Oneclass4ASD: Autism Spectrum Disorder Screening via One-class Classifier

Fan Chen, Jiansong Qi, Jie liu, Qingji Guan*

School of Computer Science and Technology, Beijing Jiaotong University, Beijing, China
23722041@bjtu.edu.cn, qjguan@bjtu.edu.cn

Abstract. Eye tracking, as a non-contact technology, has shown great potential for Autism Spectrum Disorder (ASD) diagnosis by distinguishing fixation patterns between healthy and ASD individuals. Though promising, eye-tracking–based method still faces several data-related issues, due to both the difficulty of acquiring sufficient ASD samples and the substantial intra-class variability of abnormal gaze behaviors caused by diverse individual preferences. These factors jointly pose significant challenges to traditional binary classification methods that rely on separating healthy and ASD populations. Therefore, in this paper, we propose a novel perspective that treats ASD as an abnormal detection, which only requires healthy individuals' data during training. Moreover, an uncertainty-aware weighting scheme is designed to address the intra-variance among individuals, assigning images different weights based on their separability between ASD and healthy groups. In comparison to previous research that adopts the binary classification method, our approach achieves comparable results yet requires only healthy subjects' data for training, significantly alleviating the difficulty of ASD data collection and offering a more practical solution for ASD diagnosis. The source code is available at `https://github.com/chenf0613/OneClassForASD`.

Keywords: Autism Spectrum Disorder (ASD) · One-class classification · Uncertainty-aware weighting.

1 Introduction

With Autism Spectrum Disorder (ASD) exuding Social interaction difficulties and vocal communication impairments in patients, public recognition and social interest are continually increasing. Although preceding studies [17] [8] have established that early diagnosis will lead to more effective interventions, current medical evidence still relies on subjective and time consuming diagnostic tools [2] [3]. As gaze patterns are capable of reflecting repetitive stereotyped behaviors typical in young ASD patients [19], eye-tracking holds the potential

* Corresponding author.

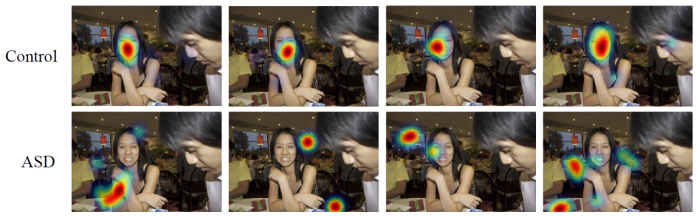

Fig. 1. Differences in the fixation pattern of healthy (denoted as control) and ASD individuals. It is obvious that healthy individuals' eye movements regarding the same picture are similar while ASD sufferers' differ significantly.

to become a leverage in objective ASD diagnosis with early research using traditional machine learning algorithm [5] generated superior results [12] [20], while the advancement of deep learning technologies with their strong representation capabilities have also led to significant breakthroughs [11] [4].

Though promising, previous studies have the disadvantage of treating healthy and ASD individuals as opposing groups, embodying binary classification models which require both healthy and ASD data, presenting two significant challenges. To begin with, the degree of social engagement in ASD patients are often low, increasing the difficulty of collecting ASD gaze patterns while also challenging its reliability. Secondly, ASD manifests differently across individuals due to personal preferences, social interactions, and other external factors, resulting in significant variations in outward appearances with patients exhibiting a wide range of fixation diagrams for a single image. As indicated in Fig. 1, healthy individuals produce nearly identical gaze patterns while ASD individuals display high variability results. In essence, the difficulties in data collection and the significant eye movement differences in children with ASD pose substantial challenges for the current binary classification methods.

To address the above issues, we treat ASD identification as abnormal detection which classifies participants as healthy or non-healthy, utilizing one-class classification methods [15] [18] [1]. By adopting this novel perspective, we only require healthy individuals' gaze pattern during training, while fixation sequences from ASD sufferers are abnormalities that appear only during testing which completes ASD diagnosis with one-class classification scheme, circumventing the variability and unreliability of ASD gaze pattern.

