Interdisciplinary Sciences: Computational Life Sciences (2022) 14:639–651
https://doi.org/10.1007/s12539-022-00510-6

ORIGINAL RESEARCH ARTICLE

Identifying Visual Attention Features Accurately Discerning Between Autism and Typically Developing: a Deep Learning Framework

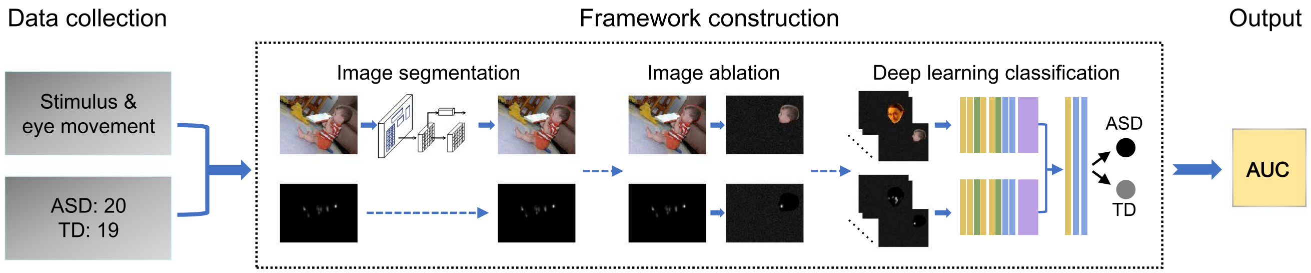
Jin Xie[1,2] · Longfei Wang[1] · Paula Webster[5] · Yang Yao[1] · Jiayao Sun[1,2] · Shuo Wang[4] · Huihui Zhou[1,3]

Received: 16 October 2021 / Revised: 10 March 2022 / Accepted: 15 March 2022 / Published online: 12 April 2022

Abstract

Atypical visual attention is a hallmark of autism spectrum disorder (ASD). Identifying the attention features accurately discerning between people with ASD and typically developing (TD) at the individual level remains a challenge. In this study, we developed a new systematic framework combining high accuracy deep learning classification, deep learning segmentation, image ablation and a direct measurement of classification ability to identify the discriminative features for autism identification. Our two-stream model achieved the state-of-the-art performance with a classification accuracy of 0.95. Using this framework, two new categories of features, Food & drink and Outdoor-objects, were identified as discriminative attention features, in addition to the previously reported features including Center-object and Human-faces, etc. Altered attention to the new categories helps to understand related atypical behaviors in ASD. Importantly, the area under curve (AUC) based on the combined top-9 features identified in this study was 0.92, allowing an accurate classification at the individual level. We also obtained a small but informative dataset of 12 images with an AUC of 0.86, suggesting a potentially efficient approach for the clinical diagnosis of ASD. Together, our deep learning framework based on VGG-16 provides a novel and powerful tool to recognize and understand abnormal visual attention in ASD, which will, in turn, facilitate the identification of biomarkers for ASD.

Graphical abstract



Keywords Autism spectrum disorder · Visual attention · Eye movement · Deep learning

1 Introduction

People with autism spectrum disorder (ASD) exhibit altered attention to specific features of visual information. Reduced attention to socially relevant stimuli [1–6], increased image center bias [7–9], and impaired joint attention [10–15] have been reported in people with ASD. For example, it has been shown that people with ASD demonstrate a stronger attention bias towards the center of images regardless of the object distribution in the images but they demonstrate

✉ Shuo Wang
shuowang@wustl.edu

✉ Huihui Zhou
zhouhh@pcl.ac.cn

Extended author information available on the last page of the article

reduced attention to faces in the stimuli [7]. These atypical attention behaviors were revealed by statistical comparisons of averaged gaze patterns on images between people with ASD and typically developing (TD) groups. However, because of the large variability of eye movement within ASD and TD groups and overlaps in the behaviors across the groups [7], there is still an inconsistency about the visual features of images associated with altered attention in ASD [16], and it is difficult to accurately separate ASD from TD at the individual level based on the averaged difference in gaze patterns on these features.

Recently, deep learning methods have been highly successful in solving these classification problems [17–22]. Machine learning has been applied to discern between ASD and TD at the individual level based on brain activity, motor behaviors, facial expressions, and body gestures [23–29]. Based on eye movement data, machine learning using support vector machine (SVM) [30], random forest [31], and shallow neural network with 2 hidden layers [32] has achieved a good accuracy close to 0.9, while test accuracy using deep learning approaches [33, 34] was around 0.6. Jiang et al. [35] designed a combined convolutional neural network (CNN) and SVM framework that achieved an accuracy of 0.85. Within this framework, the CNN was used to reconstruct the eye movement pattern differences between ASD and TD and the SVM was used for classification. However, these reconstructed eye movement differences from training subjects also served as input to the SVM during testing. Furthermore, Ruan et al. [36] designed a modified VGG-16 network to classify photos taken by individuals with ASD and TD, which revealed systematic differences related to visual attention between groups. However, there is still a need for a highly accurate deep learning framework for discriminating between ASD and TD based on eye movement data.

Built upon these models, researchers have investigated the discriminative features discerning between ASD and TD. For example, Ruan et al. [36] applied a layer-wise relevance propagation (LRP) [37] visualization method to identify features with a positive and negative contribution to the predication of ASD. Li et al. [32] adopted a SHAP (Shapley Additive exPlanations) value to identify features with high influence on their model's predication probabilities. Regarding traditional methods such as SVM, the weights of the linear SVM classifier [35] and the difference of mean feature histograms [30] were used to identify discriminative features for the classification of ASD and TD. However, there is still a lack of direct evaluation of the information from these features during the classification in these studies, especially using deep learning models.

To address the above issues and investigate the discriminative features discerning between ASD and TD, we reanalyzed a dataset from our previous study [7] using a newly developed deep learning framework that combined deep learning classification, deep learning segmentation, image ablation and a direct measurement of classification ability. We designed a novel two-stream deep learning network based on VGG-16 using 700 natural scene images and corresponding human fixation maps (HFMs) and achieved a classification accuracy of 0.95. Twelve categories of features were segmented from the images by a deep learning segmentation method of Mask R-CNN [38] model, and the area under curve (AUC) of each category during the classification was calculated based on information only from the feature through an image ablation method. The AUC of the combined top-9 features was 0.92, which allowed an accurate classification of ASD and TD at the individual level, suggesting important roles of these features in the classification. We also obtained a small but informative dataset including 12 images to achieve an AUC of 0.86 through a recursive feature elimination (RFE) method [39]. Together, we have developed a systemic approach to identify and interpret atypical visual attention in ASD.

2 Methods

2.1 Participants

We reanalyzed an eye-tracking dataset from our previous study [7]. Twenty high-functioning participants with ASD and 19 matched participants who are Typically Developing (TD) were recruited. We assessed IQ for participants using the Wechsler Abbreviated Scale of Intelligence (WASI). The ASD group had a full-scale IQ of 108.0 ± 15.6 (mean ± SD) and a mean age of 30.8 ± 11.1 years, while the TD group had a comparable full-scale IQ of 108.2 ± 9.6 and a comparable mean age of 32.3 ± 10.4 years. The two groups were also matched on gender, race, and education. Autism was evaluated using the Autism Diagnostic Observation Schedule (ADOS) [40] and the autism diagnostic interview-revised (ADI-R) [41] or social communication questionnaire (SCQ) when an informant was available (see [7] for details). All ASD participants met the DSM-IV/ICD-10 diagnostic criteria. Participants gave written informed consent, and the experiments were approved by the Caltech Institutional Review Board.

2.2 Task and Dataset

To use a novel computational framework to identify visual attention features accurately discerning between ASD and TD at the individual level, we reanalyzed our dataset collected previously [7]. In the eye-tracking experiment, subjects freely viewed 700 static natural scene images from the OSIE database [42]. Each image was viewed for 3 s, with a random presentation order. Eye movement data were collected using a non-invasive infrared remote Tobii ×300 system with a sampling rate of 300 Hz. The dimensions of the images were 800 × 600 pixels. Based on eye movement data, an 800 × 600 human fixation map (HFM) containing the total gaze time at each pixel location was constructed, smoothed with a Gaussian filter and normalized to the range from 0 to 1 according to [7]. We calculated HFMs based on eye movement data during observations of all 700 images by all subjects. Figure 1 shows HFMs of all subjects while observing a natural scene image, which showed a strong within-group variability in eye movement patterns, and substantial overlap of the patterns between ASD and TD. Because each participant viewed the images differently, we created an HFM for each participant and each image, totaling 27,300 HFMs for 20 participants with ASD and 19 TD participants.

