Deep Learning to Interpret Autism Spectrum Disorder Behind the Camera

Shi Chen† Ming Jiang Qi Zhao

Department of Computer Science and Engineering,
University of Minnesota
{chen4595, mjiang, qzhao}@umn.edu

***Abstract*—There is growing interest in understanding the visual behavioral patterns of individuals with Autism Spectrum Disorder (ASD) based on their attentional preferences. Attention reveals the cognitive or perceptual variation in ASD, and can serve as a biomarker to assist diagnosis and intervention. The development of machine learning methods for attention-based ASD screening shows promises, yet it has been limited by the need for high-precision eye trackers, the scope of stimuli, and black-box neural networks, making it impractical for real-life clinical scenarios. This study proposes an interpretable and generalizable framework for quantifying atypical attention in people with ASD. Our framework utilizes photos taken by participants with standard cameras, to enable practical and flexible deployment in resource-constrained regions. With an emphasis on interpretability and trustworthiness, our method automates human-like diagnostic reasoning, associates photos with semantically plausible attention patterns, and provides clinical evidence to support ASD experts. We further evaluate models on both in-domain and out-of-domain data, and demonstrate that our approach accurately classifies individuals with ASD and generalizes across different domains. The proposed method offers an innovative, reliable, and cost-effective tool to assist the diagnostic procedure, which can be an important effort toward transforming clinical research in ASD screening with artificial intelligence systems. Our code is publicly available at https://github.com/szzexpoi/proto_asd.**

***Index Terms*—Autism Spectrum Disorder, Visual Attention, Deep Neural Networks, Interpretable Model**

I. Introduction

Autism Spectrum Disorder (ASD) is a complex and heritable neurodevelopmental disorder with a global prevalence, affecting approximately one in 54 children in the United States [1], [2]. Timely diagnosis and intervention are recognized as the most effective clinical route to ASD treatment [3], [4], [5]. However, the current diagnostic process, which relies on clinical experts and subjective assessments through questionnaires and interviews, often results in significant delays in accessing care. In this context, our research explores the potential of deep learning to enhance the interpretability and generalizability of computer-aided ASD assessment, all of which can play key roles in enhancing the efficiency and effectiveness of ASD screening.

A collection of distinct attention patterns of people with autism have been documented, including altered attention to social stimuli (*e.g.*, facial expressions and social interactions), preference for nonsocial objects (*e.g.*, patterns, electronics, and tools) [6], [7], [8], [9], [10], [11], as well as challenges in attentional disengagement and oculomotor control [12]. Based on these findings, advanced machine learning techniques, such as deep neural networks (DNNs) [13], [14], [15], [16], [17], [18], [19], [20] have been developed to detect ASD traits from people's attentional preferences. In particular, these methods analyze eye-tracking data of subjects observing synthetic [6], [21], [22] or naturalistic [12], [23], [24], [25] stimuli and classify people with ASD based on the hierarchical features learned with DNNs. Despite their promising results, DNN-based methods typically tackle the ASD classification task by learning direct and implicit mappings from input eye-tracking data to ASD labels. Using these black-box DNNs as a diagnostic prediction mechanism is discouraged due to their lack of interpretability [26], [27]. In practice, ASD experts prefer explicit representations from machines that they can perceive and comprehend to understand the rationales behind their decisions. In addition, existing methods rely on data acquired from high-precision eye trackers, and are only tested with limited stimuli from the same domain as training data. How they would generalize toward real-world clinical settings is an open question. Therefore, the practical deployment of attention-based ASD classification remains a substantial challenge.

Our research aims to increase the acceptance of DNN-based ASD assessment in clinical settings and align machine-assisted diagnoses with those made by human experts. We fill the research gap with an interpretable and generalizable deep learning framework. The proposed method acquires people's attentional preferences from photos they take in daily environments using general cameras, which lifts the constraints on expensive instruments and extensive diagnosis [17], [28], [29], [30], [31], [32], [33]. The ASD classification is powered by DNNs optimized to detect fine-grained visual behaviors that differentiate people with ASD and healthy controls. Instead of only predicting the probabilities of ASD, we represent the implicit DNN features with prototypical attention patterns, which are interpretable to human experts.