Furthermore, healthy individuals also exhibit a small degree of inconsistency in their gaze patterns, but can still affect the diagnostic performances, necessitating the identification of high-performing images. Although Fisher scores are frequently used for images ranking in prior works [11] [4], the selected images are still treated equally during model training. To better account for these variations and adjust the contribution of each image, we propose an uncertainty-aware weighting strategy, using the similarity between healthy individuals' fixation map for a certain image as the weight during feature fusion, where higher similarity indicates lower uncertainty and greater discriminative power. Through

uncertainty-aware weighting, we are able to mitigate the intra-group variability among healthy individuals while modulating the contributions of different images to the training process. This strategy also enables image selection using the uncertainty score of each image, with a subset of appropriate images implemented for training. Additionally, experiment results demonstrate that our method excels in both prediction performance and model efficiency, highlighting the reliability and potential of our work.

In summary, this paper has three main contributions:

- By formulating ASD identification task as an anomaly detection problem, we offer a novel perspective and creatively develop an one-class classification method, relying solely on healthy individuals' gaze pattern data, This overcomes the challenges posed by insufficient and inconsistent ASD fixations.
- We propose an uncertainty-aware strategy that leverages the similarity of fixation patterns among different healthy individuals, helping identify high discriminating stimuli and adjust their contributions dynamically.
- Our experiment is conducted on the public dataset Saliency4ASD. Compared with previous research that employs binary classification methods, our proposed method achieves superior results using only healthy individuals' data, offering a more feasible implementation for ASD diagnosis.

2 Method

In this section, we provide a clear description of our proposed model for ASD screening, namely Oneclass4ASD. As depicted in Fig. 2, we adopt a two-stage procedure, involving feature learning and one-class classification. In the first stage, we target to train a model capable of predicting healthy individuals' fixation maps, which can capture distinctive features to represent the typical eye patterns of healthy individuals while further promoting interpretability. The second stage builds on the first stage where we derive fixation features then adopt the one-class classification method to determine whether a gaze feature belongs to the healthy individual group.

2.1 Feature Learning

To extract discriminative features, we develop a model which can predict the gaze patterns of healthy individuals for any image input. The corresponding fixation map reflects the principal sections that healthy individuals are more likely to focus on, which can be further interpreted as the probability distribution of the fixation points of the viewer.

In order to capture multi-scale features, two augments of the input image, high- and the low-resolution image, are used in the feature learning stage, with the high-resolution image doubling in size [11]. Both augments are separately processed using the VGG-16 model [16]. Their outputs are upsampled to a uniform size due to their resolution difference, then concatenated to form a single

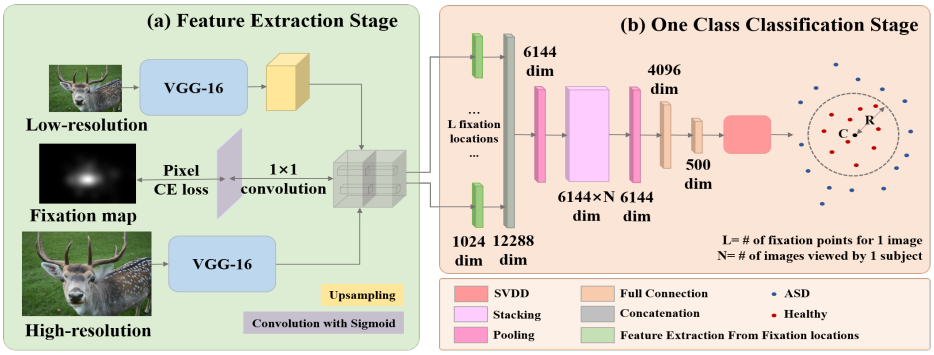

Fig. 2. An overview of our Oneclass4ASD method consisting of two stages, feature learning and one-class classification. In the green box, (a) depicts the feature learning stage, which achieves healthy individual fixation map prediction and feature extraction for input images using VGG-16 model. For the second stage in orange, (b) represents how one-class classification is conducted based on the features obtained previously. The feature vectors assist in developing a hyper-sphere using uncertainty weighting and feature fusion to singularly encompass healthy individuals.

feature map and processed using a 1×1 convolution to produce the final predicted single-channel fixation map.