2.3 Deep Learning Model for ASD Recognition

We proposed a two-stream VGG-16 network architecture, which was inspired by the previous models for eye-movement classification or prediction [35, 43–45]. In our network architecture, the first stream extracted deep features of natural scene images, which represented visual stimulus information, and the second stream was used for extracting deep features of human fixation maps, which accounted for eye-movement information. Our two-stream network (Fig. 2b) included two of the same VGGNets modified from the VGG-16 network[46]. We removed the last fully connected layer from the VGG-16, therefore, the VGGNet included 13 convolutional layers (Conv), 5 Max-pooling layers, and 2 fully connected layers (Fc). The two feature maps from the two VGGNets were combined and fed into a three-layer network (ASDNet) including 1 convolutional layer and 2 fully connected layers. In the ASDNet, the convolutional layer had a 1 × 1 filter, and the length of two fully connected layers were 512 and 2, respectively. Together, our two-stream network contained 18 layers excluding the Max-pooling layers. At the end of our two-stream network, a softmax layer was applied to transform real values into probabilities. The activation function of a rectified linear unit (ReLU) [22] was used in the two-stream network. A sample from each subject included 700 images and their corresponding HFMs, and each image and its corresponding HFM were fed into the network simultaneously. The features exacted by the VGGNets from 700 images and 700 HFMs were concatenated by the Concatenation in our two-stream network and were integrated by the first convolutional layer in the ASDNet. In this way, information from 700 images and 700 HFMs were integrated for classification.

We adapted a transfer learning strategy by using the VGGNet [46] pre-trained on the ImageNet challenge dataset [47]. During training, we fixed weights in the two VGGNets, and trained the ASDNet from scratch. The authors tried the commonly used optimizers, including stochastic gradient descent (SGD) [48] and Adam [49]. To obtain the best performance, we finally chose the SGD optimizer with the base learning rate of 10^{-5} and the learning rate policy of "inv" [50]. The weight initializers of the ASDNet were "xavier" and "gaussian" [50]. A 0.5 dropout rate was applied to avoid overfitting. The maximum number of training iterations was 1000. All network training and testing were conducted on the Caffe platform [50].

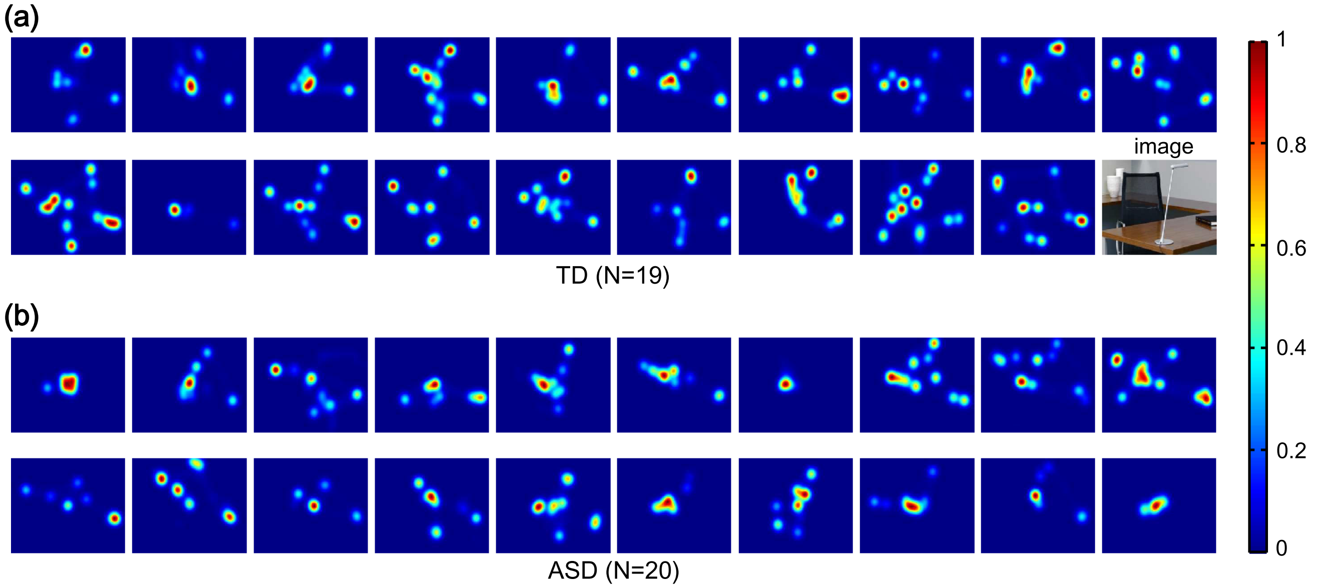


Fig. 1 HFMs during observing a natural scene image. **a** TD subjects. **b** ASD subjects

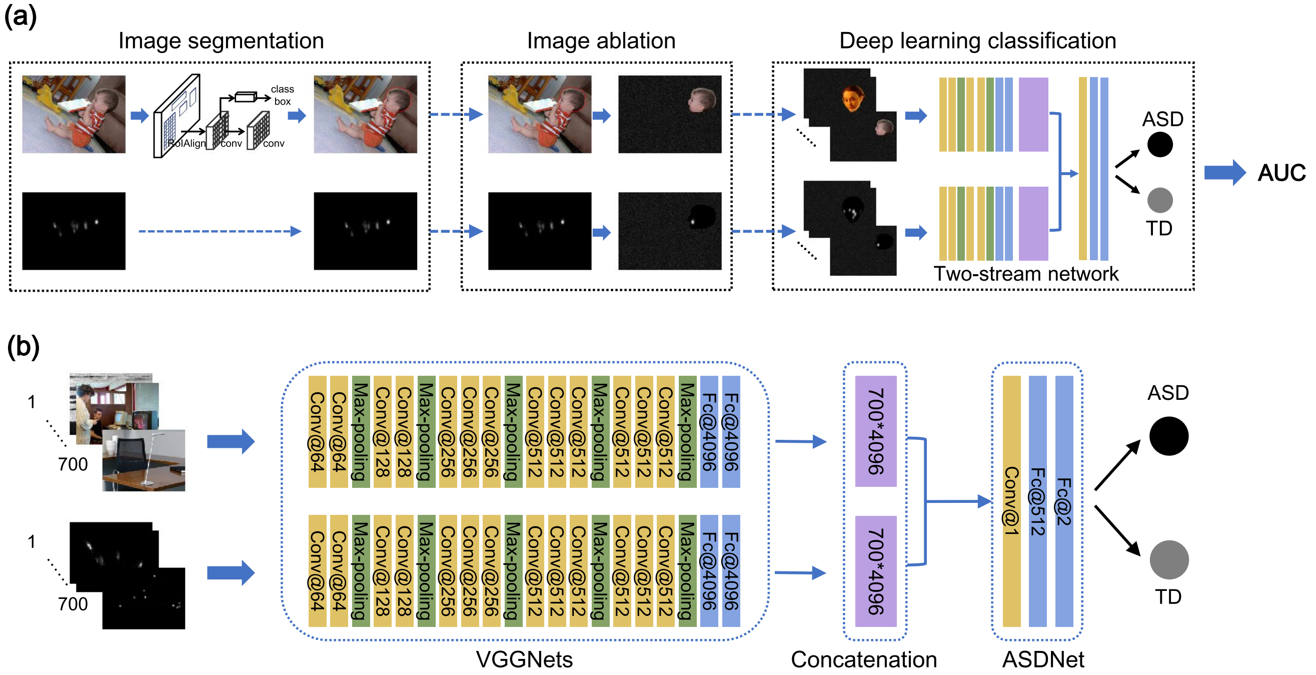


Fig. 2 The architectures of the deep learning framework and classification model. **a** The deep learning framework for identification of visual attention features discerning between ASD and TD that included deep learning segmentation (Mask R-CNN[38]), image ablation, and deep learning classification. **b** The deep learning classification model in (**a**)

2.4 Data Augmentation

To improve the performance of our two-stream network, we applied data augmentation by cropping and flipping the original dataset with the label preserved [22]. The pre-trained VGGNet required a fixed-size input of 224 × 224. The original size of the natural scene images and Human Fixation Maps (HFMs) were 800 × 600, which were resized to 256 × 256. We extracted five 224 × 224 crops from each 256 × 256 image covering its 4 corners and its center, respectively [51]. Flipping augmentation was conducted by mirroring all crops across their vertical axes. Together, we obtained a ten-fold augmentation of the original dataset. We applied the same augmentation procedures in our training and test data.

2.5 Cross-validation

Leave-one-out and 13-fold cross-validation were used in this study. During the leave-one-out cross-validation, one out of 39 subjects was selected sequentially for testing, and the model was trained on the dataset from the remaining 38 subjects, which returned the probabilities of that subject being designated as ASD and TD. There was no overlap between the training and test datasets. This process was repeated for 39 rounds. During the 13-fold cross-validation, the 39 subjects were randomly split into 13 3-subject subsets. One of the 13 subsets was selected sequentially for testing, and the model was trained on the dataset from the 36 subjects of the remaining 12 subsets. We repeated this process for 13 rounds.