Specifically, given a set of photos taken by participants (Fig. 1A, Step 1), our method associates their DNN features with prototypes that represent discriminative patterns of people's attentional preferences (Fig. 1A, Step 2). The correspondence between the photos and the prototypes helps to understand why the DNN made certain decisions by identifying

†: Work done when studying as a graduate student at the University of Minnesota.

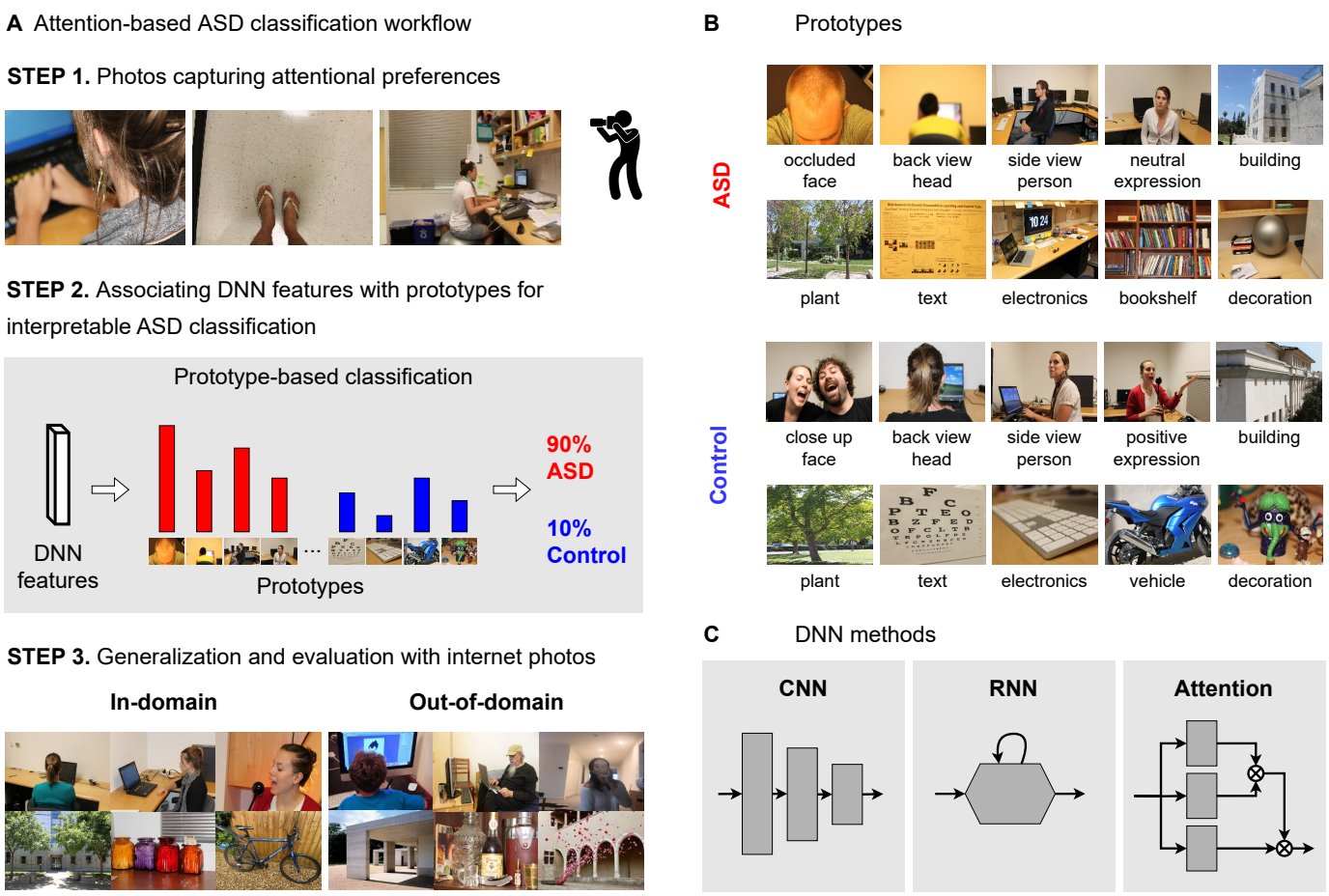

Fig. 1: **A.** Our ASD classification workflow. Participants use general cameras to freely take photos, where the photos encode their attentional preferences. Given photos taken by a participant, deep neural networks developed under our prototypical framework are utilized to associate the photos with semantically plausible prototypes. The prototypes represent discriminative attention patterns and characterize the visual behaviors of the participant with an interpretable interface. To estimate DNNs' generalizability across diverse scenarios, data from different domains are used to provide a comprehensive evaluation. **B.** Visualization of prototypes and their semantic labels. Different prototypes from the two classes (i.e., ASD and Control) may share the same semantic labels but with diverse appearances. **C.** DNNs included in our prototypical ASD classification method.

the areas of focus for people with ASD and those without the condition (Fig. 1B). Our method provides direct evidence for human experts to review and examine visually, with the goal of increasing the acceptance of DNN-based ASD classification in clinical settings and aligning the diagnoses made by human experts with machine assistance. Apart from the enhanced accuracy and interpretability, we also go beyond in-domain evaluation and generalize DNNs to handle data acquired from broader domains (Fig. 1A, Step 3). The inclusion of diverse types of data lays the foundation for developing more general and practical tools for ASD classification. These components work together to enable the development and evaluation of interpretable and generalizable ASD classifiers, and they are applicable to various architectural designs (*i.e.,* CNN, RNN, Attention model shown in Fig. 1C).