As mentioned before, the output of the model is a probability map, and we conduct a binary classification on each individual pixel to determine whether it is a fixation or not. Therefore, we adopt the pixel-wise cross-entropy loss to optimize the model:

$$L(M) = -\frac{1}{HW}\sum_{i=1}^{W}\sum_{j=1}^{H}\left(M_{i,j}\log(\hat{M}_{i,j}) + (1-M_{i,j})\log(1-\hat{M}_{i,j})\right) \tag{1}$$

where M and $\hat{M}$ represents the ground-truth fixation map and the predicted fixation map. H and W denotes the height and width of the fixation map in which we use two dimensional coordinates to indicate individual pixels. Therefore, $M_{i,j}$ and $\hat{M}_{i,j}$ represents the values of the (i, j) pixel of M and $\hat{M}$ respectively.

2.2 One-class Classification

After obtaining the predicted and concatenated fixation sequence, we extract the gaze features that represents an individual's gaze pattern. For this procedure, we specifically select L fixation locations for each image ($L = 12$ in our experiments) and if the target fixation sequence contains fewer then L, zeros will be appended. Then, a 1024-d visual feature vector is extracted from the l-th ($l = 1, 2, \ldots, L$) fixation location and all L fixation feature vector are concatenate together to forms a 12288-d dimensional feature vector representing the gaze pattern for a specific image. We half its dimension into 6144-d by using average pooling (size $= 2$, stride $= 2$) to reduce the dimensionality and avoid overfitting. As we aim

to distinguish healthy individuals, we stack all the N trials (*i.e.*, the number of images viewed by the given subject) to form a weighted $6144 \times N$ dimension vector which serves as the final gaze feature of a specific subject.

Due to the difficulty of collecting eye movement data from ASD sufferers, obtaining ASD gaze features proves to be a significant challenge. To get around this barrier, our method treats ASD as an abnormality detection task that only uses healthy individuals' data. To this end, we employ a Deep Support Vector Data Descriptor (SVDD) method [14], which is capable of finding a minimum hyper-sphere that encompasses the gaze pattern features of healthy individuals. For ASD individuals, their features will fall outside the sphere, thus successfully distinguishing between the two groups. Technically, we adopt two fully connected layers to learn the distinctive features while reducing the risk of overfitting.

For the hyper-sphere developed by SVDD, there is the possibility of obtaining zero-radius or undifferentiated results, which are trivial solutions. To prevent this, we follow [14], which sets the center C as the mean of the output of the initialized fully connected network for all features. There are no bias terms but an unbounded activation Leaky ReLU function [13] is adopted. For the optimized loss function of SVDD, when the output falls outside the hyper-sphere, we incorporate the minimum distance to the boundary while zero is used when the output falls into the sphere. The process is formulated by:

$$L(\hat{O}) = R^2 + \frac{1}{\nu N} \sum_{i=1}^{N} \max \left\{ 0, \left\| \hat{O}_i - C \right\|^2 - R^2 \right\}, \tag{2}$$

where $\hat{O}$ describes the output of the fully connected network and $\hat{O}_i$ is the $i\text{-}th$ output. The number of images viewed by a healthy subject is denoted by N, and we set it to $N = 300$ during our experiment. C and R indicate the center and radius of the hyper-sphere respectively, and R^2 is used to prevent over fitting. $v \in (0, 1]$ is the proportion of features that is allowed to fall out of the hyper-sphere to control the soft boundary as we wish to discard special features. Additionally, R will be updated after every epoch of training.

2.3 Uncertainty-aware Image Selection and Fusion

As previously mentioned, even healthy individuals' gaze attention toward the same image can exhibit differences, and different images have varying degrees of these differences. Therefore, it is reasonable to conclude that images resulting in more uniformly distributed fixation maps and less variation for healthy subjects could be more distinctive. Given the imbalances in discrimination abilities, we propose that these images should be treated unequally and weighed accordingly, which corresponds to our uncertainty-aware scheme adopted in the training process. To provide a clearer understanding of this approach, we divide it into three sections, uncertainty estimation, feature fusion on image set, and image selection, which will be separately discussed.