2.6 Performance Metric

A test was defined as correct if the probability of the subject belonging to its true group was >0.5. Each subject was tested 10 times because of the ten-fold augmentation, and a classification score was the proportion that the subject was correctly predicted in those 10 times. The model-level accuracy was the averaged classification score across all test subjects in the leave-one-out or 13-fold cross-validation. A subject was correctly recognized when the classification score of the subject was ≥ 0.6. The subject-level Accuracy was the ratio of correctly recognized test subjects to all test subjects in the leave-one-out or 13-fold cross-validation. We also calculated the Sensitivity and Specificity according to [52]. The sensitivity measured the rate of ASD subjects correctly predicted as ASD, while the specificity measured the rate of TD subjects correctly predicted as TD. We plotted the receiver operating characteristic (ROC) curve and computed the corresponding AUC [53] to evaluate the classification performance of our two-stream network.

2.7 t-distributed Stochastic Neighbor Embedding (t-SNE) Visualization

To visualize internal features learned by the two-stream CNN, we applied a t-SNE method to convert high-dimensional features in the two-stream network into 2-dimensional maps. t-SNE is a variation of Stochastic Neighbor Embedding (SNE) [54], a commonly used method for multiple class high-dimensional data visualization [55]. We applied t-SNE visualization for the last two fully connected layers of our two-stream network, converting the 512-dimensional and 2-dimensional representations in the two layers to two 2-dimensional maps, respectively.

2.8 Image Segmentation

We ran a pre-trained Mask R-CNN [38] to segment multiple categories of objects from images on a TensorFlow platform. Next, we wrote a Matlab program to manually fine-tune the segmentation of the Mask R-CNN. In this study, we segmented 12 categories of features: (1) Center-object: the central 2-degree circular area of images with semantic objects; (2) Center-non-object: the central 2-degree circular area of images without semantic objects; (3) Animals: animal faces/heads and bodies; (4) Human-faces: profile and frontal faces of human; (5) Upper-bodies: the human body below the neck and above the waist; (6) Lower-bodies: the human body below the waist; (7) Action-objects: objects that interacted with persons including gaze and operation; (8) Food & drink: anything that can be eaten and drunk except (1)-(7) features; (9) Text: digits, letters, words, and sentences; (10) Indoor-objects: indoor objects except (1–9) features; (11) Outdoor-objects: open-air objects except (1–9) features; (12) Uniform-background: uniform regions without any object.

2.9 Image Ablation

With the parameters of the trained two-stream network fixed, we applied an image ablation method to investigate the contribution of the 12 categories of local features segmented from images to the classification. We retained the regions that contained one category of local feature and occluded all other portions of input images and HFMs with noise masks (Gaussian noise of 0 mean and 0.05 variance) to remove information from the masked areas [56–58]. We constructed the input from all 700 images and HFMs retaining only one category of local feature and passed the input through the trained two-stream network to calculate the AUC score, which was used to measure the classification ability of information from the retained local feature. Classification abilities of the 12 categories of features were quantitatively evaluated by the image ablation using the AUC score.

Noise masks could also be applied to whole images to remove information from the masked images. We retained one or multiple images and masked all other images and HFMs within the 700 images and HFMs to evaluate the classification ability of the retained single image or multiple images.

2.10 Classification Ability Based on Three Levels of Features

The AUC that depicts the tradeoff between hit rates and false alarm rates of classifiers has been commonly used as an objective measure of the classification ability of classifiers [59], with a higher AUC reflecting better classification ability of a classifier. With the classifier fixed, these highly discriminative input data should give rise to high AUC values. With our two-stream network fixed, we used the AUC to measure the classification ability of three levels of features: single-image level, multi-image level, and local-feature level. We used the sklearn library in Python to calculate the AUC.

In our study, we constructed the input from the integrated information of all 700 images and HFMs, and passed them through the trained two-stream network to calculate AUC scores. To investigate the classification ability based on the single image, we calculated the AUC based on a single image, in which we retained the stimulus image and its corresponding HFM, and masked the remaining 699 images and HFMs with Gaussian noise. At the multi-image level, we retained the multiple images and their corresponding HFMs, and masking other images and HFMs with Gaussian noise. At the local-feature level, we retained the regions of images and HFMs containing the same category of the local feature and masked the remaining regions with Gaussian noise.

3 Results

Figure 2a shows the deep learning framework to investigate the discriminative features discerning between ASD and TD. Visual features were segmented from the natural scene images using a Mask R-CNN model [38]. Image ablation retained information only from the area of the feature in the image and corresponding HFM with the remaining areas replaced by Gaussian noise. Data containing information of the same category of the feature with multiple images and HFMs was constructed and passed through a two-stream deep learning model to obtain the AUC, which quantitatively evaluated the classification ability of the information from the feature.

3.1 Performance of The Two-stream Classification Model

The model-level accuracy and cross-entropy loss were plotted against the training iterations in the leave-one-out and 13-fold cross-validation (Fig. 3a–d). The classification accuracies on training and test datasets gradually increased to stable levels, whereas the corresponding losses gradually decreased to stable levels as the training iterations increased. Our two-stream network achieved 0.92 and 0.84 model-level accuracy in the leave-one-out and 13-fold cross-validation tests, respectively (Fig. 3a, b). As shown in Fig. 3e, f, we plotted the ROC curves of the leave-one-out cross-validation (Fig. 3e) and 13-fold cross-validation (Fig. 3f). In the leave-one-out cross-validation, our model obtained 0.95 subject-level Accuracy, 1.00 Sensitivity, 0.89 Specificity, and 0.93 AUC. In the 13-fold cross-validation, the subject-level Accuracy reached 0.85, with 0.80 Sensitivity, 0.89 Specificity and 0.91 AUC, respectively. As shown in Table 1.

We compared our method with other state-of-the-art methods [30, 31, 35]. Liu et al. [30] used a K-means + SVM method while Jiang et al. adopted a CNN + SVM framework [35] and random forest method [31] to recognize those with ASD using eye movement data. Our new two-stream deep learning network outperformed other models in Accuracy, Sensitivity, Specificity (as shown in Table 1; and the AUC was only 0.01 lower than that from [31]).

3.2 t-SNE Visualization

To visualize internal features in the two-stream CNN associated with the classification, we applied t-SNE to visualize the high-dimensional features in the last two layers of the ASDNet (Fig. 3g–j). Figure 3G–H shows two t-SNE plots of the two layers from a typical leave-one-out model, respectively. Figure 3i, j show the t-SNE plots from a typical 13-fold cross-validation model. Each point in these t-SNE plots represented a subject. In the leave-one-out model, t-SNE distribution of features from ASD and TD subjects

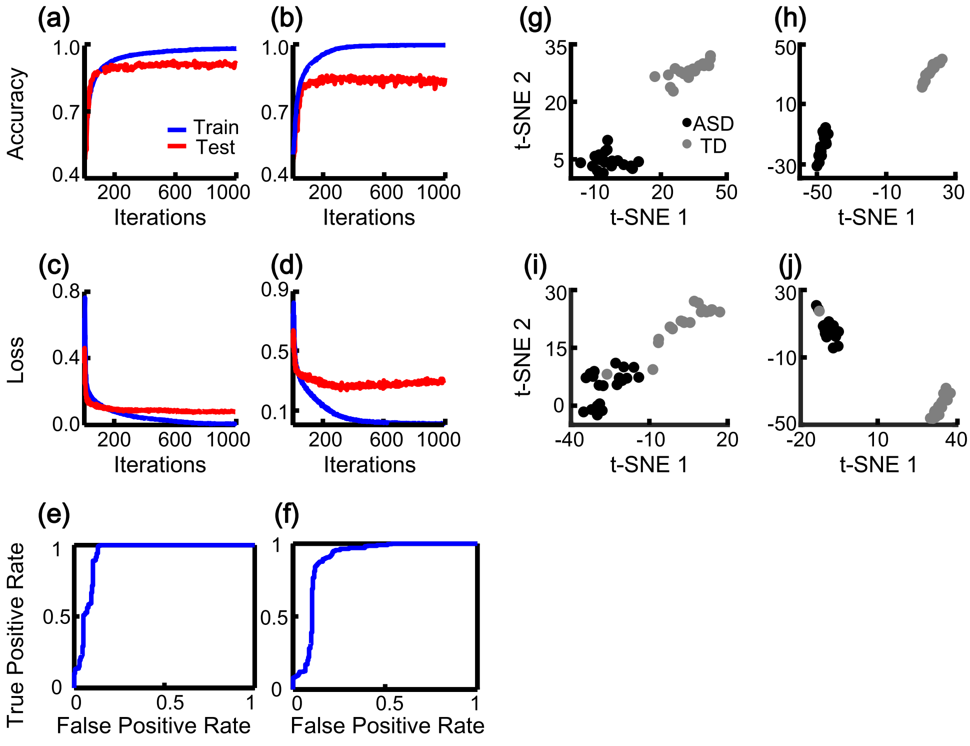


Fig. 3 The performance of the two-stream deep learning network. **a** The accuracy curves of the network in the leave-one-out validation. **b** The accuracy curves in the 13-fold validation. **c** The loss curves in the leave-one-out validation. **d** The loss curves in the 13-fold validation. **e–f** The ROC curves of our two-stream network in the leave-one-out cross-validation (**e**) and the 13-fold cross-validation (**f**). **g–h** The t-SNE visualization of high-dimensional features of the second last (**g**) and the last layer (**h**) in a leave-one-out validation. Each dot represents a subject. **i–j** The t-SNE visualization of the second last (**i**) and the last layer (**j**) in a 13-fold validation

