Visual Attention Analysis and Prediction on Human Faces for Children with Autism Spectrum Disorder

HUIYU DUAN, XIONGKUO MIN, YI FANG, LEI FAN, XIAOKANG YANG, and GUANGTAO ZHAI, MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University

The focus of this article is to analyze and predict the visual attention of children with Autism Spectrum Disorder (ASD) when looking at human faces. Social difficulties are the hallmark features of ASD and will lead to atypical visual attention toward various stimuli more or less, especially on human faces. Learning the visual attention of children with ASD could contribute to related research in the field of medical science, psychology, and education. We first construct a Visual Attention on Faces for Autism Spectrum Disorder (VAFA) database, which consists of 300 natural scene images with human faces and corresponding eye movement data collected from 13 children with ASD. Compared with matched typically developing (TD) controls, we quantify atypical visual attention on human faces in ASD. Statistics show that some high-level factors such as face size, facial features, face pose, and facial emotions have different impacts on the visual attention of children with ASD. Combining the feature maps extracted from the state-of-the-art saliency models, we get the visual attention model on human faces for individuals with ASD. The proposed model shows the best performance among all competitors. With the help of our proposed model, researchers in related fields could design specialized education contents containing human faces for the children with ASD or produce the specific model for rapidly screening ASD using their eye movement data.

CCS Concepts: • **Computing methodologies → Model development and analysis**;

Additional Key Words and Phrases: Visual attention, saliency prediction, human faces, autism spectrum disorder (ASD)

ACM Reference format:
Huiyu Duan, Xiongkuo Min, Yi Fang, Lei Fan, Xiaokang Yang, and Guangtao Zhai. 2019. Visual Attention Analysis and Prediction on Human Faces for Children with Autism Spectrum Disorder. *ACM Trans. Multimedia Comput. Commun. Appl.* 15, 3s, Article 90 (October 2019), 23 pages.
https://doi.org/10.1145/3337066

1 INTRODUCTION

Autism Spectrum Disorder (ASD) is one type of neurodevelopmental disorder. Sensory symptoms are identified to be the core characteristics of the neurobiology of autism [52]. As an important

This work was supported in part by the National Natural Science Foundation of China under Grants 61831015 and 61527804, in part by the National Key Research and Development Program of China under Grant 2016YFB1001003, in part by the STCSM under Grant 18DZ1112300, in part by the Shanghai Municipal Commission of Health and Family Planning under Grant 2018ZHYL0210, and in part by the China Postdoctoral Science Foundation under Grants BX20180197 and 2019M651496.
Authors' address: H. Duan, X. Min, Y. Fang, L. Fan, X. Yang, and G. Zhai (corresponding author), MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, 800 Dongchuan Rd., Shanghai 200240, China; emails: huiyuduan@sjtu.edu.cn, minxiongkuo@gmail.com, {yifang, lei.fan, xkyang, zhaiguangtao}@sjtu.edu.cn.

1551-6857/2019/10-ART90 $15.00
https://doi.org/10.1145/3337066

aspect of sensory perception, atypical visual attention is often observed in people with ASD [56]. Eye movements encode rich information about attention, oculomotor control, and psychological factors of an individual. Cognition condition could be reflected from the eye movements as well. Several attention differences between individuals with ASD and healthy people have been reported in the literature [18], including reduced joint-attention behaviors [48], reduced attention to social scenes [12], and preference to low-level features of the stimuli [64]. In brief, individuals with autism show reduced attention to social stimuli (i.e., faces, conversations, etc.) but pay more attention to non-social stimuli (i.e., vehicles, electronics, etc.) [17, 54].

Since faces are important social cues, research on the face processing for ASD has attracted the attention of many researchers. Compared to the normal population, individuals with ASD have impairments in face recognition or discrimination [26, 34, 36]. Existing eye-tracking experiments consistently demonstrate that people with ASD have reduced visual attention to faces compared to the controls [24]. Regarding visual attention on core facial features, some studies found reduced visual attention of people with ASD to these regions [14, 35], while other studies reported no differences in gaze patterns between ASD and typically developing (TD) individuals [6, 23]. Moreover, the influence of facial emotions on the visual attention between ASD individuals and healthy people is different [2].

Previous research in the visual attention of ASD individuals on human faces used restricted stimuli, i.e., faces in isolation or on a similar background. Moreover, the small number of images in previous research is not sufficient to conduct a systematic analysis for the visual attention of ASD patients on human faces from low-level visual features to high-level semantic features. Thus, in this article, we first establish a Visual Attention on Faces for Autism Spectrum Disorder (VAFA) database. We collect 300 images from an open face database [45]. These images contain faces of various ages, genders, sizes, poses, expressions, and so forth. Then we perform eye-tracking experiments on 13 individuals with ASD and 15 healthy people, respectively. The ASD group and TD group are matched on race, age, gender, and education. A facial behavior analysis toolkit [5] is used to detect facial landmark, face pose [68], and Action Units (AUs) [4]. 300 face images, the corresponding eye-tracking data of ASD individuals and healthy individuals, the facial landmark localization, face pose, and AU detection results together constitute the VAFA database. As far as we know, the VAFA database is the largest database of its kind which contains natural scene images with one face as the main content of the image.

With the advent of deep neural networks (DNNs), state-of-the-art deep saliency models could automatically incorporate semantic features and achieve great performance. Nevertheless, people with ASD have atypical visual attention; the visual attention model should be retraining or redesigned. Due to the difficulty of obtaining the eye movement data of individuals with ASD, we only have 300 images with the fixation data of 13 ASD participants. And it is inadequate to train an end-to-end deep neural network. Thus, based on our VAFA database, we first analyze the visual attention differences between individuals with ASD and healthy people on human faces from four influencing factors, which are facial proportion, core facial features, face pose, and facial expressions. Then we extract special feature maps for ASD and integrate these feature maps adaptively with features extracted from the CASNet [25], which is a state-of-the-art saliency prediction model for healthy individuals. Finally, through fine-tuning the network, we get the visual attention model for ASD on human faces. The construct model has achieved the best performance so far.

Analysis and prediction of the visual attention of ASD on human faces have great significance to related research fields. With the help of our VAFA database and saliency model designed for ASD, researchers could characterize the visual attention traits of ASD on human faces and understand ASD better [63]. Prediction of the visual attention of ASD on human faces has many application scenarios as well. Based on our VAFA database and proposed saliency model, special-

ized textbooks containing human faces could be designed for individuals with ASD. Similar to the applications of general visual attention models in signal processing research [20, 43, 44, 46, 69, 72], we can also use the saliency models designed for the ASD to develop relevant signal processing techniques specifically for the ASD. These techniques could be used to evaluate the mental condition of ASD individuals during their interaction with people and give specific suggestions. Furthermore, learning the visual attention of people with ASD could be used to classify the gaze patterns of ASD individuals and healthy people [32]. Since the diagnostic procedure of ASD is expensive, subjective, and time-consuming, it can be of great value to use visual attention methods to assist the diagnosis of ASD. Moreover, deep CNNs have been used in visual related fields for many years. Nowadays, interpretable representations of CNNs are widely studied. Disentangled representation learning or visually interpretable representation learning for the visual attention of individuals with ASD on human faces may be a future research direction [58, 67, 71].

The remainder of this article is organized as follows. In Section 2, we briefly review the related works. Section 3 describes the procedure of our eye-tracking experiments and detailed information of our database. In Section 4, we analyze the difference of visual attention on human faces between individuals with ASD and healthy controls. In Section 5, we propose and evaluate our model. We present our conclusions in Section 6.

2 RELATED WORK

2.1 Visual-Attention Model of Healthy People

Humans have a remarkable ability to focus on the salient regions in a scene [10, 41], which allows us to handle large amounts of visual information efficiently. This neural mechanism of the human visual system (HVS) is known as visual attention. Most traditional visual saliency models belong to a bottom-up mechanism, which usually consists of three cascaded steps: visual feature extraction, saliency inference, and saliency integration. These bottom-up visual saliency models mainly adopted various hand-designed features, including low-level features [30, 39], middle-level features [40], high-level features [8, 33], audio-visual features [47], and motion features [28]. Deep neural networks (DNNs) are state-of-the-art architectures in machine learning. With the help of DNNs, the saliency prediction task has achieved significant improvement [37, 60–62]. Huang et al. [29] computed the saliency map through concatenating fine and coarse features extracted from two stream convolutional networks. Cornia et al. [15] proposed to combine multi-level features extracted from the VGG net and then obtain the saliency map. Cornia et al. [16] also used a convolutional LSTM to enhance the feature maps extracted from a Dilated Convolutional Network. Pan et al. [49] proposed to use a generative adversarial network (GAN) to calculate the saliency map. Fan et al. [25] proposed to add a subnetwork for contextual saliency weight computation after the SALICON net [29].

2.2 Visual-Attention Analysis of Individuals with ASD

Since it is important to characterize the visual attention of ASD, many studies related to this topic have been conducted [56]. Birmingham et al. [7] found reduced attention to social scenes in ASD using a natural scene as stimuli. Sasson and Touchstone [55] researched the preference of the visual attention of individuals with ASD using competing social and object images and reported that ASD individuals attended less than controls to faces. Wang et al. [64] demonstrated that ASD individuals have impaired social attention in visual search. Wang et al. [63] also quantified the atypical visual attention of ASD through multi-level features. They proposed that ASD individuals are attracted more by low-level features (such as contrast, color, and orientation). Chevallier et al. [13] measured social attention of ASD through multi-task analysis. Mcpartland et al. [42] studied

atypical visual attention patterns in individuals with ASD when faces and objects were used as stimuli. Rice et al. [51] studied the visual scanning strategies of social scenes in school-aged children with ASD and their relationship to early social disability measurement. Samad et al. [53] also reported spontaneous visual responses to stimuli in individuals with ASD may be used as behavior markers for them. In the previous work [19], we constructed a saliency prediction for children with the ASD (SPCA) dataset. The SPCA dataset includes 500 images with corresponding fixation data of subjects with ASD and TD subjects. The images in the SPCA dataset cover various types of contents. However, the number of images for each type is limited. This limited number is not sufficient to analyze the atypical traits of the visual attention of the people with ASD to specific types of stimuli.