In summary, this paper carries out five major contributions:

1) We propose a novel deep learning framework that consists of a prototype-based method, three DNN models, and a data collection paradigm for interpretable and generalizable ASD screening.
2) By matching input photos with interpretable semantic prototypes, our method develops models that quantitatively measure the contributions of different prototypes and naturally explain their reasoning behind the classification results.
3) Based on our proposed semantic prototypes and conventional DNN architectures (*i.e.*, convolutional neural network (CNN), recurrent neural network (RNN), and Transformer), we design three ASD classifiers that outperform the previous methods with increased interpretability.
4) We for the first time study the model generalizability to different environments by experimenting with both in-domain and out-of-domain data.
5) Our extensive experiments demonstrate the effectiveness, interpretability, and generalizability of the proposed work.

II. Related Works

Our paper is related to studies about atypical attention in ASD and learning-based ASD classification.

Atypical Attention in ASD. A body of research has studied the atypical attention patterns in ASD and shows that individuals with ASD have reduced attention towards social stimuli (*e.g.,* faces and hand gestures) but more attention to nonsocial objects (*e.g.,* gadgets, electronics, and devices) [6], [7], [8], [9], [10]. These atypical attention patterns are evident in early infancy [34], young children [9], and adolescents [35]. Eye-tracking experiments with naturalistic stimuli reveal fine-grained differences in visual behaviors between individuals with ASD and healthy controls [12], [23], [24], [25], [28], [36], [37], [38]. For instance, individuals with ASD show reduced attention to faces [39], [40] in tasks involving social stimuli, and have a diminished ability to attend spontaneously to people and their activities [25]. Image viewing experiments [12] show that the impairment of attention in ASD can be modeled and interpreted with different levels of visual semantics [6], [22], [24], [41]. The atypical attention in ASD is also studied with attention data collected from a photo-taking experiment [16], [28].