Table 1 A quantitative comparison between our model and other state-of-the-art models

Methods	Models	Cross-validation	Accuracy	Sensitivity	Specificity	AUC
Liu et al.[30]	K-means + SVM	Leave-one-out	0.89	0.93	0.86	0.90
Jiang et al.[35]	CNN + SVM	Leave-one-out	0.85	0.83	0.87	0.89
Jiang et al.[31]	Random forest	Leave-one-out	0.86	0.91	0.83	0.94
Ours	Two-stream network	Leave-one-out	**0.95**	**1.00**	**0.89**	**0.93**
		13-fold (three-out)	**0.85**	**0.80**	**0.89**	**0.91**

Bold values indicate results of two-stream network was well separated in the last two layers, demonstrating that category representation in the two-stream network was unambiguous. In the 13-fold model, the vast majority of data from ASD and TD subjects were again well separated, with only one TD subject misclassified (Fig. 3i), which was consistent with the lower accuracy of 13-fold cross-validation compared with the results of leave-one-out. In addition, better separation was found in the last layer in the two examples, suggesting that ASD and TD features became more distinguishable along the network.

3.3 Classification Ability Based on Single-image Level

The previous classification results relied on all 700 images and corresponding HFMs. Here, we evaluated the classification ability of information from a single image (see details in Sect. 2). Figure 4a shows the AUC values based on information from the 700 single images in a descending order. Figure 4b shows images with the top 10 and bottom 5 AUC values. These AUC values in Fig. 4a (<0.55) were far smaller than the AUC value of our two-stream network (0.93) based on all 700 images and HFMs, suggesting the contribution of information from a single image was limited and the combined information from multiple images played an important role in our model.

3.4 Classification Ability Based on Multi-image Level

We calculated AUCs based on multiple images and HFMs (Fig. 4c). The AUC was about 0.74 based on data from the top 50 images and gradually increased to 0.96 with the top 250 images. It stayed around this high level with more data integrated until all 700 images and HFMs were included. In fact, the AUC value based on all 700 images and HFMs was 0.93, which was smaller than the value based on information from the top 250 images, suggesting that the combined

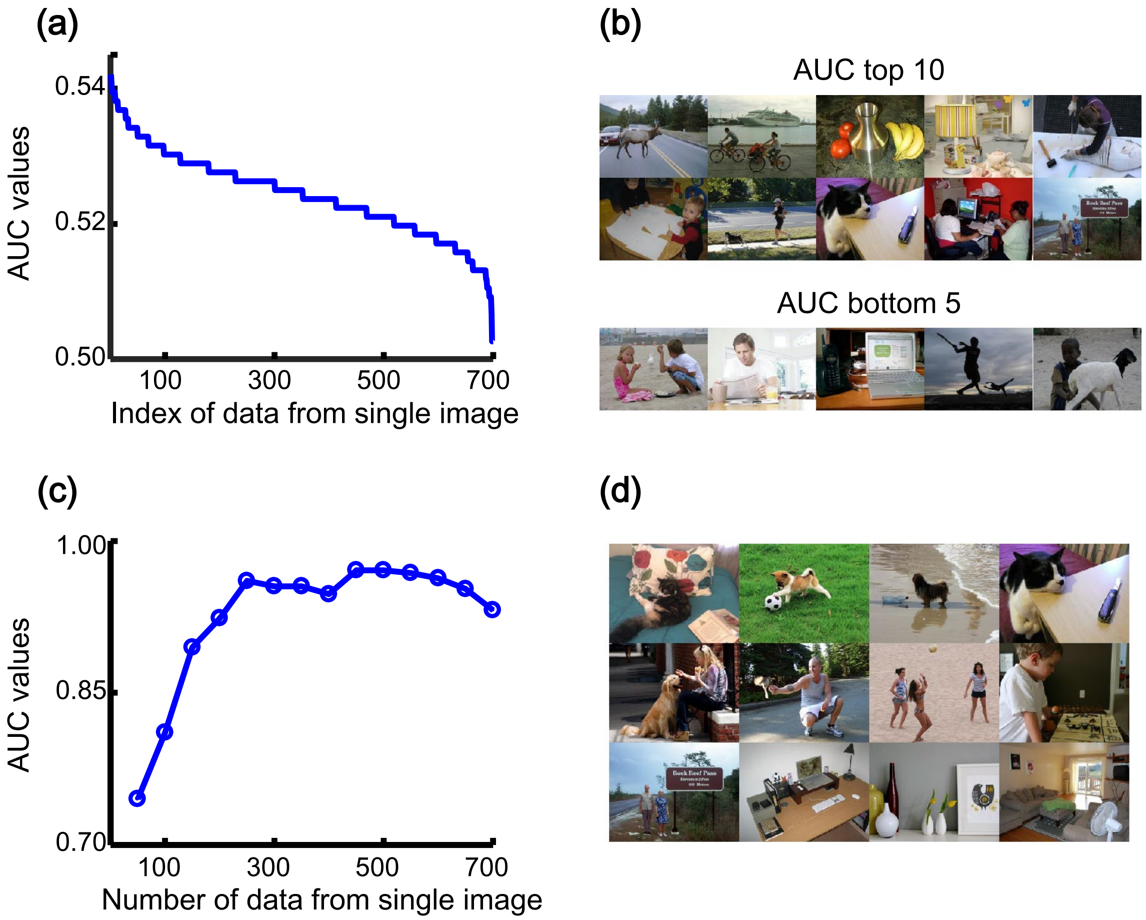


Fig. 4 Classification ability of information from single and multiple images. **a** The AUCs based on information from a single image. X: the index of data from a single image according to its AUC values in a descending order; Y: AUC values of information from a single image. **b** Images with the top 10 and bottom 5 AUCs. **c** The AUCs based on data from multiple images. X: the number of data from a single image combined; Y: the AUC values of the combined data. **d** The 12 image dataset with an AUC of 0.86

information from all 700 images was redundant, and there were both synergistic and antagonistic interactions between information from single images during the integration. To find an efficient image dataset for the classification by a RFE method [39], we started from the top 250 images, and each time replaced data of one image with Gaussian noise if this replacement caused the smallest AUC reduction. We continued this process and found a dataset including 12 images with an AUC of 0.86 (Fig. 4d), suggesting the potential to find a highly efficient image dataset to discern between ASD and TD.

3.5 Classification Ability Based on Local-feature Level

We segmented the top 250 images through a pre-trained Mask R-CNN model [38]. After manually fine-tuning the segmentation, we classified the segmented features into 12 categories (Fig. 5). To investigate the classification ability of information from these features, we calculated the AUCs based on each category of these features while replacing all remaining parts of the images with Gaussian noise. Figure 6a showed these features with the other parts of images replaced with Gaussian noise. A baseline AUC was calculated for each feature by randomly retaining image parts with the same size of the feature in each image. The classification ability based on a feature was higher than that of randomly selected areas if its AUC was greater than its baseline. As shown in Fig. 6b, there were 10 features with AUCs higher than baseline, including Center-object, Center-non-object, Animals, Human-faces, Lower-bodies, Food & drink, Action-objects, Text, Indoor-objects, and Outdoor-objects. However, AUCs of Upper-bodies and Uniform-background were lower than their baselines. The averaged AUC of the 11 features (excluding the Uniform-background) was significantly higher than the average of their baselines (Rank-sum test, $p = 0.015$), as shown in the rightmost of Fig. 6b. We combined information from the 11 features sequentially according to their AUCs in a high-to-low sequence. As the number of combined features increased, the values of AUC gradually increased (Fig. 6c) and reached its peak of 0.92 with the top-9 features combined (excluding the Action-objects and Lower-bodies) that consisted of 44% of the top 250 images in size. The AUC based on the 11 combined features was 0.91. These AUC values were very close to the AUC value of 0.93 based on data from all 700 images, suggesting that these features played an important role in distinguishing ASD from TD during the classification.