2.3 Visual Attention of the People with ASD on Human Faces

One of the most-studied areas of visual processing in ASD is face processing, because ASD is characterized as a social deficit, and faces are believed to be the most "social" visual stimuli. Amso et al. [1] investigated the bottom-up attention orienting to faces across various developmental participants and showed the difference across various ages. Vabalas and Freeth [59] analyzed the eye movements' patterns during a face-to-face interaction and suggested that the individual differences were related to the amount of traits of individuals with ASD. Åsberg Johnels et al. [2] analyzed the gaze patterns of ASD individuals on various emotional faces. They showed that different emotional content causes diverse gaze behavior and the effect is different between individuals with ASD and healthy people. Yi et al. analyzed the difference of gaze pattern on human faces for the ASD and TD using multi-method analysis [65, 66]. Although this topic is important, in our previous work [19] or other researchers' works, the limited number of images containing human faces is not sufficient to analyze the atypical visual attention of the people with ASD to human faces. Thus, in this article, we constructed the visual attention on faces for the Autism Spectrum Disorder (VAFA) database.

3 EYE-TRACKING EXPERIMENTS

Detailed procedures of eye-tracking experiments are introduced in this section. First, we describe the stimuli and apparatus we used in the experiments. Then the basic information of subjects including children with ASD and their matched controls are presented. Next, we introduce the detailed experimental procedures. Finally, we provide detailed construction information for the VAFA database.

3.1 Stimuli and Apparatus

The images we used as stimuli are collected from an open face database [45]. This database has 481 source images with resolutions of 1,280 × 960, 1,024 × 1,024, or 768 × 1,024 (width × height). The collected images contain faces of various sizes, poses, emotions, ages, genders, and so forth. We select 300 images from this database considering the balance of various emotions and whether the content is appropriate for children. Finally, the selected images are classified into six expressions. The six expressions are generally positive, very positive, neutral, generally negative, very negative, and complex expressions, respectively, as shown in Figure 1. Each expression has 50 images in our database.

The apparatus we used to display stimuli and record eye movement data is Tobii T120 Eye Tracker, which has a 17-inch display with the resolution of 1,280 × 1,024 (width × height) pixels. Subjects are seated around 65cm from the eye tracker. The sampling rate is set to 120Hz.

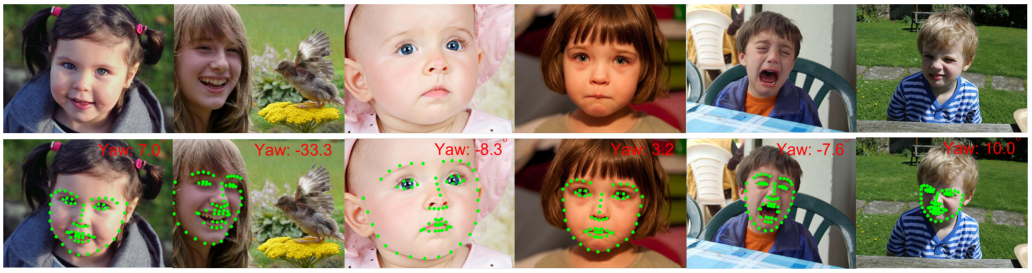


Fig. 1. Six different sample expressions with their corresponding face landmarks and estimated face poses (yaw angles). The first line shows raw images. The image from the left to the right represents the expressions of generally positive, very positive, neutral, generally negative, very negative, and complex, respectively. The second line shows the face landmarks and estimated face poses of the images above.

Table 1. Descriptive Characteristics of All Participants

Group	Total: Male	Handed	Age (Years)		Performance IQ	
			Range	M (SD)	Range	M (SD)
ASD	13:11	12 Right 1 Left	5.0–12.3	7.8 (2.1)	80–113	97.1 (9.9)
TD	15:12	13 Right 2 Left	5.2–12.1	8.0 (2.0)	90–120	105.6 (8.9)

3.2 Subjects

Nineteen high-functioning children with ASD were recruited. All ASD participants met DSM-V diagnostic criteria for autism [3]. Because it is difficult for the children with ASD to concentrate on the screen, only 13 subjects could complete the calibration step and obtain effective eye movement data. Among the six participants whom cannot complete the experiments, four of them did not look at the screen at all, and the other two individuals have large errors in calibration. We measured errors of the calibration using the calibration error vectors in the calibration result figure of "Tobii Studio." We would recalibrate it if the length of the calibration error vectors is more than twice the diameter of the standard circle given by "Tobii Studio." The two individuals who have large errors in calibration cannot provide good calibration data in any session of the experiment. Thus, we exclude the eye movement data of these two participants in this study. The age of the remaining participants with ASD ranged from 5 years old to 12 years old and the mean age of the subjects was 7.8 years old. Fifteen healthy children were recruited as controls. The age of the healthy children also ranged from 5 years old to 12 years old and the mean age was 8 years old. Besides the age, to guarantee the generalization of our database and saliency model, we also matched the gender, handedness, age, and performance IQ of the two groups. Table 1 presents the descriptive information of all participants in detail. All participants had normal or correct-to-normal visual acuity. Before the experiments, the parents of all participants gave written informed consent.

3.3 Experimental Procedures

Three hundred images were shuffled into a random sequence. Due to the lack of patience of ASD participants, the experiment was split into 10 recording sessions. Each session has 30 randomly selected images. Test images were displayed in a random order at full resolution for 3 seconds. Each image was followed by a 1-second gray screen mask. At the beginning of each session, the eye-tracking calibration process was conducted to ensure the reliability of the data. And before

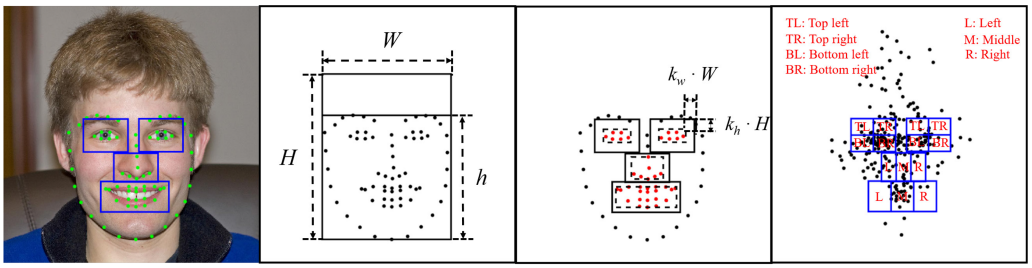


Fig. 2. Definition of some facial regions on human faces. (a) Example image with facial landmark and ROIs on it. (b) Definition of face size. W and H are face width and height, respectively. $H = \frac{4}{3} \cdot h$, and face size $S = \sqrt{W \cdot H}$. (c) ROIs creation. Dashed rectangles show the smallest rectangles cover the key landmark (marked as red). Solid rectangles show the enlarged ROIs. Enlarging factors are different for different ROIs. Detailed enlarging factors can be found in Table 2 in [45]. (d) Segmentation of ROIs. The eye region is segmented into four regions: Top Left (TL), Top Right (TR), Bottom Left (BL), and Bottom Right (BR), respectively. The nose region and mouth region are segmented into three regions: Left (L), Middle (M), and Right (R), respectively.

the experiments, subjects were told to look at images freely. However, because of difficulty concentrating, we had to remind the ASD participants to look at images during the experiments. The same experimental procedures are also conducted for the healthy controls. The experiments lasted 10 weeks with one week one session.

3.4 The VAFA Database

From the eye movement data obtained in the experiments, the fixation map is generated. We overlay the fixation points of all ASD participants into one map and get the fixation map of individuals with ASD. With the same methods, we get the fixation map of controls. The fixation map is then smoothed with a Gaussian kernel to generate the fixation density map (FDM, also called visual attention map). We set the standard deviation of the Gaussian kernel to 1° of visual angle. In our experimental condition, the standard deviation of the Gaussian kernel is set as $\sigma = 40$.

We use facial landmarks as the key points on faces, and then label the regions of interest (ROIs) according to these facial landmarks. The difference of visual attention on human faces between individuals with ASD and controls could be analyzed on these ROIs. A facial behavior analysis toolkit [5] is used to detect facial landmark and estimate the pose and emotion of faces in this article. We use the CE-CLM approach [68] in this toolkit to localize 66 landmark points and estimate face pose in this article. It is a state-of-the-art face landmark detection approach. Figure 1 shows the landmarks and face poses in yaw detected using this method. And we use the AU detection system as stated in [4] to detect facial expressions.

Some facial areas are defined and included in the VAFA database. As shown in Figure 2(a), W and h are the width and height of the rectangles which could cover all landmarks. W is also defined as the face width, and the face height is defined as $H = \frac{4}{3} \cdot h$. Thus, the face size S is defined by $S = \sqrt{W \cdot H}$. Facial ROIs are defined in Figure 2(b). Dashed rectangles show the smallest rectangles cover the key landmark (marked as red) and solid rectangles show the enlarged ROIs. The dashed rectangle is enlarged by adding $2 \cdot k_w \cdot W$ to the width and $2 \cdot k_h \cdot H$ to the height, where W, H are face width and height, and k_w, k_h are enlarging factors. Detailed enlarging factors can be found in [45]. Figure 2(d) shows the segmentation methods we used for each ROI. We segment eyes region into four parts: Top Left (TL), Top Right (TR), Bottom Left (BL), and Bottom Right (BR). Nose and

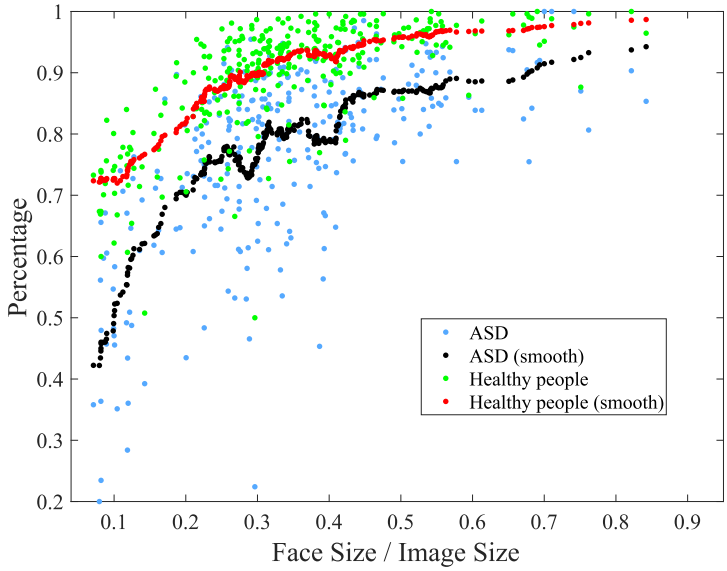


Fig. 3. Percentage of fixation points on the human faces versus facial proportion in the image. Green points represent the percentage of healthy controls' fixation points on faces. Blue points represent the percentage of ASD subjects' fixation points on faces. Red points are obtained by filtering the green points (attribute to healthy controls) with a moving average filter. Black points are obtained by filtering the blue points (attribute to ASD group) with a moving average filter. Vertical axis represents the percentage of fixation points on the human faces. Horizontal axis represents facial proportion.

mouth regions are segmented into three parts: Left (L), Middle (M), and Right (R). We segment the ROI for the following detailed analysis.