Uncertainty estimation. In our approach, we specifically define the images that yield more similar results for healthy individuals as less uncertain and more

distinctive. To quantitatively measure the uncertainty, the Euclidean similarity is employed to calculate the similarity and weight of specific images, and the similarity can be obtained through:

$$Sim(F_i, F_j) = \frac{1}{1 + \|F_i - F_j\|}, \tag{3}$$

where F_i and F_j represent the two target features being measured. $\|\cdot\|$ denotes the Euclidean distance, which ensures $Sim(F_i, F_j) \in (0, 1]$.

We now adopt this measurement to the fixation maps of healthy individuals corresponding to a specific image. Specifically, assuming that K healthy individuals are involved for viewing each image, we can generate the average fixation maps of the healthy group denoted by $\bar{M}$. For the fixation map of the *k-th* subject, denoted by M_k, we then compute its similarity with the average map $\bar{M}$. Adding each similarity together, we obtain the total similarity as:

$$S = \frac{1}{K} \sum_{k=1}^{K} Sim\left(M_k, \bar{M}\right), \tag{4}$$

where $\bar{M}$ denotes the average fixation map of the healthy people for an image.

Feature fusion on image set. Building upon total similarity S, the weight of the specific *i-th* image in the dataset can be determined. By utilizing the softmax function, we can obtain $W_i = \text{Softmax}(S_i)$. To assign the weight of the images into the pooling process, we fuse the obtained $6144 \times N$-d vectors denoted by $V = \sum_{i=1}^{N} W_i V_i$ and it is easy to see that V_i corresponds to the 6144-d feature of the *i-th* image.

Image selection. After obtaining the weight for each specific image, a subset of more discriminating images, which is the ones with higher similarity scores and weights, can be selected to form the stimuli for eye-tracking experiments. Examples of images with different similarities are shown in Fig. 3.

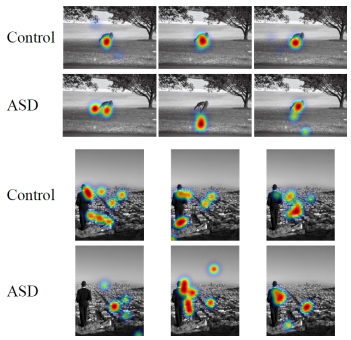

Fig. 3. The gaze patterns for different pictures. The top two rows present image with high similarity score while the lower two rows shows image with low similarity.

3 Experiments

3.1 Experimental Setup

Dataset: We utilize the Saliency4ASD dataset [7] for evaluation, which contains 300 images and 14 ASD as well as 14 healthy subjects. Each subject is asked to view all 300 images, giving 58,800 fixation maps. For these fixation data, the

fixation point index, coordinates of the fixation point, and duration of fixation are all provided. The IDs of the subject are not given, so we assume that the *i-th* fixation map is the result of the *i-th* participant with ($i = 1, 2, \cdots, 28$).

Training details: For all data, we resize the images to 800×800 and perform normalization before training. During feature extraction, VGG-16 and 1×1 convolution initialization adopt the pre-trained parameters on ImageNet [6] and Xavier initialization [9]. The Adam optimizer is utilized with learning rate initialized to 10^{-3} and weight decay set to 10^{-6}, while adopting gradient clip. The model is trained for 100 epochs with learning rate halved every 25 epochs. For one-class classification stage, we similarly adopt the Adam optimizer but with a lower initial learning rate of 10^{-7}. This process involves 1000 epochs and uses different selection of images for training. Learning rate is halved every 300 epochs and we incorporate $v = 10^{-4}$ as the threshold for the allowed portion of healthy individual features that falls outside the hyper-sphere.

Evaluation: We adopt the leave-one-subject-out protocol with each experiment consists of 14 rounds. During each round, 13 healthy subjects data is used for training while one healthy individual is combined with 14 ASD subjects as testing data. Four metrics, accuracy, sensitivity, specificity, and Area Under the ROC Curve (AUC) are used for model evaluation. For all 14 rounds of training and testing, we record and analyze the final aggregated results of these metrics.