4 Discussion

In this study, we developed a new systematic framework to identify discriminative features for the classification of ASD and TD, which combined deep learning classification, deep learning segmentation, image ablation and a direct measurement of classification ability (AUC). In the framework, we designed a novel two-stream deep learning network based on VGG-16 combining image and eye movement information for discerning between ASD and TD, and obtained accuracies of 0.95 and 0.85 in the leave-one-out and 13-fold cross-validation, respectively. We used VGG-16 [46] in our two-stream network, because it has been one of the most popular CNN models [44], and

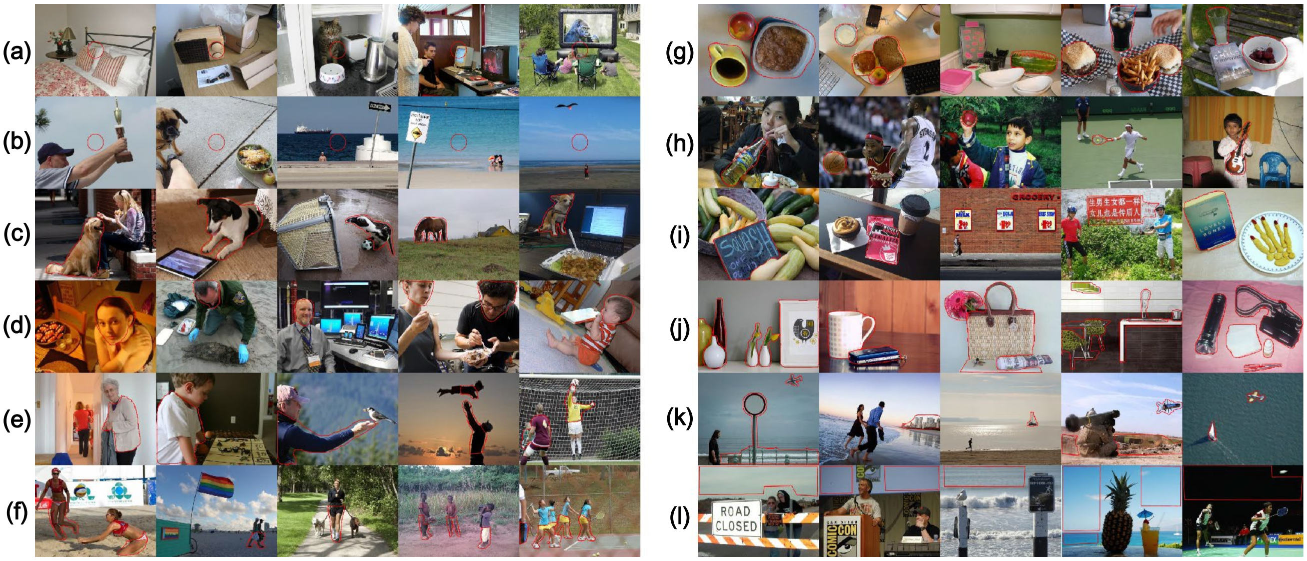


Fig. 5 The 12 categories of features segmented from images (outlined in red lines). **a** Center-object. **b** Center-non-object. **c** Animals. **d** Human-faces. **e** Upper-bodies. **f** Lower-bodies. **g** Food & drink. **h** Action-objects. **i** Text. **j** Indoor-objects. **k** Outdoor-objects. **l** Uniform-background

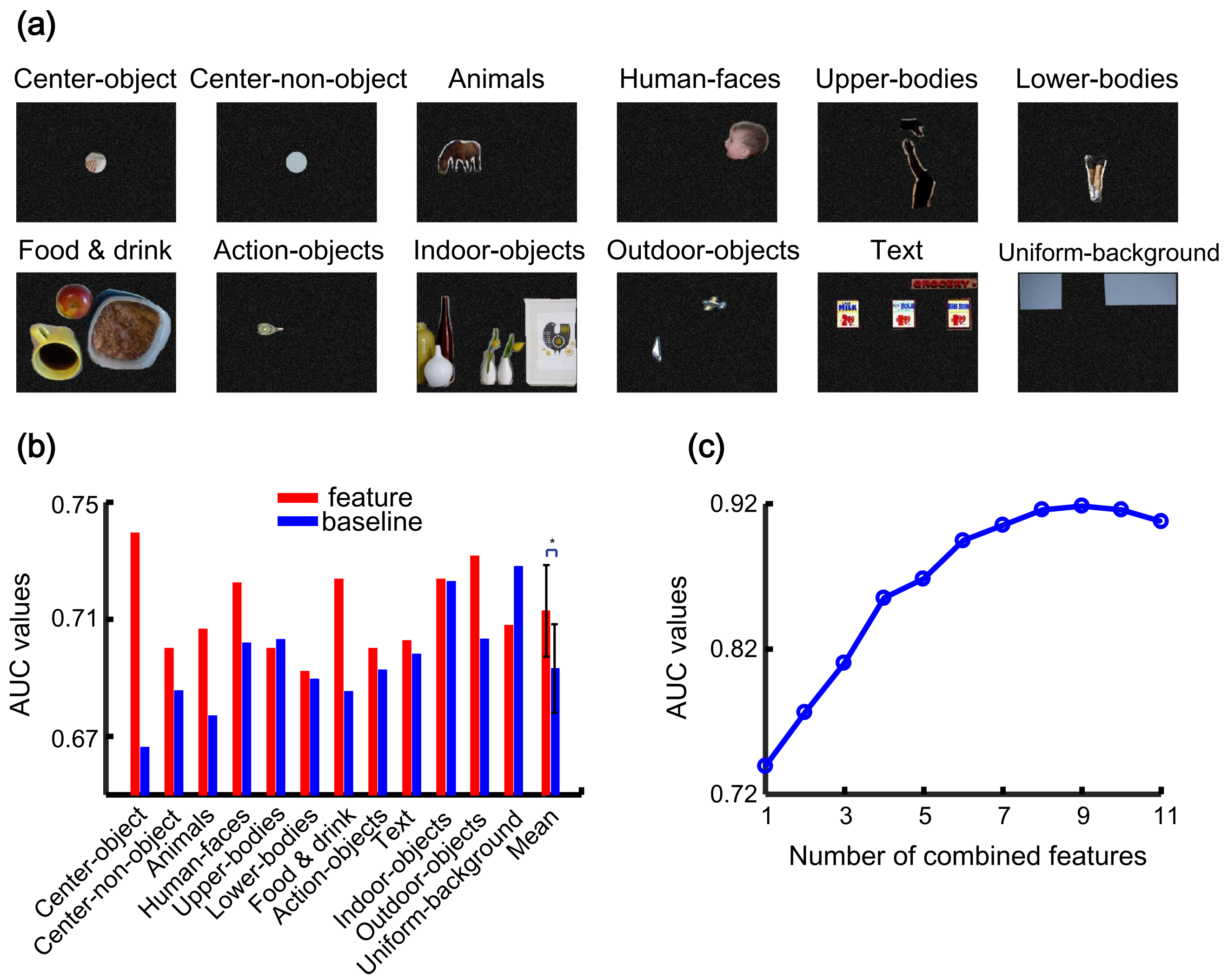


Fig. 6 Classification ability of information from features segmented from images. **a** Examples of 12 categories of features segmented from images with the remainder of the images masked by Gaussian noise. **b** The AUCs of the 12 categories of features (red) and corresponding baselines (blue), and the averaged AUC of the 11 categories of features (Mean) with the Uniform-background excluded. **c** The AUCs based on combined features. X: the number of features combined; Y: the AUC values of combined features

has been used in studies on recognition of atypical eye-movement behavior in ASD [35, 36, 60] and prediction of the saliency maps based on eye-movements in control subjects [43–45, 61–63]. Our two-stream network achieved the state-of-the-art performance with a gain of 6% in classification accuracy [30, 31, 35]. It is worth noting that although our VGG-16 based model outperformed prior models in predicting eye movement, our present model was more complex (in both its structure and number of parameters) than prior models and was thus more likely to fit the data. Deep learning networks have also been developed for ASD recognition based on fMRI signals including auto-encoders [19, 64–67], one or multi-stream CNNs [29, 68] with accuracies at about 0.70–0.96. The accuracy of our model is comparable with the best accuracy of these deep learning models.