4 VISUAL ATTENTION ANALYSIS ON HUMAN FACES: COMPARING BETWEEN INDIVIDUALS WITH ASD AND HEALTHY CONTROLS

4.1 Effect of Face Size

In [45], Normalized Scanpath Saliency (NSS) [50], the shuffled version of Area Under Curve (sAUC) [70], and Correlation Coefficient (CC) are used to evaluate the performance of the GaussFC model. The GaussFC model is simply set a 2D Gaussian kernel at the face center of the image. The results show that with the face size increase, the performance of the GaussFC model decreases. It means that we need a more accurate face saliency model when the face occupies most of the image. Although state-of-the-art deep saliency models could automatically incorporate face features, the eye movement data in our VAFA database is not enough to train an end-to-end specific saliency model for ASD. To get an accurate face saliency model for individuals with ASD, we should extract special feature maps.

Figure 3 illustrates the tendency between the percentage of fixation points on the human faces and facial proportion in the image. The definition of face size is described in Figure 2(b). The green points represent the percentage of healthy controls' fixation points on faces. The blue points represent the percentage of ASD subjects' fixation points on faces. To get an overall view of the tendency, we filter the data by a moving average filter and then scatter the results in Figure 3. The span of the moving average filter is set to 30. The red points represent the tendency of healthy controls and the black points represent the tendency of ASD subjects. The percentage of fixation points on the human faces increases along with the increase of facial proportion in the image, no

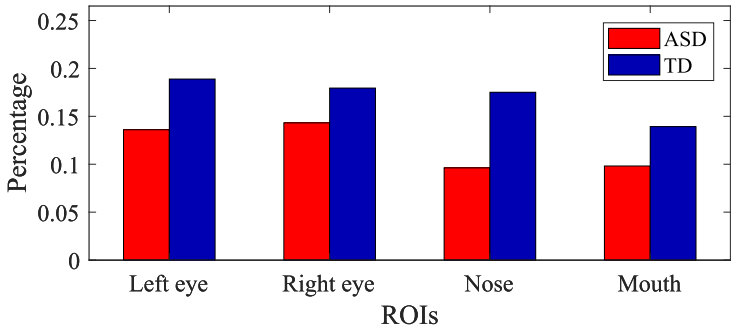


Fig. 4. Overall fixation distribution on faces. Vertical axis denotes the percentage of fixation points falling on corresponding ROIs. Horizontal axis denotes ROIs as defined in Figure 2(c). Red bars represent the percentages of fixation points of the ASD group. Blue bars represent the percentages of fixation points of the TD group.

matter for ASD participants or healthy controls. By comparing the red points and black points, it is obvious that healthy people fixate more on human faces than the people with ASD. The percentage difference is about 10% on average. Thus, we use a smaller Gaussian kernel at the center of faces and extract a rough face feature map.

4.2 Effect of Facial Features

Facial features such as eyes, nose, and mouth have influences on the visual attention. In this section, we quantify the differences of visual attention between ASD participants and healthy controls in these facial ROIs.

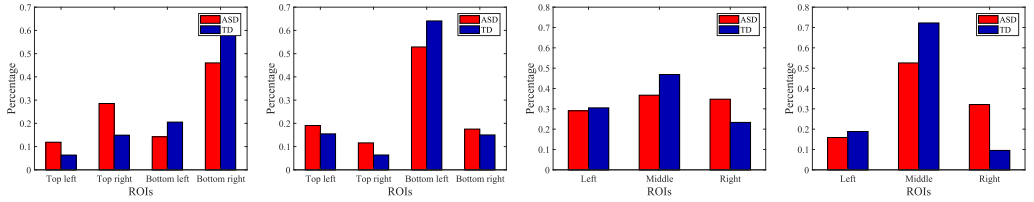
4.2.1 Overall Fixation Distribution Differences. ROIs are defined as shown in Figure 2(c). Based on the definition, we could analyze the overall fixation distribution differences on human faces. Figure 4 illustrates the percentage of fixations in each ROI for ASD participants and healthy controls. Red bars represent the ASD group while blue bars represent the TD group. The vertical axis denotes the percentage of fixations falling into corresponding ROIs. The horizontal axis denotes ROIs. From this figure, we can see that the percentage of fixation points of ASD subjects in each ROI is less than that of healthy controls in each ROI. We believe it is due to the atypical visual attention in ASD. Note that for the ASD group, the percentage of fixations in eyes ROIs is more than it is in nose or mouth ROIs. The hypothesis of excess mouth viewing in autism did not receive support in this study. This phenomenon is in keeping with the majority of studies in the related field of ASD [2, 22, 27]. This phenomenon may be caused by the different definitions of ROIs. Comparing between each ROI, we can see that the percentages of fixations of people with ASD in nose regions and mouth regions are almost the same, while healthy people fixate more on nose regions than on mouth regions. As shown in Table 2, T-statistics show significant differences in these comparisons.

4.2.2 Differences of Fixation Distribution Bias in Specific ROIs. To analyze the detailed fixation distribution differences in each ROI, we segment each ROI into several small symmetrical ROIs. As shown in Figure 2, the eyes region is segmented into four parts: top left, top right, bottom left, and bottom right. The nose and mouth region is segmented into three parts: left, middle, and right. Figure 5 illustrates the differences of fixation distribution bias in each ROI between the people with ASD and healthy controls. Red bars in Figure 5 show the detailed fixation distribution of ASD participants in the left eye region, right eye region, nose region, and mouth region, respectively. Blue bars in Figure 5 show the detailed fixation distribution of healthy subjects in the left eye

Table 2. t-Test for the Between-Group Comparison Over 300 Images for Items in Figure 4

ROIs	Left eye	Right eye	Nose	Mouth
H	1	1	1	1
P	1.67e-36	3.38e-14	2.45e-69	3.57e-24

$H = 1$ indicates that t-test rejects the null hypothesis at the 5% significance level (i.e., the difference between two groups is significant), while $H = 0$ indicates that t-test dose not reject the null hypothesis at the 5% significance level. H: test decision for the null hypothesis; P: p-value for the T-statistic.



(a) Left Eye (b) Right Eye (c) Nose (d) Mouth

Fig. 5. Fixation distribution in each ROI. Vertical axes denote the percentage of fixation points belonging to the small part. Horizontal axes denote the names of small parts. Red bars represent the percentages of fixation points of individuals with ASD. Blue bars represent the percentages of fixation points of the TD group. (a) Fixation distribution in the left eye region. (b) Fixation distribution in the right eye region. (c) Fixation distribution in the nose region. (d) Fixation distribution in the mouth region.

region, right eye region, nose region, and mouth region, respectively. Comparing red bars and blue bars in Figure 5(a) and (b), both ASD participants and healthy subjects look most at the bottom right part of the left eye region and the bottom left part of the right eye region. But for the second salient part, people with ASD tend to look more at the top right of the left eye region and the top left of the right eye region, while healthy subjects are likely to look more at the bottom left of the left eye region and the bottom right of the right eye region. Moreover, it is obvious that the fixation distribution of the people with ASD in the eye region is more dispersed than that of healthy controls. A similar phenomenon also appears in the nose region and mouth region, as shown in Figure 5(c) and (d).

Figure 5(c) and (d) illustrate fixation distribution of ASD participants and healthy controls in the nose and mouth regions, respectively. As shown in Figure 5(c) and (d), for healthy controls, the middle part of both nose and mouth regions are much more salient than other parts. A similar phenomenon also appears in the ASD group. Nevertheless, compared with healthy controls, ASD participants fixate more on the nose or mouth ROIs, while the healthy controls fixate more on the middle part. Table 3 shows the t-test for the between-group comparison over 300 images for small parts in each ROI. Statistics show that nearly all of these differences in corresponding small parts of ROIs are significant.

4.3 Effect of Face Pose

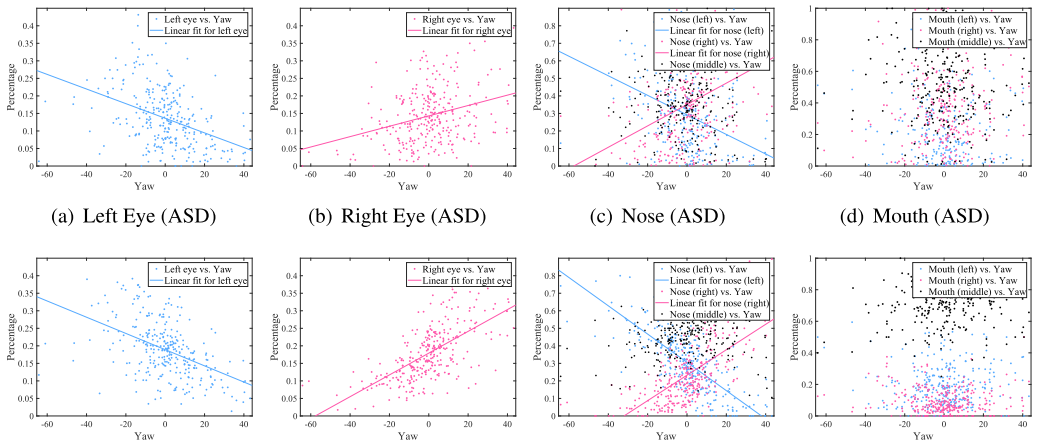
In [45], the authors discussed the effect of face pose on the visual attention of healthy people. In this article, we compare the difference of this effect between ASD subjects and healthy subjects. Figure 6 illustrates the influence of face pose on the fixation distribution of ASD and healthy controls. We mainly consider the yaw of face pose in this article. Figure 6(a)–(d) show the influence of yaw

Table 3. t-Test for the Between-Group Comparison Over 300 Images for Items in Figure 5

ROIs	Left Eye (tl)	Left Eye (tr)	Left Eye (bl)	Left Eye (br)	Right Eye (tl)	Right Eye (tr)	Right Eye (bl)
H	1	1	1	1	1	1	1
P	8.61e-09	3.65e-20	4.41e-10	7.19e-17	1.50e-03	5.75e-09	7.85e-15

ROIs	Right Eye (br)	Nose (l)	Nose (m)	Nose (r)	Mouth (l)	Mouth (m)	Mouth (r)
H	1	0	1	1	1	1	1
P	2.19e-02	2.43e-01	1.36e-10	1.55e-12	1.91e-02	5.95e-26	7.07e-35

$H = 1$ indicates that t-test rejects the null hypothesis at the 5% significance level (i.e., the difference between two groups is significant), while $H = 0$ indicates that t-test does not reject the null hypothesis at the 5% significance level. tl: top lef; tr: top right,; bl: bottom left; br: bottom right; l: left; m: middle; r: right; H: test decision for the null hypothesis; P: p-value for the T-statistic.