3.2 Results

Using previous binary-classification methods as the baseline [4], we conducted various experiments and compared our Oneclass4ASD model results. By utilizing the similarity score S, we rank the 300 images from highest to lowest similarity to select sub-datasets and conduct corresponding experiments: **Oneclass4ASD-ALL**, **-T100**, **-T200**, **-M100**, and **-B100**, which uses all, top 100, top 200, middle 100, and bottom 100 ranked images in training.

Table 1. The comparative results of the five Oneclass4ASD models with different image selection and the baseline model.

Model	Data	Acc. (%)	Sen. (%)	Spe. (%)	AUC
Chen et al. (IV) [4]	TD+ASD	89.00	86.00	**93.00**	0.92
Oneclass4ASD-ALL	TD	90.48	91.32	78.57	0.93
Oneclass4ASD-T200	TD	89.52	90.31	78.57	0.89
Oneclass4ASD-T100	TD	**91.43**	**91.33**	92.86	**0.95**
Oneclass4ASD-M100	TD	93.81	94.90	78.57	0.93
Oneclass4ASD-B100	TD	86.19	86.22	85.71	0.91

The results of the five experiments in comparison to the baseline are summarized in Table 1, showing that our Oneclass4ASD-T100 model demonstrates the best overall performance. When comparing it with the baseline model, although the baseline slightly outperforms Oneclass4ASD-T100 in terms of specificity, our

Table 2. The results of Oneclass4ASD-ALL with and without uncertainty weighting.

Method	Acc. (%)	Sen. (%)	Spec. (%)	AUC
w/ uncertainty	90.48	91.32	78.57	0.93
w/o uncertainty	86.19	86.73	78.57	0.91

Table 3. The result of Oneclass4ASD-T100 using different backbone.

Backbone	Acc. (%)	Sen. (%)	Spec. (%)	AUC
VGG-16	91.43	91.33	92.86	0.95
ResNet-34	62.85	61.73	78.57	0.72

model achieves the best results regarding all other indicators, with accuracy increasing 2.4%, sensitivity increasing 5.3%, and AUC increasing 3%. It is worth noting that when comparing Oneclass4ASD-ALL and Oneclass4ASD-T100, with high similarity images alleviating the intra-group variance among the healthy group, specificity will experience a significant increase of 14.3%.

3.3 Ablation Studies

Effectiveness of Uncertainty estimation. We construct a comparative experiment without the uncertainty weighting strategy to demonstrate its usefulness and contributions. As shown in Table 2, a decline in accuracy and sensitivity is evident when uncertainty weighting is removed.

Backbone. In the first stage of our Oneclass4ASD a VGG-16 model is used. As VGG-16 and Resnet-34 [10] both serve as feature extractors and their outputs have the same channels, we compare their effects to show the validity of our choice. Table 3 presents the results of Oneclass4ASD-T100 when incorporating both models, with VGG-16 demonstrating a much better performance.

The choice of similarity measurement. In our approach, we use the Euclidean distance on 800×800 fixation maps for similarity calculation. To verify its validity, we examine the effects of different similarity metrics using Oneclass4ASD-T100. Specifically, we use the cosine metric for similarity measurement in comparison with the Euclidean distance, and explore the impact of using 50×50 fixation maps. Besides, we also evaluate the effect of using the extracted gaze features for calculation, *i.e.,* the 6144-d vector before being stacked. The results are shown in Table 4, and it is evident that using Euclidean distance and 800×800 fixation map for similarity measurement yields the best results.

Table 4. The results of Oneclass4ASD-T100 under different similarity strategies.

Similarity	Features Used	Acc.(%)	Sen.(%)	Spe.(%)	AUC
Cosine	800×800 Fixation	83.81	83.67	85.71	0.90
	50×50 Fixation	88.10	88.78	78.57	0.90
	Gaze Features	94.76	97.96	50.00	0.87
Euclidean	800×800 Fixation	**91.43**	**91.33**	**92.86**	**0.95**
	50×50 Fixation	89.05	89.29	85.71	0.94
	Gaze Features	94.29	95.92	71.43	0.93

Table 5. Train and test time.