Learning-Based ASD Classification. A series of studies propose machine learning methods for automatic and objective ASD assessment. Early studies typically apply simplified linear models on handcrafted features, including quotient-based ASD diagnostic tools [42], acoustic data of early language [43], magnetic resonance imaging data [44], and kinematic data [45], [46]. Lately, there is a trending research interest in classifying individuals with ASD and healthy controls with eye-tracking. Duan *et al.* [47] constructs a public-available eye-tracking dataset with attention annotations collected from children with autism. Several studies [48], [49] model the gaze patterns based on the distributions of eye fixations in different facial regions to characterize children with ASD. Jiang *et al.* [50] analyze eye-tracking data collected in a facial emotion recognition task to measure the social responsiveness of individuals with ASD. Jiang and Zhao [13] learn discriminative features of attention and uses a support vector machine (SVM) to classify individuals with or without ASD. Several works [14], [15] also take into account the temporal dynamics in attention deployment for ASD classification. Instead of characterizing attention patterns with high-end eye-tracker, two recent studies [16], [17] propose to classify photos taken by individuals with ASD [28], which reveals their visual preferences from the first-person perspective. Our study is built on top of this new paradigm to increase the accessibility of ASD classification models in real-world clinical applications.

Despite achieving promising performance, the aforementioned attention-based ASD classification methods mainly rely on black-box models that implicitly characterize attention patterns. Drawing inspiration from previous studies on concept-based explanations [51] and prototypical models [52], our work differentiates itself from existing methods by explicitly representing attention patterns in a photo-taking experiment with a prototype-based method designed with knowledge about ASD [12]. It helps understand models' reasoning with quantitative measurements, which promotes the trust between clinical experts and machines and enables generalization to data from broader environments.

III. Methods

In this work, we identify effectiveness, interpretability, and generalizability as critical needs of real-world clinical applications. The core of our method is a principled prototypical ASD classification paradigm for capturing atypical attention patterns with semantic prototypes for ASD classification (see Fig. 2). Firstly, it represents the attention patterns of both the ASD and control groups by employing a set of interpretable semantic prototypes. These prototypes serve as visual representations of specific semantic concepts. Secondly, it establishes a loss function based on prototype matching, which optimizes the alignment between input photos and semantic prototypes. This process associates photos with attention patterns that are focused on distinguishing semantics. Lastly, ASD classification is accomplished by examining the similarity features derived from photos and the corresponding matched prototypes. These components are integrated seamlessly with three DNN baselines to yield comprehensive diagnostic outcomes. We elaborate on the specifics of these components below.

Definition of semantic prototypes. The semantic prototypes are obtained by first clustering training photos based on their visual appearance and then assigning their semantic meanings based on the majority of photos in each cluster. To classify a photo with interpretability, it is important to obtain high-level semantic features of the photos. The features can be extracted from different layers of a CNN encoder. It is a common practice to use the average-pooled final convolutional layer features, as they focus more on semantic-level information instead of low-level details [53]. Upon obtaining the visual features, we use the K-means [54] clustering algorithm with the number of clusters equals 10 for each class (*i.e.*, 20 semantics prototypes in total). In our experiments, we find that such a setting achieves a good trade-off between the discriminative power and interpretability of prototypes, as it provides a reasonable Silhouette [55] score and can separate photos with fine-grained semantics. In practice, this setting may be adjusted for broader clinical applications. With the initial clusters, we then perform a further examination and cleaning by manually removing outliers, such as photos that are too close to an object (*e.g.*, a dark computer screen), or those with semantics different from the majority of photos in the same cluster, *e.g.*, photos with people smiling in the cluster for a neutral expression. Note that the manual cleaning of photos is optional for modeling. Finally, our semantic prototypes are computed by averaging the visual features of all photos in each cluster, and their corresponding meanings are determined based on the semantics of the majority of photos as well as the attention patterns for the ASD and control groups [16], [12], [28]. With the aforementioned paradigm, we can derive discriminative prototypes based on the similarity of semantics in images, where each prototype is represented with images correlated semantics.