(a) Left Eye (ASD) (b) Right Eye (ASD) (c) Nose (ASD) (d) Mouth (ASD)

(e) Left Eye (TD) (f) Right Eye (TD) (g) Nose (TD) (h) Mouth (TD)

Fig. 6. Influence of face pose on the fixation distribution of ASD and healthy controls. For each point in each scatter plot, it represents the percentage of fixation points of the ASD group or the TD group falling into corresponding ROIs. (a)–(d) Scatter plot of the percentages of ASD participants' fixations on the left eye region, right eye region, small parts in the nose region, and small parts in the mouth region, respectively, with varying yaw angle. (e)–(h) Scatter plot of the percentages of healthy subjects' fixations on the left eye region, right eye region, small parts in the nose region, and small parts in the mouth region, respectively, with varying yaw angle. Note that the vertical axes in (a), (b), (e), and (f) denote the ratio of fixations belonging to the left or right eye region to all fixations in the test image, while the vertical axes in (c), (d), (g), and (h) denote the ratio of fixations belonging to the small part (i.e., Left, Middle, and Right) to the fixations belonging to specific ROIs (i.e., nose and mouth).

angle on percentages of fixations of ASD in the left eye region, right eye region, nose region, and mouth region, respectively. Figure 6(e)–(h) show the influence of yaw angle on percentages of fixations of healthy subjects in the left eye region, right eye region, nose region, and mouth region, respectively. We apply a linear model to fit the scatter points in Figure 6:

$$f(Y) = p_0 + p_1 \cdot Y, -56.4 \leq Y \leq 47.0, \tag{1}$$

where the Y is the yaw angle, $f(Y)$ represents the percentage of fixations, and p_0, p_1 represent two fitting parameters obtained by a least-square fitting. Table 4 gives the fitting results, in which *value* represents the estimated value, *SE* indicates standard error, and *p-value* denotes the p-value

Table 4. Parameters of the Fitted Linear Models

Groups	ROIs	p_0			p_1		
		value	*SE*	*p-value*	*value*	*SE*	*p-value*
ASD	Left eye	1.36e-1	4.47e-3	4.02e-93	−2.08e-3	2.75e-3	4.21e-13
	Right eye	1.43e-1	4.39e-3	7.80e-100	1.48e-3	2.69e-4	8.65e-8
	Nose (left)	2.90e-1	1.48e-2	6.29e-55	−5.65e-3	9.33e-4	4.48e-9
	Nose (middle)	3.67e-1	1.40e-2	2.43e-78	−4.08e-4	8.82e-4	6.44e-1
	Nose (right)	3.48e-1	1.56e-2	1.71e-64	6.01e-3	9.87e-4	3.63e-9
	Mouth (left)	1.59e-1	1.13e-2	1.58e-34	−1.73e-3	6.89e-4	1.24e-2
	Mouth (middle)	5.26e-1	1.59e-2	1.27e-98	−9.65e-4	9.74e-4	3.23e-1
	Mouth (right)	3.21e-1	1.57e-2	1.32e-57	2.56e-3	9.58e-4	7.99e-3
TD	Left eye	1.89e-1	4.03e-3	8.78e-139	−2.31e-3	2.48e-4	2.61e-18
	Right eye	1.79e-1	3.49e-3	3.97e-149	3.11e-3	2.14e-4	2.80e-36
	Nose (left)	3.06e-1	6.71e-3	1.72e-135	−8.06e-3	4.12e-4	3.53e-55
	Nose (middle)	4.69e-1	7.03e-3	5.35e-180	8.35e-4	4.32e-4	5.39e-2
	Nose (right)	2.32e-1	7.28e-3	1.01e-97	7.26e-3	4.47e-4	9.54e-43
	Mouth (left)	1.89e-1	7.28e-3	5.65e-78	1.58e-5	4.47e-4	9.72e-1
	Mouth (middle)	7.22e-1	7.58e-3	1.23e-223	4.25e-4	4.66e-4	3.62e-1
	Mouth (right)	9.55e-2	5.21e-3	1.38e-50	−5.02e-4	3.20e-4	1.18e-1

Left: left part; middle: middle part; right: right part. $value$: estimated value. SE: standard error. $p\text{-}value$: p-value for the T-statistic.

for the T-statistic. The T-statistic is used to validate the effectiveness of fitting results. If $p\text{-}value$ is less than 0.001, the fitting results are significant, otherwise, we should discard the fitting results.

For each point in each scatter plot of Figure 6, it represents the percentage of fixation points of the ASD or TD group falling into corresponding ROIs. As shown in Figure 6(a), (b), (e), and (f), in the left eye region, the percentage generally decreases with the yaw angle, and in the right eye region, the percentage generally increases with the yaw angle. In Table 4, $p\text{-}value$ for p_1 of these four subfigures is very small ($p\text{-}value < 0.001$), therefore we believe that yaw angles have an influence on fixation distribution on eyes for both people with ASD and healthy people. Moreover, comparing Figure 6(a) and (e), and (b) and (f), respectively, it is obvious that slopes in (a) and (b) are smaller than that in (e) and (f), and scatter plots are more dispersed in (a) and (b). In Table 4, for the left eye region and the right eye region, the values of p_1 of the ASD group are smaller than that of TD group, and the $p\text{-}values$ of p_1 of the ASD group are much larger than that of the TD group. The statistics in Table 4 demonstrate the phenomena we observed in Figure 6.

For the nose region, Figure 6(c) and (g) show the scatter plots of the percentages of fixations belonging to three small parts (e.g., Left, Middle, and Right) with three different colors. It is obvious that the percentage generally decreases with the yaw angle in the left part of the nose region and increases with the yaw angle in the right part of the nose region for both people with ASD and healthy people. We also perform a similar least-square fitting and list the results in Table 4. Statistics in Table 4 also demonstrate that the phenomena are significant ($p\text{-}value < 0.001$). But for the middle part of the nose region, we could not find out any trend and $p\text{-}values$ of p_1 are larger than 0.001 for both the ASD group and the TD group. For the left part and right part of the nose region, the values of p_1 of the ASD group are smaller than that of the TD group, and the $p\text{-}values$ of p_1 of the ASD group are much larger than that of the TD group. This illustrates that for the nose region, the effect of face pose on the visual attention of people with ASD is smaller than that of TD people. Comparing Figure 6(c) and (g), we can see that the black points in figure (c) are mixed

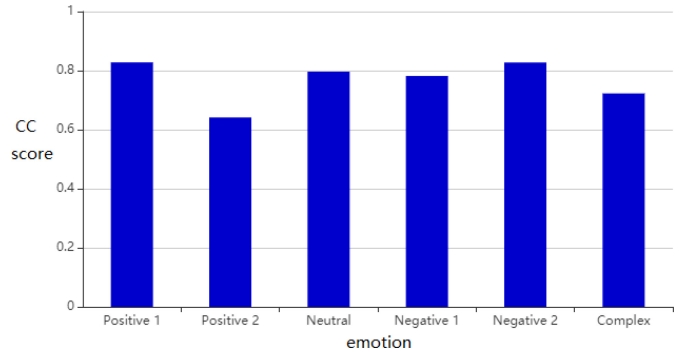


Fig. 7. Linear correlation coefficient (CC) score between ASD participants' saliency map and healthy participants' saliency map under different facial expressions (emotions). Implication of the labels of the horizontal axis are as follows. Positive 1: generally positive facial expressions (i.e., smile); Positive 2: very positive facial expressions (i.e., happy, and laugh); Neutral: neutral facial expressions; Negative 1: generally negative facial expressions (i.e., sad, disappointment, etc; Negative 2: very negative facial expressions (i.e., fear, anger, cry, etc.); and Complex: complex facial expressions (i.e., surprise, strange, etc.).

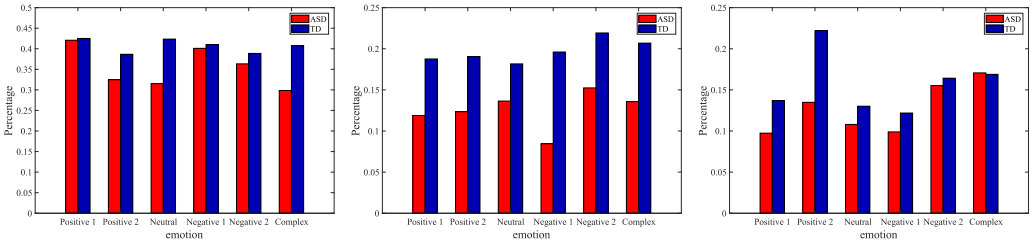
with blue points and pink points. However, in figure (g), the black points are general above the blue points and pink points. This phenomenon confirms the conclusion obtained in Section 4.2.2.

Figure 6(d) and (h) illustrate the percentages of fixations belonging to three small parts (e.g., Left, Middle, and Right) within the mouth region. It seems that for both the ASD group and the TD group, the percentages of fixations of all three parts do not change with varying yaw angle significantly. Statistics in Table 4 also confirm this phenomenon (for the p_1 of the mouth region of both the ASD group and the TD group, $p\text{-}value > 0.001$). Moreover, in Figure 6(h), the middle part (black points) of the mouth region is much more salient than other parts (blue points and pink points). Nevertheless, in Figure 6(d), the three parts have similar salient weights. This phenomenon also confirms the conclusion obtained in Section 4.2.2.