Leveraging the high predication accuracy of our two-stream model, we evaluated the classification ability (AUC) of a feature based on information only from the feature through an image ablation method that retained the area of the feature with the remaining areas replaced by Gaussian noise. These features were segmented from the images by the Mask R-CNN model and were slightly fine-tuned manually. Within this framework, the high accuracy of our model made a good basis for finding these discriminative features, and deep learning segmentation allowed us to consistently segment features from images with high localization precision, and the image ablation combined with AUC measurement allowed us to evaluate the classification ability of each feature and that of their combination directly and quantitatively. The group averaging methods have been commonly used to identify important features of ASD in previous studies, while it could not accurately classify ASD and TD at the individual level based on these group differences [69–72]. In contrast, the combined top-9 features identified in this study gave rise to an AUC of 0.92 in our network, permitting an accurate classification. Recently, the LRP method [36] and SHAP method [32] have been applied to identify pixel-wise information important for ASD identification. These methods identified areas with high influence on the predication probabilities, but the tradeoff between hit rates and false alarm rates has not been considered in these methods. Thus, here we have developed a systematic framework to identify discriminative features discerning between ASD and TD. This framework could also be applied to the identification of discriminative image-related features in a variety of brain disorders beyond ASD.

Through this method, we evaluated 12 categories of features in our natural scene images. Previous studies using natural scenes have found fixation pattern differences in: human faces[4, 7, 70, 73, 74]; faces of animals and cartoons [7]; person and people [2, 72, 75–77]; bodies [4, 71, 78]; gazed objects [79]; motion, smell, touch objects [7]; non-uniform background [4, 7, 72]; and image center [7] by group averaging and statistical analysis. Consistent with previous findings, Human-faces, Center-object, Center-Non-object, Action-objects, and Text were identified as discriminative features in this study, but we find that Center-object was more discriminative than the Center-non-object, which was not distinguished in previous studies. Animal faces were included in the Face category in previous studies, this time we identified the Animals including face/head and the whole body as a discriminative feature, and further analysis showed that animal face was not more discriminative than the body of the animal (AUC of face vs body: 0.68 vs 0.70). To evaluate the discriminability of these features more accurately, a baseline AUC was calculated for each feature by randomly retaining image parts with the same size of the feature in each image. We found that AUCs of 2 categories (the Uniform-background and Upper-bodies) were lower than their baselines, suggesting that the two features were not more discriminative than the randomly selected areas in the images and played a less important role in the classification. We found two new categories of discriminative features: Food & drink and Outdoor-objects, which have not been clearly identified in previous studies. Food selectivity has been described as a common feature of ASD [80], such as eating a narrow variety of foods, requiring specific presentations of foods and specific utensils, and eating only low texture foods [81, 82], which have also been classified as part of repetitive behaviors in ASD [83]. Our result suggests that altered attention to stimuli of Food & drink may contribute to the food selectivity in ASD, which may stem from different food selectivity. The altered attention to the Outdoor-objects might be related to less outdoor experience in children with ASD, who spend more than twice as much time indoors compared to TD children [84–86]. The value of AUC based on the combined top-9 features reached 0.92, which was very close to the AUC value of 0.93 based on data from all 700 images, confirming the importance of these features during the classification. Together, the Center-object, Food & drink, Outdoor-objects, Animals, and Human-faces were the most discriminative features by considering both the AUC values of these features and corresponding baselines.

We also characterized the classification ability of information from single image, and found that information from multiple images was necessary for accurate ASD identification. Recently, Liaqat et al. [34] developed deep-learning approaches to predict ASD with an accuracy of 0.62 based on a single image and scan-path data. The high accuracy of that study might result from more information being derived from a scan-path than that from an HFM or a difference between the image-based division of training and test datasets in their study and the subject-based division in our study. We found a twelve-image dataset with an AUC of 0.86, suggesting that it was possible to recognize abnormal eye movements in ASD based on a small but informative dataset. A limitation is that the size of our dataset (39 subjects) is still small. The scarce availability and difficulty of acquiring eye-movement datasets have been a key challenge in ASD research [87]. In the present study, we applied the data augmentation methods of cropping and flipping to increase the dataset size as in previous studies [60, 87, 88] for eye-movement data. But visual attention behaviors may be changed with the images cropped or flipped. Further investigation is required to test the robustness and generalization of these discriminative features. However, it is worth noting that one advantage of our deep learning model is to exact features from a large dataset, and the performance of our model will be improved with more data from ASD and TD participants. Together, our present approach will not only provide a novel and powerful tool to identify and interpret abnormal visual attention in ASD but also facilitate the identification of eye-movement biomarkers for ASD.

Acknowledgements We are grateful for the financial support from the National Key R&D Program of China (2017YFC1307500), the National Natural Science Foundation of China (62027804 and 31671108) as well as the Scientific Instrument Innovation Team of the Chinese Academy of Sciences (GJJSTD20180002).

Data Availability All data needed to evaluate the conclusions are present in the paper. Additional data related to this paper may be requested from the authors.

Declarations

Conflict of Interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Ethical Approval Participants gave written informed consent and the experiments were approved by the Caltech Institutional Review Board.

References

1. Birmingham E, Cerf M, Adolphs R (2011) Comparing social attention in autism and amygdala lesions: effects of stimulus and task condition. Soc Neurosci 6(5–6):420–435. https://doi.org/10.1080/17470919.2011.561547
2. Chawarska K, Macari S, Shic F (2013) Decreased spontaneous attention to social scenes in 6-month-old infants later diagnosed with autism spectrum disorders. Biol Psychiat 74(3):195–203. https://doi.org/10.1016/j.biopsych.2012.11.022
3. Shic F, Bradshaw J, Klin A, Scassellati B, Chawarska K (2011) Limited activity monitoring in toddlers with autism spectrum disorder. Brain Res 1380:246–254. https://doi.org/10.1016/j.brainres.2010.11.074
4. Rice K, Moriuchi JM, Jones W, Klin A (2012) Parsing heterogeneity in autism spectrum disorders: visual scanning of dynamic social scenes in school-aged children. J Am Acad Child Adolesc Psychiatry 51(3):238–248. https://doi.org/10.1016/j.jaac.2011.12.017
5. Dawson G, Webb SJ, McPartland J (2005) Understanding the nature of face processing impairment in autism: insights from behavioral and electrophysiological studies. Dev Neuropsychol 27(3):403–424. https://doi.org/10.1207/s15326942dn2703_6
6. Sasson NJ, Elison JT, Turner-Brown LM, Dichter GS, Bodfish JW (2011) Brief report: circumscribed attention in young children with autism. J Autism Dev Disord 41(2):242–247. https://doi.org/10.1007/s10803-010-1038-3
7. Wang S, Jiang M, Duchesne XM, Laugeson EA, Kennedy DP, Adolphs R et al (2015) Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking. Neuron 88(3):604–616. https://doi.org/10.1016/j.neuron.2015.09.042
8. Duan H, Zhai G, Min X, Che Z, Fang Y, Yang X et al (2019) A dataset of eye movements for the children with autism spectrum disorder. In: Proceedings of the 10th ACM Multimedia Systems Conference: pp. 255–260. https://doi.org/10.1145/3304109.3325818
9. Arru G, Mazumdar P, Battisti F (2019) Exploiting visual behaviour for autism spectrum disorder identification. In: 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW): IEEE pp. 637–640. https://doi.org/10.1109/ICMEW.2019.00123
10. Osterling J, Dawson G (1994) Early recognition of children with autism: a study of first birthday home videotapes. J Autism Dev Disord 24(3):247–257. https://doi.org/10.1007/bf02172225
11. Mundy P, Sigman M, Kasari C (1994) The theory of mind and joint-attention deficits in autism. Understanding other minds: perspectives from autism, pp 181–203. https://psycnet.apa.org/record/1993-98373-008. Accessed 9 Mar 2022
12. Leekam SR, Ramsden CAH (2006) Dyadic orienting and joint attention in preschool children with autism. J Autism Dev Disord 36(2):185–197. https://doi.org/10.1007/s10803-005-0054-1
13. Brenner LA, Turner KC, Mueller R-A (2007) Eye movement and visual search: are there elementary abnormalities in autism? J Autism Dev Disord 37(7):1289–1309. https://doi.org/10.1007/s10803-006-0277-9
14. Mundy P, Sullivan L, Mastergeorge AM (2009) A parallel and distributed-processing model of joint attention, social cognition and autism. Autism Res 2(1):2–21. https://doi.org/10.1002/aur.61
15. Chevallier C, Kohls G, Troiani V, Brodkin ES, Schultz RT (2012) The social motivation theory of autism. Trends Cogn Sci 16(4):231–239. https://doi.org/10.1016/j.tics.2012.02.007
16. Guillon Q, Hadjikhani N, Baduel S, Roge B (2014) Visual social attention in autism spectrum disorder: insights from eye tracking studies. Neurosci Biobehav Rev 42:279–297. https://doi.org/10.1016/j.neubiorev.2014.03.013
17. Graves A, Mohamed A-r, Hinton G (2013) Speech recognition with deep recurrent neural networks. In: IEEE International Conference on Acoustics: speech and signal processing pp. 6645–6649. https://doi.org/10.1109/ICASSP.2013.6638947
18. Kather JN, Pearson AT, Halama N, Jaeger D, Krause J, Loosen SH et al (2019) Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nature Med. https://doi.org/10.1038/s41591-019-0462-y
19. Hazlett HC, Gu H, Munsell BC, Kim SH, Styner M, Wolff JJ et al (2017) Early brain development in infants at high risk for autism spectrum disorder. Nature. https://doi.org/10.1038/nature21369
20. Kim J, Calhoun VD, Shim E, Lee J-H (2016) Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: evidence from whole-brain resting-state functional connectivity patterns of schizophrenia. Neuroimage 124:127–146. https://doi.org/10.1016/j.neuroimage.2015.05.018
21. Choi H, Jin KH (2018) Predicting cognitive decline with deep learning of brain metabolism and amyloid imaging. Behav Brain Res 344:103–109. https://doi.org/10.1016/j.bbr.2018.02.017
22. Krizhevsky A, Sutskever I, Hinton GE (2017) ImageNet classification with deep convolutional neural networks. Commun ACM 60(6):84–90. https://doi.org/10.1145/3065386
23. Crippa A, Salvatore C, Perego P, Forti S, Nobile M, Molteni M et al (2015) Use of machine learning to identify children with autism and their motor abnormalities. J Autism Dev Disord 45(7):2146–2156. https://doi.org/10.1007/s10803-015-2379-8
24. Deshpande G, Libero LE, Sreenivasan KR, Deshpande HD, Kana RK (2013) Identification of neural connectivity signatures of autism using machine learning. Front Hum Neurosci 7(670):1–15. https://doi.org/10.3389/fnhum.2013.00670
25. Duda M, Kosmicki JA, Wall DP (2014) Testing the accuracy of an observation-based classifier for rapid detection of autism risk. Transl Psychiatry 5(4):e556–e556. https://doi.org/10.1038/tp.2014.65
26. Kosmicki JA, Sochat V, Duda M, Wall DP (2015) Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning. Transl Psychiatry 5(2):e514–e514. https://doi.org/10.1038/tp.2015.7
27. Stahl D, Pickles A, Elsabbagh M, Johnson MH (2012) Novel machine learning methods for ERP analysis: a validation from research on infants at risk for autism. Dev Neuropsychol 37(3):274–298. https://doi.org/10.1080/87565641.2011.650808
28. Zhou Y, Yu F, Duong T (2014) Multiparametric MRI characterization and prediction in autism spectrum disorder using graph theory and machine learning. PLoS One 9(6):e90405. https://doi.org/10.1371/journal.pone.0090405
29. Li X, Dvornek NC, Zhuang J, Ventola P, Duncan JS (2018) Brain Biomarker Interpretation in ASD Using Deep Learning and fMRI. In: International conference on medical image computing and computer-assisted intervention: Springer pp. 206–214. https://doi.org/10.1007/978-3-030-00931-1_24
30. Liu W, Li M, Yi L (2016) Identifying children with autism spectrum disorder based on their face processing abnormality: a machine learning framework. Autism Res 9(8):888–898. https://doi.org/10.1002/aur.1615
31. Jiang M, Francis SM, Srishyla D, Conelea C, Zhao Q, Jacob S et al. (2019) Classifying Individuals with ASD Through Facial Emotion Recognition and Eye-Tracking. In: 2019 41st Annual International Conference of the Ieee Engineering in Medicine and Biology Society (EMBC): IEEE pp. 6063–6068. https://doi.org/10.1109/EMBC.2019.8857005
32. Li B, Barney E, Hudac C, Nuechterlein N, Ventola P, Shapiro L et al. (2020) Selection of Eye-Tracking Stimuli for Prediction by Sparsely Grouped Input Variables for Neural Networks: towards Biomarker Refinement for Autism. In: ACM Symposium on Eye Tracking Research and Applications: Association for Computing Machinery pp. 1–8. https://doi.org/10.1145/3379155.3391334