Fig. 1B illustrates the semantic prototypes discovered by our approach. From the photos taken by the ASD group, we

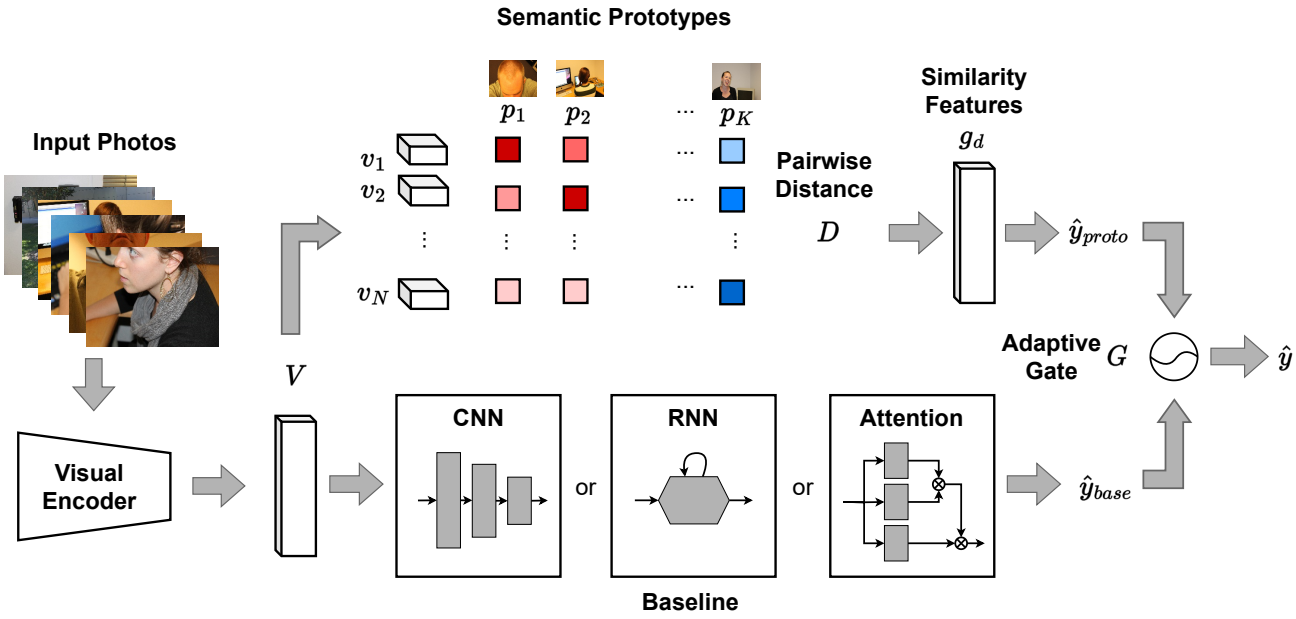

Fig. 2: Overview of the proposed deep learning methods for ASD classification.

find prototypes such as occluded faces or photos taken from odd angles, photos about people's backs or hair, and photos about electronics. From the photos taken by the Control group, we find prototypes including human faces or expressions (*e.g.*, people smiling or making funny faces, and people looking at the camera in different poses) and photos about salient objects (*e.g.*, artistic decorations). These findings align well with existing studies on the atypical attention patterns in ASD [16], [22], [28]. It is noteworthy that we discover similar prototypes in the two groups (*e.g.*, both groups take side view photos about people), yet fine-grained differences in attention patterns can be observed between groups (*e.g.*, people in side view photos taken by controls typically have stronger facial expressions or more eye contact with the camera).

Prototype matching loss. With the defined semantic prototypes, an input photo and a prototype can be matched by measuring their feature similarity. For example, given the features of the input photo, denoted as v, one can measure its minimal distance to a class of prototypes $P \in \mathbb{R}^{K^{'} \times 2048}$ ($K^{'}$ is the number of prototypes in the class (*i.e.*, prototypes for ASD or Control) as

$$d(v, P) = \min_{p_j \in P} \parallel v - p_j \parallel_2^2 . \tag{1}$$

By encouraging the alignment between features and prototypes of the same class the photo belongs to, we can learn more discriminative features and create an embedding space for DNNs to make decisions based on an individual's attentional preferences to different semantics. Drawing inspirations from [52], we define the prototype matching loss L_{proto} as

$$L_{proto} = \frac{1}{N} \sum_{i=1}^{N} d(v_i, P_i^+) - d(v_i, P_i^-) \tag{2}$$

where N is the number of input photos, v_i is the feature of the i-th photo, and P_i^+ and P_i^- are the semantic prototypes of the class (*i.e.*, ASD or Control) that the i-th photo is from.