Model	Time (h:m)
ALL	08:07
T200	04:58
T100	**02:12**
M100	02:49
B100	02:06

Time consumption. Training time is an essential indicator for diagnostic model implementation. As shown in Table 5, we are able to see that -T100 require minimum training time, further stipulating it application potential.

4 Conclusion and Limitation

In this paper, we present a novel Oneclass4ASD only using healthy individual data, mitigating the intra-variability among ASD individuals and avoid ASD data collection. We also design an uncertainty-aware weighting strategy to measure the distinctive ability of different stimuli for training. Comprehensive experiments demonstrate the supreme performance of our model. Although our method achieves promising performance on Saliency4ASD dataset, further evaluation on larger datasets is needed to explore. Besides, our method primarily aims to provide autism clinicians with a screening tool focusing on identifying neurotypical individuals in outpatient settings, and can reasonably distinguish whether a sample belongs to the neurotypical class. Further analysis of other disorders would require additional approaches, such as cross-disorder classification or out-of-distribution detection.

Acknowledgements. This work is supported by Beijing Natural Science Foundation (L259055) and National Natural Science Foundation of China (62271042, 62376021).

Disclosure of Interests. The authors have no competing interests in the paper.

References

1. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: International conference on information processing in medical imaging. pp. 146–157. Springer (2017)
2. Bodfish, J.W., Symons, F.J., Parker, D.E., Lewis, M.H.: Repetitive behavior scale–revised. Journal of Autism and Developmental Disorders (1999)
3. Bruni, T.P.: Test review: Social responsiveness scale–second edition (srs-2). Journal of Psychoeducational Assessment **32**(4), 365–369 (2014)
4. Chen, S., Zhao, Q.: Attention-based autism spectrum disorder screening with privileged modality. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1181–1190 (2019)
5. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning **20**(3), 273–297 (1995)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
7. Duan, H., Zhai, G., Min, X., Che, Z., Fang, Y., Yang, X., Gutiérrez, J., Callet, P.L.: A dataset of eye movements for the children with autism spectrum disorder. In: Proceedings of the 10th ACM Multimedia Systems Conference. pp. 255–260 (2019)

8. Fein, D., Barton, M., Eigsti, I.M., Kelley, E., Naigles, L., Schultz, R.T., Stevens, M., Helt, M., Orinstein, A., Rosenthal, M., et al.: Optimal outcome in individuals with a history of autism. Journal of child psychology and psychiatry **54**(2), 195–205 (2013)
9. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. pp. 249–256. JMLR Workshop and Conference Proceedings (2010)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
11. Jiang, M., Zhao, Q.: Learning visual attention to identify people with autism spectrum disorder. In: Proceedings of the ieee international conference on computer vision. pp. 3267–3276 (2017)
12. Liu, W., Li, M., Yi, L.: Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework. Autism Research **9**(8), 888–898 (2016)
13. Maas, A.L., Hannun, A.Y., Ng, A.Y., et al.: Rectifier nonlinearities improve neural network acoustic models. In: Proc. icml. vol. 30, p. 3. Citeseer (2013)
14. Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S.A., Binder, A., Müller, E., Kloft, M.: Deep one-class classification. In: International conference on machine learning. pp. 4393–4402. PMLR (2018)
15. Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. Neural computation **13**(7), 1443–1471 (2001)
16. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
17. Smith, T.: Outcome of early intervention for children with autism. Clinical Psychology: Science and Practice **6**(1), 33 (1999)
18. Tax, D.M., Duin, R.P.: Support vector data description. Machine learning **54**(1), 45–66 (2004)
19. Wang, Wall, B.B.M.C.S.: Promoting social attention in 3-year-olds with asd through gaze-contingent eye tracking. Autism research (2020)
20. Wang, S., Jiang, M., Duchesne, X.M., Laugeson, E.A., Kennedy, D.P., Adolphs, R., Zhao, Q.: Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking. Neuron **88**(3), 604–616 (2015)