4.4 Effect of Facial Expressions

From the demonstration in Figure 4 and the analysis in Section 4.2.1, we could get the overall fixation distribution of people with ASD on faces. However, facial expressions could also influence fixation distribution on human faces [21] and the influence of facial expressions on average fixation points in each ROI is different between ASD patient and healthy people [2]. Thus we discuss the effect of facial expressions on the visual attention differences between individuals with ASD and healthy individuals in this section. Since there are many kinds of expressions, we use positive 1, positive 2, neutral, negative 1, negative 2, and complex represent the level of these expressions. In detail, positive 1 represents generally positive facial expressions, i.e., smile. Positive 2 represents very positive facial expressions, i.e., happy, and laugh. Neutral represents neutral facial expressions. Negative 1 represents generally negative facial expressions (i.e., sad, disappointment, etc.). Negative 2 represents very negative facial expressions (i.e., fear, anger, cry, etc.). Complex represents complex facial expressions (i.e., surprise, strange, etc.).

Figure 7 illustrates linear CC scores between ASD participants' saliency map and healthy participants' saliency map under different facial expressions (emotions). For "Positive 1," "Neutral," "Negative 1," and "Negative 2," CC scores are around 0.8. Nevertheless, for "Positive 2" and "Complex," which represent very positive facial expressions and complex facial expressions, respectively, the CC scores are much smaller than others, especially for very positive facial expressions. It illustrates that people with ASD have much different visual attention than healthy people when



(a) Eye Region (b) Nose Region (c) Mouth Region

Fig. 8. Influence of facial expressions on percentage of fixations in each ROI. Red bars represent the percentages of fixation points of the ASD group. Blure bars represent the percentages of fixation points of the TD group. (a) In eye region. (b) Iin nose region. (c) In mouth region.

Table 5. *t*-Test for the Between-Group Comparison Over 300 Images for Items in Figure 8

ROIs	Eye (p1)	Eye (p2)	Eye (n)	Eye (n1)	Eye (n2)	Eye (c)	Nose (p1)	Nose (p2)	Nose (n)
H	0	1	1	0	0	1	1	1	1
P	7.67e-01	2.20e-03	5.46e-11	4.70e-01	6.54e-02	2.80e-09	1.84e-10	5.92e-08	6.03e-07

ROIs	Nose (n1)	Nose (n2)	Nose (c)	Mouth (p1)	Mouth (p2)	Mouth (n)	Mouth (n1)	Mouth (n2)	Mouth (c)
H	1	1	1	1	1	1	1	0	0
P	9.48e-17	1.04e-07	1.13e-08	1.78e-05	5.86e-10	1.00e-02	3.06e-02	4.06e-01	8.93e-01

$H = 1$ indicates that t-test rejects the null hypothesis at the 5% significance level (i.e., the difference between two groups is significant), while $H = 0$ indicates that t-test does not reject the null hypothesis at the 5% significance level. p1: positive 1; p2: positive 2;, n: neutral; n1: negative 1; n2: negative 2; c: complex; H: test decision for the null hypothesis; P: p-value for the T-statistic.

fixating on very positive facial expressions. We believe that the visual attention map of face images can be applied to distinguish ASD individuals and healthy individuals and it is more reasonable to use very positive facial expressions as the stimuli from the conclusion above.

Figure 8 illustrates the influence of facial expressions on the percentage of fixations in each ROI. As shown in Figure 8(a), we can see that under different facial expressions, percentages of fixation points of healthy people (blue bars) are around 0.4 in the eye region. However, as demonstrated with red bars in Figure 8(a), for people with ASD, in the eye region, percentages of fixations under "Positive 2," "Neutral" and "Complex" facial expressions are obviously smaller than that under "Positive 1," "Negative 1," and "Negative 2" facial expressions. In the nose region, as shown in Figure 8(b), percentages of fixations of healthy people are similar under different facial expressions; nevertheless, for people with ASD, the percentage of fixations under "Negative 1" facial expression is smaller than those under other facial expressions. In the mouth region, as shown in Figure 8(c), both people with ASD and healthy people fixate less under "Positive 1," "Neutral," and "Negative 1" facial expressions. For the other three facial expressions, which are "Positive 2," "Negative 2," and "Complex," healthy people fixate much more on the mouth region under "Positive 2" expression, while people with ASD fixate less on the mouth region under "Positive 2" facial expression. According to these phenomena, we could get special features of human face images for ASD. Table 5 shows t-test for the between-group comparison over 300 images for items in Figure 8. Statistics show that most of these differences are significant.

5 VISUAL ATTENTION PREDICTION ON HUMAN FACES FOR ASD

In Section 4, we compare the differences of visual attention on human faces between ASD participants and healthy subjects. In this section, we introduce our method for the visual attention

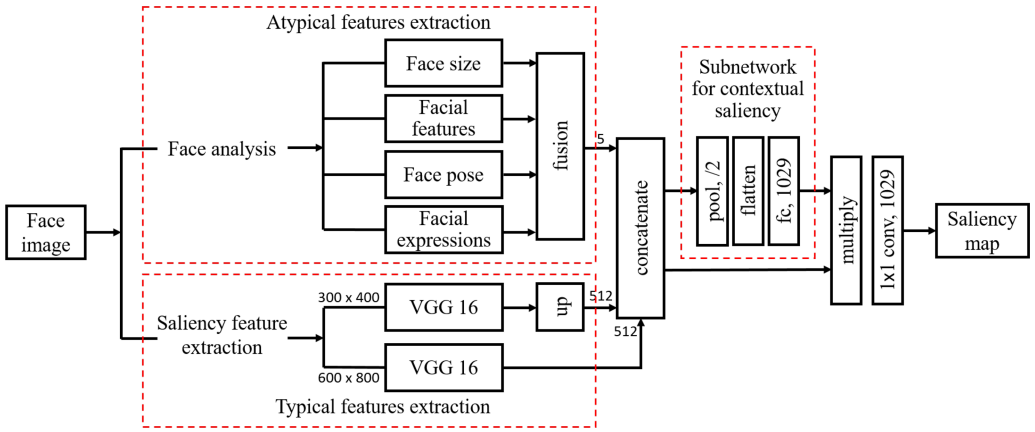


Fig. 9. Flowchart of our proposed method. "5" represents five feature maps we extracted, including face, left eye, right eye, nose, and mouth, since they are the most salient regions in natural images with faces in them.

prediction on human faces for ASD. Figure 9 illustrates the flowchart of our proposed method. In brief, we extract atypical features from face images according to the special visual attention of ASD, and then integrate the features via CASNet [25]. The subnetwork of CASNet (as shown in the dashed red rectangle on the right side of Figure 9) could capture the relative weight of semantic features of an image. In the following of this section, we first propose some atypical facial features based on the atypical visual attention analysis in Section 4. Next, we introduce our visual attention prediction methods on human faces for ASD. Then, we discuss our experimental validation process and the performance of our proposed method. Finally, we discuss the limitations of our experiments and propose possible future research directions.

5.1 Atypical Feature Extraction

As described in Section 4, the visual attention of people with ASD is much different compared with that of healthy people. Thus, in this section, we propose our method of extracting atypical features from 300 facial images. We first extract face size features from facial images by simply placing a Gaussian kernel at the center of the face. The standard deviation σ of the Gaussian kernel can be calculated by $\sigma = \sqrt{S}/4$, where S denotes the face size. Then we extract atypical facial features for the left eye region, right eye region, nose region, and mouth region, respectively, according to the preceding analysis. These five features together constitute the atypical features extracted for ASD.

In previous work, Min et al. [45] proposed a method to extract salient facial features. In this article, we extract different facial feature points because of atypical visual attention of people with ASD. Figure 10 illustrates the example of facial feature points extracted from 300 facial images. Note that for the nose region, we extract different feature points from facial images when the face pose is turned to the left or right. Then we calculate the feature map by placing a uniform Gaussian kernel at each extracted feature point and the feature map of region k can be expressed as

$$S_k(x) = \mathcal{N}\left(\sum_p \exp\left[-\frac{(x - x_p)^2}{2\sigma^2}\right]\right), p \in \mathbb{P}_k, \tag{2}$$

$$\sigma = w_p \cdot w_e \cdot 40, \tag{3}$$

where k represents the ROI (left eye, right eye, nose, or mouth region). p denotes each feature point in region k and $\mathbb{P}_k$ denotes the feature point sets in region k. x can be any two-dimensional

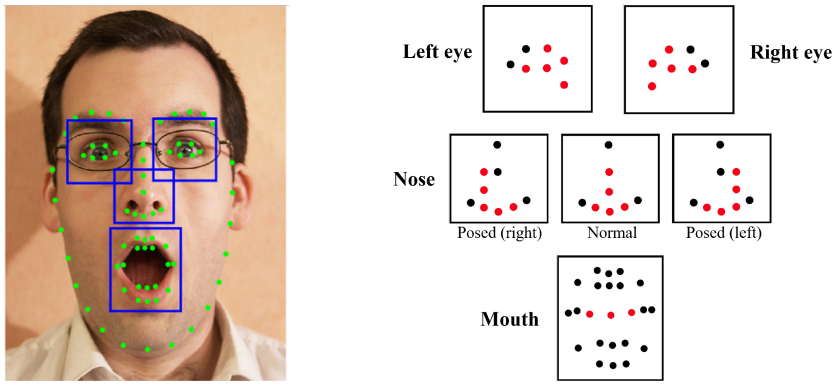


Fig. 10. Example of facial features extracted from stimuli. Chosen facial marks (marked as red points) for all ROIs are shown. Note that when the face pose changed, we use different features for the nose region. Posed toward "left" or "right" is defined based on the viewer's coordinate system.

coordinate in the image and x_p is the coordinate of feature point p. σ denotes the standard deviation of the Gaussian kernel. $\mathcal{N}$ is used to normalize the extracted feature map to the same dynamic range (i.e., [0, 1]). w_p and w_e denote the weight coefficients of the influence of face pose or facial expressions, respectively.

w_p represents face pose weight. As described in Section 4.3, we find that the influence of face poses of faces in facial images on the visual attention of people with autism is less than that of healthy people. Thus, we consider this atypical feature in this article. Similar to [45], according to the yaw angle (Y), we classify the images into three types. However, considering atypical visual attention of ASD, we define a new face pose range. Compared to the parameters used in [45], we enlarge the face pose response from 10° to 15°. When $Y > 15°$, the face pose is considered to be toward the left. When $Y < -15°$, the face pose is considered to be toward the right. When $-15° < Y < 15°$, the face pose is considered to be toward the forward (normal). As shown in Equation (3), in general condition, when the face is posed toward forward, we set $w_p = 1$. When faces turn to the left or right, for the left eye region or the right eye region, we use a lower weight for face pose as shown in Table 6(a). Moreover, we utilize new feature point sets "Posed (right)" or "Posed (left)," as shown in Figure 10, for nose region, when faces turn to the the right or left. Table 6(a) summarizes the weights for the effect of face pose in different conditions.

w_e represents facial expressions weight. As described in Section 4.4, compared to healthy people, the visual attention of individuals with ASD will be more influenced by facial expressions in images. Thus, we introduce a weight coefficient w_e corresponding to facial expressions in Equation (3). For each ROI, the value of weight w_e comes from the comparison between the percentage of fixation points of individuals with ASD falling into the ROI. Table 6(b) summarizes the weights for the effect of face pose in different conditions.