ASD classification with semantic prototypes. With the proposed prototype matching loss that optimizes the matching between visual features and interpretable semantic prototypes, DNNs can reason about the attention patterns for ASD classification. As shown in the top block of Fig. 2, the prediction of classification probabilities can be obtained by taking into account the pairwise distance between the visual features for each photo and all prototypes:

$$D = \{d_{11}, d_{12}, ..., d_{NK}\}, d_{ij} = \parallel v_i - p_j \parallel_2^2 \tag{3}$$

$$g_d = \log \frac{D+1}{D+\epsilon} \tag{4}$$

where $D \in \mathbb{R}^{NK}$ contains pairwise distance d_{ij} between visual features $v_i \in \mathbb{R}^{2048}$ for i_{th} photos and j_{th} prototype p_j. The ϵ is a small value for preventing numerical errors. Following [52], the features $g_d \in \mathbb{R}^{NK}$ are designed to be monotonically decreasing with respect to the pairwise distance, and a large value $g_{d_{ij}}$ means the i_{th} input is close to the j_{th} prototype. They represent the similarity between photos and visual semantics related to the attention patterns of people with or without ASD.

Upon obtaining the features, the prediction output can be computed as

$$\hat{y}_{proto} = \sigma(W_g g_d) \tag{5}$$

where W_g is a trainable fully-connected layer and σ denotes the Sigmoid activation function.

DNNs for ASD classification. Our prototypical paradigm is general and can be flexibly added into various deep networks (*i.e.*, the bottom block in Fig. 2) to enhance their effectiveness, interpretability, and generalizability. We achieve this with an adaptive output fusion method that combines the prediction result of our prototypical classifier with a conventional DNN baseline classifier. We consider three DNNs as our baselines, including CNN [56] operated on a single photo, RNN [57] that

sequentially processes multiple photos, and an Attention model (the original Vision Transformer [58] fails to converge on the clinical data with smaller scale) that leverages self-attention mechanism [59] to model the pairwise relationship between a collection of photos:

- **CNN**: We directly use features V extracted from the visual encoder to predict the class label, which is computed as $\hat{y}_{base} = \sigma(W_{base}V)$ and W_{base} is the fully-connected layer.
- **RNN**: The RNN baseline considers multiple photos taken by the same individual, and sequentially processes features extracted from each photo v_i. The prediction is based on the last hidden state h (*i.e.*, $\hat{y}_{base} = \sigma(W_{base}h)$).
- **Attention model**: Unlike RNN which is restricted to sequential inputs, self-attention encodes multiple inputs without constraining their temporal order. Specifically, the corresponding model captures the pairwise relationship α between features of different photos:

$$\alpha = \text{softmax}(\frac{QR^\top}{\sqrt{C}}) \tag{6}$$

  Q and R are query and key derived based on the features $Q = W_qV$, $R = W_rV$, where W_q and W_r denote trainable layers. C is the embedding size. The prediction is computed by selectively considering features from different photos $\hat{y}_{base} = \sigma(W_{base}\alpha V)$.

To determine the final prediction, we integrate the interpretable and generalizable output of our prototypical classifier $\hat{y}_{proto}$ and the output of a selected DNN baseline $\hat{y}_{base}$. In particular, we leverage an adaptive gate that dynamically balances their contributions:

$$\hat{y} = G \cdot \hat{y}_{proto} + (1 - G) \cdot \hat{y}_{base} \tag{7}$$

where $G = \sigma(W_{gate}V)$ is the value of the gate and W_{gate} is a fully-connected layer. The method allows us to take advantage of the learning power of DNNs without compromising interpretability.

We train the DNNs with both binary cross-entropy loss $L_{cls} = -y\log(\hat{y}) - (1-y)\log(1-\hat{y})$ for classification, where y is the ground truth label, and the proposed prototype matching loss:

$$L = L_{cls} + \lambda \cdot \text{ReLU}(L_{proto} + \xi) \tag{8}$$

where λ and ξ are the scale factor and margin for the prototype matching loss. The ReLU activation controls the contributions of the prototype matching loss, *i.e.,* we do not impose further regularization if the prototypes are close enough to their corresponding classes.