33. Tao Y, Shyu M-L (2019) SP-ASDNET: CNN-LSTM based asd classification model using observer scanpaths. In: IEEE International Conference on Multimedia and Expo Workshops (ICMEW): IEEE pp. 641–646. https://doi.org/10.1109/icmew.2019.00124
34. Liaqat S, Wu C, Duggirala PR, Cheung S-CS, Chuah C-N, Ozonoff S et al (2021) Predicting ASD diagnosis in children with synthetic and image-based eye gaze data. Signal Process Image Commun 94:116198. https://doi.org/10.1016/j.image.2021.116198
35. Jiang M, Zhao Q (2017) Learning Visual Attention to Identify People with Autism Spectrum Disorder. In: Proceedings of the IEEE International Conference on Computer Vision: IEEE pp. 3267–3276. https://doi.org/10.1109/iccv.2017.354
36. Ruan MD, Webster PJ, Li X, Wang S (2021) Deep neural network reveals the world of autism from a first-person perspective. Autism Res 14(2):333–342. https://doi.org/10.1002/aur.2376
37. Bach S, Binder A, Montavon G, Klauschen F, Muller KR, Samek W (2015) On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One 10(7):e0130140. https://doi.org/10.1371/journal.pone.0130140
38. He KM, Gkioxari G, Dollar P, Girshick R (2020) Mask R-CNN. IEEE Trans Pattern Anal Mach Intell 42(2):386–397. https://doi.org/10.1109/tpami.2018.2844175
39. Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. Mach Learn 46(1–3):389–422. https://doi.org/10.1023/a:1012487302797
40. Lord C, Rutter M, Goode S, Heemsbergen J, Jordan H, Mawhood L et al (1989) Autism diagnostic observation schedule: a standardized observation of communicative and social behavior. J Autism Dev Disord 19(2):185–212. https://doi.org/10.1007/bf02211841
41. Le Couteur A, Rutter M, Lord C, Rios P, Robertson S, Holdgrafer M et al (1989) Autism diagnostic interview: a standardized investigator-based instrument. J Autism Dev Disord 19(3):363–387. https://doi.org/10.1007/bf02212936
42. Xu J, Jiang M, Wang S, Kankanhalli MS, Zhao Q (2014) Predicting human gaze beyond pixels. J Vis 14(1):28–28. https://doi.org/10.1167/14.1.28
43. Huang X, Shen C, Boix X, Zhao Q (2015) SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In: 2015 IEEE International Conference on computer vision (ICCV): IEEE pp. 262–270. https://doi.org/10.1109/iccv.2015.38
44. Cornia M, Baraldi L, Serra G, Cucchiara R (2016) A deep multi-level network for saliency prediction. In: 23rd International Conference on Pattern Recognition (ICPR): pp. 3488–3493. https://doi.org/10.1109/ICPR.2016.7900174
45. Fan S, Shen Z, Jiang M, Koenig BL, Xu J, Kankanhalli MS et al (2018) Emotional Attention: a Study of Image Sentiment and Visual Attention. In: 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR): IEEE pp. 7521–7531. https://doi.org/10.1109/cvpr.2018.00785
46. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
47. Deng J, Dong W, Socher R, Li L-J, Li K, Li F-F (2009) ImageNet: A Large-Scale Hierarchical Image Database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition: IEEE pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848
48. Ruder S (2016) An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747
49. Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980
50. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R et al (2014) Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM international conference on Multimedia: pp. 675–678. https://doi.org/10.1145/2647868.2654889
51. Chatfield K, Simonyan K, Vedaldi A, Zisserman A (2014) Return of the devil in the details: Delving deep into convolutional nets. arXiv preprint arXiv:1405.3531
52. Fan X, Yao Q, Cai Y, Miao F, Sun F, Li Y (2018) Multiscaled fusion of deep convolutional neural networks for screening atrial fibrillation from single lead short ECG recordings. IEEE J Biomed Health Inform 22(6):1744–1753. https://doi.org/10.1109/jbhi.2018.2858789
53. Green DM, Swets JA, Emmerich DS (1966) Signal detection theory and psychophysics. Wiley, New York
54. Hinton GE, Roweis ST (2002) Stochastic neighbor embedding. Advances in neural information processing systems 15:857–864. https://proceedings.neurips.cc/paper/2002/file/6150ccc6069bea6b5716254057a194ef-Paper.pdf. Accessed 9 Mar 2022
55. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM et al (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature. https://doi.org/10.1038/nature21056
56. Koker T, Mireshghallah F, Titcombe T, Kaissis G (2021) U-Noise: Learnable Noise Masks for Interpretable Image Segmentation. arXiv preprint arXiv:2101.05791
57. Aminu M, Ahmad NA, Noor MHM (2021) Covid-19 detection via deep neural network and occlusion sensitivity maps. Alex Eng J 60(5):4829–4855. https://doi.org/10.1016/j.aej.2021.03.052
58. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: European conference on computer vision (ECCV): Springer pp. 818–833. https://linkspringer.53yu.com/content/pdf/10.1007/978-3-319-10590-1_53.pdf. Accessed 9 Mar 2022
59. Fawcett T (2006) An introduction to ROC analysis. Pattern Recogn Lett 27(8):861–874. https://doi.org/10.1016/j.patrec.2005.10.010
60. Nebout A, Wei W, Liu Z, Huang L, Le Meur O (2019) Predicting saliency maps for asd people. In: 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW): IEEE pp. 629–632. https://doi.org/10.1109/ICMEW.2019.00121
61. Jetley S, Murray N, Vig E, Ieee (2016) End-to-End Saliency Mapping via Probability Distribution Prediction. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR): pp. 5753–5761. https://doi.org/10.1109/cvpr.2016.620
62. Kümmerer M, Wallis TS, Bethge M (2016) DeepGaze II: reading fixations from deep features trained on object recognition. arXiv preprint arXiv:1610.01563.
63. Kruthiventi SSS, Ayush K, Babu RV (2017) DeepFix: a fully convolutional neural network for predicting human eye fixations. IEEE Trans Image Process 26(9):4446–4456. https://doi.org/10.1109/tip.2017.2710620
64. Heinsfeld AS, Franco AR, Cameron Craddock R, Buchweitz A, Meneguzzi F (2018) Identification of autism spectrum disorder using deep learning and the ABIDE dataset. Neuroimage-Clin 17:16–23. https://doi.org/10.1016/j.nicl.2017.08.017
65. Guo X, Dominick KC, Minai AA, Li H, Erickson CA, Lu LJ (2017) Diagnosing autism spectrum disorder from brain resting-state functional connectivity patterns using a deep neural network with a novel feature selection method. Front Neurosci 11:460. https://doi.org/10.3389/fnins.2017.00460
66. Li H, Parikh NA, He L (2018) A novel transfer learning approach to enhance deep neural network classification of brain functional connectomes. Front Neurosci 12:491. https://doi.org/10.3389/fnins.2018.00491
67. Xiao Z, Wu J, Wang C, Jia N, Yang X (2019) Computer-aided diagnosis of school-aged children with ASD using full frequency bands and enhanced SAE: a multi-institution study. Exp Ther Med 17(5):4055–4063. https://doi.org/10.3892/etm.2019.7448
68. Aghdam MA, Sharifi A, Pedram MM (2019) Diagnosis of autism spectrum disorders in young children based on resting-state functional magnetic resonance imaging data using convolutional