5.2 Atypical Visual Attention Prediction on Human Faces for ASD

To verify the effectiveness of our proposed method, we compare the performance of six state-of-the-art deep learning based saliency prediction models, including SALICON [29], mlnet [15], SAM-VGG [16], SAM-ResNet [16], SalGAN [49], and CASNet [25], and their corresponding fine-tuned models based on our VAFA database. Our goal in this step is to choose one visual attention model which could get better performance across various evaluation metrics after fine-tuning. Table 7 lists the performance of these models and their corresponding fine-tuned models on our

Table 6. Parameters Setting for Atypical Feature Extraction

(a) Face pose weight w_p

Conditions	Pose	w_p			
		Left eye	Right eye	Nose	Mouth
Normal	$-15° < Y < 15°$	1	1	1	1
Posed (left)	$Y > 15°$	1	1.2	1	1
Posed (right)	$Y < -15°$	1.2	1	1	1

(b) Facial expressions weight w_e

Expressions	w_e		
	Eye	Nose	Mouth
Positive 1	1	1	0.7
Positive 2	0.75	1	1
Neutral	0.75	1	0.7
Negative 1	1	0.7	0.7
Negative 2	1	1	1
Complex	0.75	1	1

Table 7. Results on Testing Set of the VAFA Database

Models	AUC		sAUC		CC		NSS	
	Original	Fine-tuned	Original	Fine-tuned	Original	Fine-tuned	Original	Fine-tuned
SALICON [29]	0.7856	0.8087	0.5419	0.5552	0.5628	0.6448	1.3816	1.4237
mlnet [15]	0.8175	0.8186	0.5509	0.5598	0.6768	0.6955	1.5957	1.6011
SAM-VGG [16]	0.8297	**0.8369**	0.5529	0.5644	0.7171	**0.7710**	1.6900	1.7594
SAM-ResNet [16]	0.8288	0.8155	0.5595	0.5585	0.7537	0.6873	**1.7764**	1.5838
SlaGAN [49]	0.8182	0.8256	0.5824	0.5752	0.6926	0.7422	1.5654	1.6811
CASNet [25]	0.8272	**0.8376**	**0.5825**	**0.5832**	0.7283	**0.7791**	1.6418	**1.7812**

"Original" represents the original model designed for healthy people. "Fine-tuned" represents the model fine-tuned based on the SPCA database and VAFA database. AUC, sAUC, CC, and NSS are used to evaluate the performance of these models. We highlight the best two results under each evaluation criterion in bold.

testing set of the VAFA database. After fine-tuning, almost all models have performance improvements except SAM-ResNet. We believe that the exception to SAM-ResNet is caused by its complex network structure. Obviously, the CASNet looks better across all criteria. With this comparison, we believe that the CASNet after fine-tuning could better match the visual attention patterns of individuals with ASD. Moreover, CASNet creatively proposed a subnetwork for contextual saliency perception, especially for emotion perception. Thus, we incorporate atypical features we extracted into CASNet and enlarge the data dimension of layers after the concatenate layer.

Figure 9 illustrates the flowchart of our methods to predict the atypical visual attention of ASD. We first extract atypical features for ASD as described in Section 5.1. In this step, we extract five-dimensional atypical facial features for individuals with ASD. Moreover, as shown in Figure 7, the visual attention of individuals with ASD and healthy people have a strong correlation. Thus, we integrate the extracted atypical features via CASNet [25], which is a state-of-the-art visual attention model of healthy people. The typical features are extracted based on a two-stream VGG net with different input size. The outputs of the two VGG networks are resized to the same spatial

dimension. We get 5-dimensional vector, 512-dimensional vector, and 512-dimensional vector from our atypical features extraction procedure and two VGG 16 net. The size of each feature map is 18×25. Then we concatenate atypical features and typical features, and then feed into the subnetwork designed for the feature weights' perception as described in [25]. After the concatenating layer, the dimension of the feature vector is 1,029. Specifically, we expand the dimension of the fully connected (fc) layer in the subnetwork to fit our features' dimension. The dimension of the output of the subnetwork is 1,029. Next, a multiple layer is used to integrate the weights captured from the subnetwork into the concatenated features. Finally, passing a convolution layer (also expanded to fit the dimension) with a 1×1 kernel, we get the final saliency map.

5.3 Experimental Validation

5.3.1 Experimental Settings. For fine-tuning six state-of-the-art deep learning based saliency prediction models, we first pre-train these models on a large saliency dataset—SALICON [31]. Next, we fine-tune these models on the saliency prediction for children with the ASD (SPCA) dataset [19]. Finally, we fine-tune these models on the training set of our VAFA database. We randomly select 240 images as the training set, and another 60 images as the testing set. For SALICON, mlnet, SAM-VGG, SAM-ResNet, and SalGAN, we set the training parameters to be the same as that in [19] when fine-tuning these networks. For CASNet, which is not mentioned in [19], SGD is applied with a learning rate of 10^{-5}, momentum of 0.9, and weight decay of 0.0005 when fine-tuning CASNet.

In order to get the prediction model for the visual attention for ASD, our proposed model should be trained first. As described in Section 5.2, we expand the dimension for the fully connected layer and the last convolutional layer to fit the dimension of our proposed method. We first initiate the added parameters of the two layers with zero weights and keep the other parameters of the original model of CASNet unchanged. Next, we fine-tune our proposed methods on a saliency prediction for children with the ASD (SPCA) dataset [19] using a relatively large learning rate of 10^{-3}. Then we use a learning rate of 10^{-5} to fine-tune our model based on our VAFA database. During the fine-tuning procedure, we freeze the network before the "concatenate" layer and only fine-tune the following network.

The feasibility of using pre-trained networks (VGG) and fine-tuning with small datasets is as follows. First, as shown in Figure 7, CC comparisons illustrate that individuals with ASD and healthy controls have similar visual attention patterns on human faces, though differences exist. And carefully comparing the visual attention maps of ASD patients and healthy controls, we find that they are generally similar, but for relatively more salient regions, they are different. Furthermore, as far as we know, there is no similar large open datasets related to the visual attention of individuals with ASD, which could be used to train an end-to-end DNN. Therefore, we use pre-trained networks in this article. Moreover, in our previous work [19], we have fine-tuned the neural networks on a small dataset and obtained relatively good performance. Thus, the fine-tuning method is feasible.

5.3.2 Evaluation Criteria. We use six criteria, which are AUC-Judd [33], AUC-Borji [9], sAUC [70], CC [38], NSS [50], and SIM [57], to evaluate the performance of the visual attention models. The Area Under receiver operating characteristic (ROC) Curve (AUC) criteria treats the saliency map as a binary classifier of human fixations with various thresholds. Three kinds of AUC criteria including AUC-Judd, AUC-Borji, and sAUC are used in this article. In particular, shuffled-AUC (sAUC) could alleviate the effects of center bias of the fixations. Linear Correlation Coefficient (CC) computes the linear correlation coefficient between saliency map and visual attention map. Normalized Scanpath Saliency (NSS) calculates the average normalized saliency at each fixation point. Similarity Metric (SIM) is calculated by summing the minimum values at each pixel for

Table 8. Results of Our Proposed Method

Models	AUC Judd	AUC Borji	sAUC	CC	NSS	SIM
CASNet	0.8307	0.7956	0.6220	0.7432	1.6431	0.6339
CASNet fine-tuned	0.8350	0.8052	0.6166	0.7800	1.7492	0.6488
The proposed method	**0.8480**	**0.8234**	**0.6232**	**0.8272**	**1.8239**	**0.6729**

AUC-Judd, AUC-Borji, sAUC, CC, NSS, and SIM are used to evaluate the performance of these models. We highlight the best performance under each criterion in bold.

normalized input maps. CC, NS, and SIM are similarity criteria which are used to measure the similarity between the predicted saliency map and the ground-truth visual attention map. Detailed introduction of these criteria could be found in [11].

5.3.3 Results. As shown in Table 7 and discussed in Section 5.2, after fine-tuning, nearly all saliency prediction models get better performance on our VAFA database. Moreover, CASNet fine-tuned based on our database looks better across all criteria. Therefore, to illustrate that our proposed method has better performance, we further compare the results of our proposed method and the results of the fine-tuned CASNet.

We adopt a 10-fold cross-validation method to evaluate the performance of CASNet, fine-tuned CASNet, and our proposed method. In each run of cross-validation, we randomly select 240 images from the VAFA database as the training set and the other 60 images as the testing set. We fine-tune the model for 100 epochs on the training set. Then we calculate the performance of CASNet, fine-tuned CASNet, and our proposed method on the testing set. Table 8 lists the averaged performance after 10-fold cross-validation. The best performance under each criterion is highlighted in bold. As shown in the table, our proposed method has the best performance under almost all criteria. In particular, for three similarity metrics, which are CC, NSS, and SIM, our proposed method has a great improvement in performance. Figure 11 illustrates the saliency map of all models. Apparently, most models have improvements in performance after fine-tuning and our proposed method appears to perform best. For example, for the image in the fifth column, the expression of the face is anger. In the visual attention map of individuals with ASD, the mouth region is the most salient area, but no similar phenomenon shows for healthy controls. In the output saliency map of our proposed method, mouth regions are more emphasized compared to other methods. Carefully comparing the output saliency map of our proposed method and visual the attention map of individuals with ASD, we find that our method is very robust. However, possible failure or less accurate results may happen when the expressions on human faces are strange or the background of the image is complex.

