B. M. Stoesz, and L. S. Jakobson, "Gaze patterns during scene processing in typical adults and adults with autism spectrum disorders," *Research in Autism Spectrum Disorders*, vol. 25, pp. 24–36, 2016.

[62] R. M. Jones, A. Southerland, A. Hamo, C. Carberry, C. Bridges, S. Nay, E. Stubbs, E. Komarow, C. Washington, J. M. Rehg *et al.*, "Increased eye contact during conversation compared to play in children with autism,"

Journal of autism and developmental disorders, vol. 47, no. 3, pp. 607–614, 2017.

[63] S. R. Edmunds, A. Rozga, Y. Li, E. A. Karp, L. V. Ibanez, J. M. Rehg, and W. L. Stone, "Brief report: using a point-of-view camera to measure eye gaze in young children with autism spectrum disorder during naturalistic social interactions: a pilot study," *Journal of autism and developmental disorders*, vol. 47, no. 3, pp. 898–904, 2017.

[64] A. Vernetti, A. Senju, T. Charman, M. H. Johnson, T. Gliga, B. Team *et al.*, "Simulating interaction: Using gaze-contingent eye-tracking to measure the reward value of social signals in toddlers with and without autism," *Developmental cognitive neuroscience*, vol. 29, pp. 21–29, 2018.

[65] E. L. Ajodan, E. Clark-Whitney, B. Silver, M. R. Silverman, A. Southerland, E. Barnes, S. Dikker, C. Lord, J. M. Rehg, A. Rozga *et al.*, "Increased eye contact during parent-child versus clinician-child interactions in young children with autism," 2019.

[66] M.-K. Kwon, A. Moore, C. C. Barnes, D. Cha, and K. Pierce, "Typical levels of eye-region fixation in toddlers with autism spectrum disorder across multiple contexts," *Journal of the American Academy of Child & Adolescent Psychiatry*, 2019.

[67] D. P. Kennedy and R. Adolphs, "Perception of emotions from facial expressions in high-functioning adults with autism," *Neuropsychologia*, vol. 50, no. 14, pp. 3313–3319, 2012.

[68] S. Wang and R. Adolphs, "Reduced specificity in emotion judgment in people with autism spectrum disorder," *Neuropsychologia*, vol. 99, pp. 286–295, 2017.

[69] M. H. Black, N. T. 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.

[70] N. N. Capriola-Hall, A. T. Wieckowski, D. Swain, V. Tech, S. Aly, A. Youssef, A. L. Abbott, and S. W. White, "Group differences in facial emotion expression in autism: Evidence for the utility of machine classification," *Behavior therapy*, vol. 50, no. 4, pp. 828–838, 2019.

[71] A. C. Miu, S. E. Pană, and J. Avram, "Emotional face processing in neurotypicals with autistic traits: implications for the broad autism phenotype," *Psychiatry research*, vol. 198, no. 3, pp. 489–494, 2012.

[72] R. C. Leung, E. W. Pang, D. Cassel, J. A. Brian, M. L. Smith, and M. J. Taylor, "Early neural activation during facial affect processing in adolescents with autism spectrum disorder," *NeuroImage: Clinical*, vol. 7, pp. 203–212, 2015.

[73] P. Shah, G. Bird, and R. Cook, "Face processing in autism: Reduced integration of cross-feature dynamics," *cortex*, vol. 75, pp. 113–119, 2016.

[74] G. Dawson, S. J. Webb, and J. McPartland, "Understanding the nature of face processing impairment in autism: insights from behavioral and electrophysiological studies," *Developmental neuropsychology*, vol. 27, no. 3, pp. 403–424, 2005.

[75] P. H. J. M. Vlamings, L. M. Jonkman, E. van Daalen, R. J. van der Gaag, and C. Kemner, "Basic abnormalities in visual processing affect face processing at an early age in autism spectrum disorder," *Biological psychiatry*, vol. 68, no. 12, pp. 1107–1113, 2010.

[76] U. Rutishauser, O. Tudusciuc, S. Wang, A. N. Mamelak, I. B. Ross, and R. Adolphs, "Single-neuron correlates of atypical face processing in autism," *Neuron*, vol. 80, no. 4, pp. 887–899, 2013.

[77] 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. 1–20, 2020.

[78] J. A. A. van Rentergem, M. K. Deserno, and H. M. Geurts, "Validation strategies for subtypes in psychiatry: A systematic review of research on autism spectrum disorder," *Clinical Psychology Review*, p. 102033, 2021.

[79] A. Saranya and R. Anandan, "Figs-deaf: an novel implementation of hybrid deep learning algorithm to predict autism spectrum disorders using facial fused gait features," *Distributed and Parallel Databases*, pp. 1–26, 2021.

[80] L. R. Qualls and B. A. Corbett, "Examining the relationship between social communication on the ados and real-world reciprocal social communication in children with asd," *Research in autism spectrum disorders*, vol. 33, pp. 1–9, 2017.

[81] J. R. Pruette, "Autism diagnostic observation schedule-2 (ados-2)," *Google Scholar*, 2013.

[82] V. Hus and C. Lord, "The autism diagnostic observation schedule, module 4: revised algorithm and standardized severity scores," *Journal of autism and developmental disorders*, vol. 44, no. 8, pp. 1996–2012, 2014.

[83] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[84] B. Jiang, M. F. Valstar, and M. Pantic, "Action unit detection using sparse appearance descriptors in space-time video volumes," in *Face and Gesture 2011*. IEEE, 2011, pp. 314–321.

[85] Z. Jiang, Z. Lin, and L. S. Davis, "Learning a discriminative dictionary for sparse coding via label consistent k-svd," in *CVPR 2011*. IEEE, 2011, pp. 1697–1704.

[86] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on signal processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[87] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 1, pp. 40–51, 2007.

[88] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM review*, vol. 51, no. 1, pp. 34–81, 2009.

[89] M. Elad, *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer Science & Business Media, 2010.

[90] R. Rubinstein, T. Faktor, and M. Elad, "K-svd dictionary-learning for the analysis sparse model," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 5405–5408.

[91] S. Yang, L. Liu, and M. Xu, "Free lunch for few-shot learning: Distribution calibration," *arXiv preprint arXiv:2101.06395*, 2021.

[92] T. Ramalho and M. Garnelo, "Adaptive posterior learning: few-shot learning with a surprise-based memory module," *ICLR*, 2019.