neural networks. J Digit Imaging 32(6):899–918. https://doi.org/10.1007/s10278-019-00196-1

69. Griffin JW, Scherf KS (2020) Does decreased visual attention to faces underlie difficulties interpreting eye gaze cues in autism? Mol Autism 11(1):1–14. https://doi.org/10.1186/s13229-020-00361-2
70. Riby DM, Hancock PJB (2009) Do faces capture the attention of individuals with williams syndrome or autism? Evidence from tracking eye movements. J Autism Dev Disord 39(3):421–431. https://doi.org/10.1007/s10803-008-0641-z
71. Jones W, Klin A (2013) Attention to eyes is present but in decline in 2–6-month-old infants later diagnosed with autism. Nature. https://doi.org/10.1038/nature12715
72. Fletcher-Watson S, Leekam SR, Benson V, Frank MC, Findlay JM (2009) Eye-movements reveal attention to social information in autism spectrum disorder. Neuropsychologia 47(1):248–257. https://doi.org/10.1016/j.neuropsychologia.2008.07.016
73. Startsev M, Dorr M (2019) Classifying autism spectrum disorder based on scanpaths and saliency. In: 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW): IEEE pp. 633–636. https://doi.org/10.1109/ICMEW.2019.00122
74. Constantino JN, Kennon-McGill S, Weichselbaum C, Marrus N, Haider A, Glowinski AL et al (2017) Infant viewing of social scenes is under genetic control and is atypical in autism. Nature. https://doi.org/10.1038/nature22999
75. Karlsson MF, Galazka MA, Gillberg C, Gillberg C, Miniscalco C, Billstedt E et al (2019) Social scene perception in autism spectrum disorder: an eye-tracking and pupillometric study. J Clin Exp Neuropsychol 41(10):1024–1032. https://doi.org/10.1080/13803395.2019.1646214
76. Chawarska K, Ye S, Shic F, Chen L (2016) Multilevel differences in spontaneous social attention in toddlers with autism spectrum disorder. Child Dev 87(2):543–557. https://doi.org/10.1111/cdev.12473
77. Zantinge G, van Rijn S, Stockmann L, Swaab H (2017) Psychophysiological responses to emotions of others in young children with autism spectrum disorders: correlates of social functioning. Autism Res 10(9):1499–1509. https://doi.org/10.1002/aur.1794
78. Hanley M, McPhillips M, Mulhern G, Riby DM (2013) Spontaneous attention to faces in Asperger syndrome using ecologically valid static stimuli. Autism 17(6):754–761. https://doi.org/10.1177/1362361312456746
79. Nystrom P, Thorup E, Bolte S, Falck-Ytter T (2019) Joint attention in infancy and the emergence of autism. Biol Psychiat 86(8):631–638. https://doi.org/10.1016/j.biopsych.2019.05.006
80. Bandini LG, Anderson SE, Curtin C, Cermak S, Evans EW, Scampini R et al (2010) Food selectivity in children with autism spectrum disorders and typically developing children. J Pediatr 157(2):259–264. https://doi.org/10.1016/j.jpeds.2010.02.013
81. Schreck KA, Williams K, Smith AF (2004) A comparison of eating behaviors between children with and without autism. J Autism Dev Disord 34(4):433–438. https://doi.org/10.1023/b:jadd.0000037419.78531.86
82. Ahearn WH, Castine T, Nault K, Green G (2001) An assessment of food acceptance in children with autism or pervasive developmental disorder-not otherwise specified. J Autism Dev Disord 31(5):505–511. https://doi.org/10.1023/a:1012221026124
83. American Psychiatric Association, D. S., American Psychiatric Association (2013) Diagnostic and statistical manual of mental disorders: DSM-5, vol 5. American Psychiatric Association, Washington, DC. https://www.amberton.edu/media/Syllabi/Spring%202022/Graduate/CSL6798_E1.pdf. Accessed 7 Apr 2022
84. Chonchaiya W, Nuntnarumit P, Pruksananonda C (2011) Comparison of television viewing between children with autism spectrum disorder and controls. Acta Paediatr 100(7):1033–1037. https://doi.org/10.1111/j.1651-2227.2011.02166.x
85. Kheir NM, Ghoneim OM, Sandridge AL, Hayder SA, Al-Ismail MS, Al-Rawi F (2012) Concerns and considerations among caregivers of a child with autism in Qatar. BMC Res Notes 5:290–290. https://doi.org/10.1186/1756-0500-5-290
86. Orsmond GI, Kuo H-Y (2011) The daily lives of adolescents with an autism spectrum disorder: discretionary time use and activity partners. Autism 15(5):579–599. https://doi.org/10.1177/1362361310386503
87. Elbattah M, Loughnane C, Guérin J-L, Carette R, Cilia F, Dequen G (2021) Variational autoencoder for image-based augmentation of eye-tracking data. J Imaging 7(5):83. https://doi.org/10.3390/jimaging7050083
88. Carette R, Elbattah M, Cilia F, Dequen G, Guérin J-L, Bosche J (2019) Learning to predict autism spectrum disorder based on the visual patterns of eye-tracking scanpaths. In: HEALTHINF, pp 103–112. https://doi.org/10.5220/0007402601030112

Authors and Affiliations

Jin Xie[1,2] · Longfei Wang[1] · Paula Webster[5] · Yang Yao[1] · Jiayao Sun[1,2] · Shuo Wang[4] · Huihui Zhou[1,3]

Jin Xie
jin.xie@siat.ac.cn

Longfei Wang
lf.wang@siat.ac.cn

Paula Webster
Paulajmwebster@gmail.com

Yang Yao
yaoyang720@qq.com

Jiayao Sun
jy.sun@siat.ac.cn

[1] The Brain Cognition and Brain Disease Institute, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen-Hong Kong Institute of Brain Science-Shenzhen Fundamental Research Institutions, Shenzhen 518055, China

[2] University of Chinese Academy of Sciences, Beijing 100049, China

[3] The Research Center for Artificial Intelligence, Peng Cheng Laboratory, No. 2 Xingke First Street, Nanshan District, Shenzhen 518000, China

[4] Department of Radiology, Washington University in St. Louis, St. Louis, MO 63130, USA

[5] Department of Chemical and Biomedical Engineering and Rockefeller Neuroscience Institute, West Virginia University, Morgantown, WV 26506, USA