5.4 Further Discussion

Our article only focuses on the still images; studies about the visual attention on videos may contribute more discoveries for the reason that videos have temporal information. With dynamic face stimuli, the visual behaviors of individuals with ASD may have differences with that on still stimuli. The visual attention of individuals with ASD may have more differences with healthy controls as the variation of continuous facial motion. Predicting the visual attention of individuals with ASD on dynamic human faces may be more significant. This is another interesting research direction.

6 CONCLUSION

In this article, we analyze and predict the visual attention on human faces for the children with ASD. We first construct a Visual Attention on Faces for Autism Spectrum Disorder (VAFA) database. Based on the database, we analyze the differences of the visual attention between people with

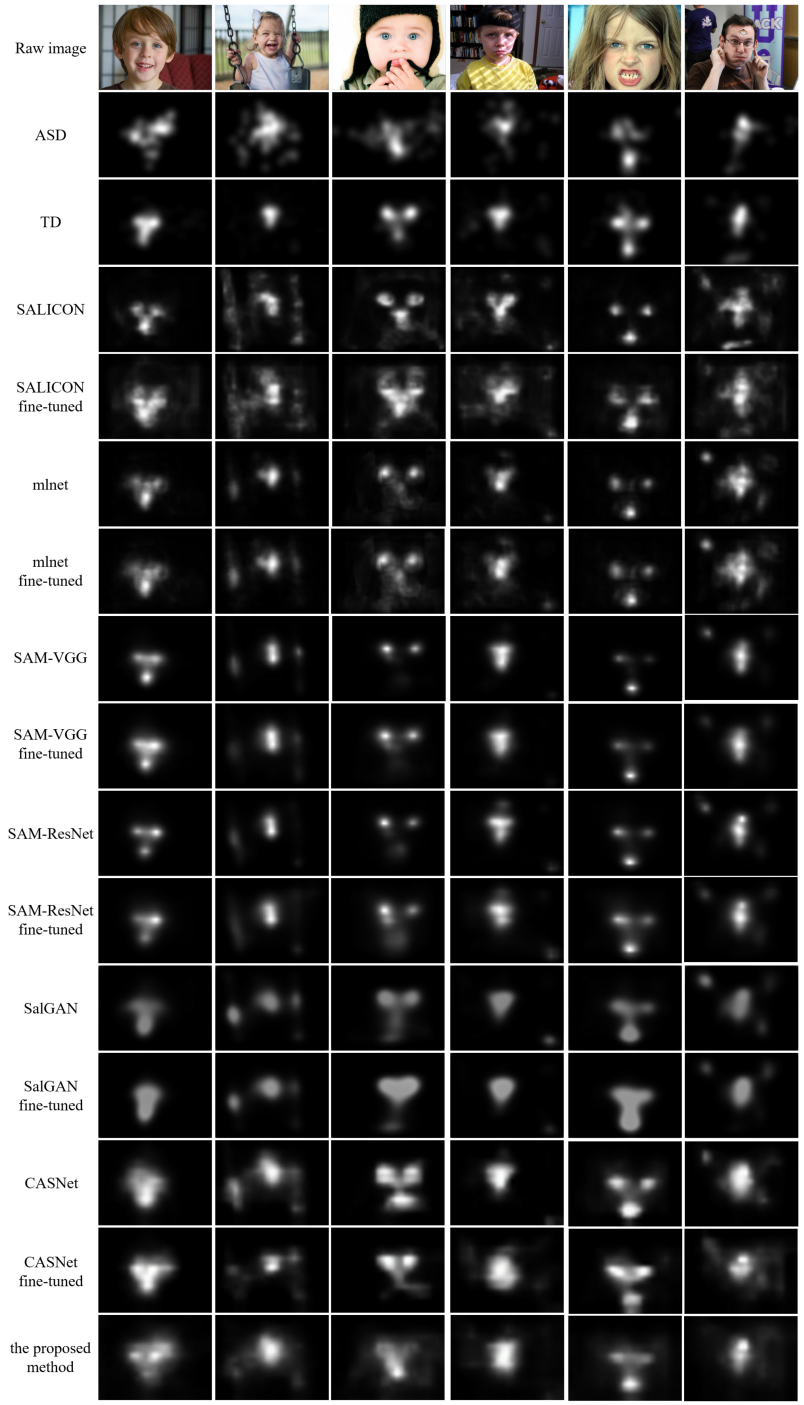


Fig. 11. Saliency map of sample images. The first row: sample images. The second row: visual attention map of people with ASD. The third row: visual attention map of TD people. The 4th row to the 15th row: saliency map calculated from the original and fine-tuned models. The 16th row: saliency map calculated from our proposed model.

ASD and healthy people. Four factors are found to make a difference, which are facial proportion in images, facial features, face pose, and facial expressions, respectively. Based on the differences, we propose to extract atypical features for ASD, and integrate these feature maps adaptively with features extracted from the CASNet. After fine-tuning, our proposed method could get the best performance. Learning to predict the visual attention of children with ASD contributes to related research in the field of medical science, psychology, and education. This research also has many application scenarios. With the help of our proposed model, we could design specialized education contents with human faces for the children with ASD. And this study could also contribute to the related research on rapid screening for ASD.

REFERENCES

[1] Dima Amso, Sara Haas, and Julie Markant. 2014. An eye tracking investigation of developmental change in bottom-up attention orienting to faces in cluttered natural scenes. *PLoS One* 9, 1 (2014), e85701.

[2] Jakob Åsberg Johnels, Daniel Hovey, Nicole Zürcher, Loyse Hippolyte, Eric Lemonnier, Christopher Gillberg, and Nouchine Hadjikhani. 2017. Autism and emotional face-viewing. *Autism Research* 10, 5 (2017), 901–910.

[3] American Psychiatric Association et al. 2013. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. American Psychiatric Pub.

[4] Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson. 2015. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (FG'15)*, Vol. 6. 1–6.

[5] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. OpenFace 2.0: Facial behavior analysis toolkit. In *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (FG'18)*. 59–66.

[6] Yair Bar-Haim, Cory Shulman, Dominique Lamy, and Arnon Reuveni. 2006. Attention to eyes and mouth in high-functioning children with autism. *Journal of Autism and Developmental Disorders* 36, 1 (2006), 131–137.

[7] Elina Birmingham, Moran Cerf, and Ralph Adolphs. 2011. Comparing social attention in autism and amygdala lesions: Effects of stimulus and task condition. *Social Neuroscience* 6, 5–6 (2011), 420–435.

[8] Ali Borji. 2012. Boosting bottom-up and top-down visual features for saliency estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'12)*. 438–445.

[9] Ali Borji, Dicky N. Sihite, and Laurent Itti. 2013. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing* 22, 1 (2013), 55–69.

[10] Claus Bundesen, Signe Vangkilde, and Anders Petersen. 2015. Recent developments in a computational theory of visual attention (TVA). *Vision Research* 116 (2015), 210–218.

[11] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. 2018. What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2018).

[12] Katarzyna Chawarska, Suzanne Macari, and Frederick Shic. 2013. Decreased spontaneous attention to social scenes in 6-month-old infants later diagnosed with autism spectrum disorders. *Biological Psychiatry* 74, 3 (2013), 195–203.

[13] Coralie Chevallier, Julia Parish-Morris, Alana McVey, Keiran M. Rump, Noah J. Sasson, John D. Herrington, and Robert T. Schultz. 2015. Measuring social attention and motivation in autism spectrum disorder using eye-tracking: Stimulus type matters. *Autism Research* 8, 5 (2015), 620–628.

[14] Ben Corden, Rebecca Chilvers, and David Skuse. 2008. Avoidance of emotionally arousing stimuli predicts social–perceptual impairment in Asperger's syndrome. *Neuropsychologia* 46, 1 (2008), 137–147.

[15] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. 2016. A deep multi-level network for saliency prediction. In *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR'16)*. 3488–3493.

[16] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. 2018. Predicting human eye fixations via an LSTM-based saliency attentive model. *IEEE Transactions on Image Processing* (2018).

[17] Geraldine Dawson, Sara Jane Webb, and James McPartland. 2005. Understanding the nature of face processing impairment in autism: Insights from behavioral and electrophysiological studies. *Developmental Neuropsychology* 27, 3 (2005), 403–424.

[18] Huiyu Duan, Guangtao Zhai, Xiongkuo Min, Zhaohui Che, Yi Fang, Xiaokang Yang, Jesús Gutiérrez, and Patrick Le Callet. 2019. A dataset of eye movements for the children with autism spectrum disorder. In *Proceedings of the ACM Multimedia Systems Conference (MMSys'19)*.

[19] Huiyu Duan, Guangtao Zhai, Xiongkuo Min, Yi Fang, Zhaohui Che, Xiaokang Yang, Cheng Zhi, Hua Yang, and Ning Liu. 2018. Learning to predict where the children with ASD look. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'18)*. 704–708.

[20] Huiyu Duan, Guangtao Zhai, Xiongkuo Min, Yucheng Zhu, Yi Fang, and Xiaokang Yang. 2018. Perceptual quality assessment of omnidirectional images. In *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS'18)*. 1–5.

[21] Hedwig Eisenbarth and Georg W. Alpers. 2011. Happy mouth and sad eyes: Scanning emotional facial expressions. *Emotion* 11, 4 (2011), 860.

[22] Terje Falck-Ytter, Sven Bölte, and Gustaf Gredebäck. 2013. Eye tracking in early autism research. *Journal of Neurodevelopmental Disorders* 5, 1 (2013), 28.

[23] Terje Falck-Ytter, Elisabeth Fernell, Christopher Gillberg, and Claes Von Hofsten. 2010. Face scanning distinguishes social from communication impairments in autism. *Developmental Science* 13, 6 (2010), 864–875.

[24] Terje Falck-Ytter and Claes von Hofsten. 2011. How special is social looking in ASD: A review. In *Progress in Brain Research*. Vol. 189. Elsevier, 209–222.

[25] Shaojing Fan, Zhiqi Shen, Ming Jiang, Bryan L. Koenig, Juan Xu, Mohan S. Kankanhalli, and Qi Zhao. 2018. Emotional attention: A study of image sentiment and visual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'18)*. 7521–7531.

[26] Bruno Gepner, Beatrice de Gelder, and Scania de Schonen. 1996. Face processing in autistics: Evidence for a generalised deficit? *Child Neuropsychology* 2, 2 (1996), 123–139.

[27] Quentin Guillon, Nouchine Hadjikhani, Sophie Baduel, and Bernadette Rogé. 2014. Visual social attention in autism spectrum disorder: Insights from eye tracking studies. *Neuroscience & Biobehavioral Reviews* 42 (2014), 279–297.

[28] Chenlei Guo and Liming Zhang. 2010. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE Transactions on Image Processing* 19, 1 (2010), 185–198.

[29] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. 2015. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'15)*. 262–270.

[30] Laurent Itti, Christof Koch, and Ernst Niebur. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 11 (1998), 1254–1259.

[31] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. 2015. Salicon: Saliency in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15)*. 1072–1080.

[32] Ming Jiang and Qi Zhao. 2017. Learning visual attention to identify people with autism spectrum disorder. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'17)*. 3267–3276.

[33] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. 2009. Learning to predict where humans look. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'09)*. 2106–2113.

[34] Chawarska Katarzyna, Volkmar Fred, and Klin Ami. 2010. Limited attentional bias for faces in toddlers with autism spectrum disorders. *Archives of General Psychiatry* 67, 2 (2010), 178–185.

[35] Ami Klin and Warren Jones. 2008. Altered face scanning and impaired recognition of biological motion in a 15-month-old infant with autism. *Developmental Science* 11, 1 (2008), 40–46.

[36] Ami Klin, Sara S. Sparrow, Annelies De Bildt, Domenic V. Cicchetti, Donald J. Cohen, and Fred R. Volkmar. 1999. A normed study of face recognition in autism and related disorders. *Journal of Autism and Developmental Disorders* 29, 6 (1999), 499–508.

[37] Srinivas S. S. Kruthiventi, Kumar Ayush, and R. Venkatesh Babu. 2017. Deepfix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing* 26, 9 (2017), 4446–4456.

[38] Olivier Le Meur, Patrick Le Callet, and Dominique Barba. 2007. Predicting visual fixations on video based on low-level visual features. *Vision Research* 47, 19 (2007), 2483–2498.

[39] Zhiqiang Li, Tao Fang, and Hong Huo. 2010. A saliency model based on wavelet transform and visual attention. *Science China: Information Sciences* 53, 4 (2010), 738–751.

[40] Ming Liang and Xiaolin Hu. 2015. Predicting eye fixations with higher-level visual features. *IEEE Transactions on Image Processing* 24, 3 (2015), 1178–1189.

[41] René Marois and Jason Ivanoff. 2005. Capacity limits of information processing in the brain. *Trends in Cognitive Sciences* 9, 6 (2005), 296–305.

[42] James C. McPartland, Sara Jane Webb, Brandon Keehn, and Geraldine Dawson. 2011. Patterns of visual attention to faces and objects in autism spectrum disorder. *Journal of Autism and Developmental Disorders* 41, 2 (2011), 148–157.

[43] Xiongkuo Min, Ke Gu, Guangtao Zhai, Jing Liu, Xiaokang Yang, and Chang Wen Chen. 2017. Blind quality assessment based on pseudo-reference image. *IEEE Transactions on Multimedia* 20, 8 (2017), 2049–2062.

[44] Xiongkuo Min, Kede Ma, Ke Gu, Guangtao Zhai, Zhou Wang, and Weisi Lin. 2017. Unified blind quality assessment of compressed natural, graphic, and screen content images. *IEEE Transactions on Image Processing* 26, 11 (2017), 5462–5474.

[45] Xiongkuo Min, Guangtao Zhai, Ke Gu, Jing Liu, Shiqi Wang, Xinfeng Zhang, and Xiaokang Yang. 2017. Visual attention analysis and prediction on human faces. *Information Sciences* 420 (2017), 417–430.

[46] Xiongkuo Min, Guangtao Zhai, Ke Gu, Yutao Liu, and Xiaokang Yang. 2018. Blind image quality estimation via distortion aggravation. *IEEE Transactions on Broadcasting* 64, 2 (2018), 508–517.

[47] Xiongkuo Min, Guangtao Zhai, Ke Gu, and Xiaokang Yang. 2017. Fixation prediction through multimodal analysis. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 13, 1 (2017), 6.

[48] Julie Osterling and Geraldine Dawson. 1994. Early recognition of children with autism: A study of first birthday home videotapes. *Journal of Autism and Developmental Disorders* 24, 3 (1994), 247–257.

[49] Junting Pan, Cristian Canton Ferrer, Kevin McGuinness, Noel E. O'Connor, Jordi Torres, Elisa Sayrol, and Xavier Giro-i Nieto. 2017. Salgan: Visual saliency prediction with generative adversarial networks. *Arxiv Preprint Arxiv:1701.01081* (2017).

[50] Robert J. Peters, Asha Iyer, Laurent Itti, and Christof Koch. 2005. Components of bottom-up gaze allocation in natural images. *Vision Research* 45, 18 (2005), 2397–2416.

[51] Katherine Rice, Jennifer M. Moriuchi, Warren Jones, and Ami Klin. 2012. Parsing heterogeneity in autism spectrum disorders: Visual scanning of dynamic social scenes in school-aged children. *Journal of the American Academy of Child & Adolescent Psychiatry* 51, 3 (2012), 238–248.

[52] Caroline E. Robertson and Simon Baron-Cohen. 2017. Sensory perception in autism. *Nature Reviews Neuroscience* 18, 11 (2017), 671.

[53] Manar D. Samad, Norou Diawara, Jonna L. Bobzien, John W. Harrington, Megan A. Witherow, and Khan M. Iftekharuddin. 2018. A feasibility study of autism behavioral markers in spontaneous facial, visual, and hand movement response data. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 26, 2 (2018), 353–361.

[54] Noah J. Sasson, Jed T. Elison, Lauren M. Turner-Brown, Gabriel S. Dichter, and James W. Bodfish. 2011. Brief report: Circumscribed attention in young children with autism. *Journal of Autism and Developmental Disorders* 41, 2 (2011), 242–247.

[55] Noah J. Sasson and Emily W. Touchstone. 2014. Visual attention to competing social and object images by preschool children with autism spectrum disorder. *Journal of Autism and Developmental Disorders* 44, 3 (2014), 584–592.

[56] David R. Simmons, Ashley E. Robertson, Lawrie S. McKay, Erin Toal, Phil McAleer, and Frank E. Pollick. 2009. Vision in autism spectrum disorders. *Vision Research* 49, 22 (2009), 2705–2739.

[57] Michael J. Swain and Dana H. Ballard. 1991. Color indexing. *International Journal of Computer Vision* 7, 1 (1991), 11–32.

[58] Luan Tran, Xi Yin, and Xiaoming Liu. 2017. Disentangled representation learning GAN for pose-invariant face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1415–1424.

[59] Andrius Vabalas and Megan Freeth. 2016. Brief report: Patterns of eye movements in face to face conversation are associated with autistic traits: Evidence from a student sample. *Journal of Autism and Developmental Disorders* 46, 1 (2016), 305–314.

[60] Eleonora Vig, Michael Dorr, and David Cox. 2014. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'14)*. 2798–2805.

[61] Lijun Wang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. 2015. Deep networks for saliency detection via local estimation and global search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15)*. 3183–3192.

[62] Linzhao Wang, Lijun Wang, Huchuan Lu, Pingping Zhang, and Xiang Ruan. 2016. Saliency detection with recurrent fully convolutional networks. In *Proceedings of the European Conference on Computer Vision (ECCV'16)*. Springer, 825–841.

[63] Shuo Wang, Ming Jiang, Xavier Morin Duchesne, Elizabeth A. Laugeson, Daniel P. Kennedy, Ralph Adolphs, and Qi Zhao. 2015. Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking. *Neuron* 88, 3 (2015), 604–616.

[64] Shuo Wang, Juan Xu, Ming Jiang, Qi Zhao, Rene Hurlemann, and Ralph Adolphs. 2014. Autism spectrum disorder, but not amygdala lesions, impairs social attention in visual search. *Neuropsychologia* 63 (2014), 259–274.

[65] Li Yi, Yuebo Fan, Paul C. Quinn, Cong Feng, Dan Huang, Jiao Li, Guoquan Mao, and Kang Lee. 2013. Abnormality in face scanning by children with autism spectrum disorder is limited to the eye region: Evidence from multi-method analyses of eye tracking data. *Journal of Vision* 13, 10 (2013), 5–5.

[66] Li Yi, Cong Feng, Paul C. Quinn, Haiyan Ding, Jiao Li, Yubing Liu, and Kang Lee. 2014. Do individuals with and without autism spectrum disorder scan faces differently? A new multi-method look at an existing controversy. *Autism Research* 7, 1 (2014), 72–83.

[67] Bangjie Yin, Luan Tran, Haoxiang Li, Xiaohui Shen, and Xiaoming Liu. 2018. Towards interpretable face recognition. *Arxiv Preprint Arxiv:1805.00611* (2018).

[68] Amir Zadeh, Yao Chong Lim, Tadas Baltrusaitis, and Louis-Philippe Morency. 2017. Convolutional experts constrained local model for 3D facial landmark detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'17) Workshops*. 2519–2528.

[69] Guangtao Zhai, Xiaolin Wu, Xiaokang Yang, Weisi Lin, and Wenjun Zhang. 2011. A psychovisual quality metric in free-energy principle. *IEEE Transactions on Image Processing* 21, 1 (2011), 41–52.
[70] Lingyun Zhang, Matthew H. Tong, Tim K. Marks, Honghao Shan, and Garrison W. Cottrell. 2008. SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision* 8, 7 (2008), 32–32.
[71] Xiuzhuang Zhou, Kai Jin, Yuanyuan Shang, and Guodong Guo. 2018. Visually interpretable representation learning for depression recognition from facial images. *IEEE Transactions on Affective Computing* (2018).
[72] Wenhan Zhu, Guangtao Zhai, Xiongkuo Min, Menghan Hu, Jing Liu, Guodong Guo, and Xiaokang Yang. 2019. Multi-channel decomposition in tandem with free-energy principle for reduced-reference image quality assessment. *IEEE Transactions on Multimedia* 21, 9 (2019), 2334–2346.

Received November 2018; revised March 2019; accepted May 2019