IV. Experiment

A. Implementation

Model configuration. Each DNN model developed under our framework is trained end-to-end with the Adam optimizer [60], where the weight decay and gradient clipping are set to 10^{-5} and 0.1, respectively. The batch size is set to 12. We train the CNN that takes in a single photo for 30 epochs. For RNN and the Attention model, we consider a set of 14 photos randomly sampled from an individual's photo pool as a single sample, and train each DNN for 180 epochs to ensure an equivalent number of iterations on each sample. The learning rate is initialized as 10^{-4} and divided by 2 every 30 epochs. λ and ξ in Equation 8 are empirically set to 10^{-2} and 80, respectively. Please see the supplementary files for code implementation.

Experiment data. With an emphasis on ASD classification in real-world settings, our method is designed to generalize across broader domains with photos of different characteristics. We validate it on two sets of data: In-domain photos are from the dataset introduced in [17], [28] (see Fig. 1A). It contains 1672 photos taken by 22 participants with ASD and 23 healthy controls. All ASD participants meet the diagnostic criteria for different autism diagnostic protocols, including DSM-V/ICD-10 diagnostic criteria for autism spectrum disorder, Autism Diagnostic Observation Schedule2 (ADOS-2) [61], and Autism Diagnostic InterviewRevised (ADI-R) [62], [63], and have matched IQ with the healthy controls. The participants are provided with a camera and instructed to take photos freely in three different environments, including scenes involving people, indoor scenes, and outdoor scenes. They are told that they could keep any of the photos and also have the option to delete any photos they had taken. Please refer to [28] for additional details on data collection. Moreover, to test the generalizability of ASD classifiers, out-of-domain photos are collected with Google Image Search using the in-domain photos as the queries and assigned the same class label as the query images.

Evaluation. Following [13], [16], [17], [48], we perform a leave-one-subject-out cross-validation to train and evaluate DNNs. They are evaluated with four popular evaluation metrics widely used in clinical assessments, *i.e.*, accuracy (Acc.), sensitivity (Sen.), specificity (Spe.), and area Under the ROC Curve (AUC). By training and evaluating one classifier for each holdout subject, the cross-validation returns an almost unbiased estimate of the probability of error [64].

B. Quantitative Results

We demonstrate the effectiveness of our models (*i.e.*, CNN-Proto, RNN-Proto, and Attention-Proto) by comparing them with their corresponding baselines, previous computational methods [16], [17], as well as human experts [28].

Results in Table I show that our method achieves better accuracy than all compared state-of-the-art approaches. For example, CNN-proto achieves an 87% accuracy and 94% AUC (area under the ROC curve, see Fig. 3A) on in-domain photos, outperforming human experts and existing models by $10\% - 30\%$ absolute gains in performance. With our prototypical method, all three DNNs acquired significant and consistent performance gains. Further, when validated on out-of-domain photos, our prototypical methods also demonstrate better generalizability. When considering photos from different domains, conventional DNNs suffer from an over 10% drop in performance (*e.g.*, accuracy and AUC) on the out-of-domain photos. The large performance gap suggests that implicitly learning the mapping between the input photos and output

TABLE I: Quantitative results of the human experts, previous ASD classifiers, conventional DNN baselines, and the proposed prototypical DNNs. Four evaluation metrics are used, including accuracy (Acc.), sensitivity (Sen.), specificity (Spe.), and AUC. Best results are highlighted in bold text.

	In-domain				Out-of-domain			
	Acc.	Sen.	Spe.	AUC	Acc.	Sen.	Spe.	AUC
Experts [28]	0.65	-	-	0.60	-	-	-	-
Ruan *et al.* [16]	0.59	-	-	0.64	-	-	-	-
Chen *et al.* [17]	0.76	0.77	0.74	0.82	-	-	-	-
CNN	0.76	0.68	0.83	0.81	0.69	0.53	0.83	0.71
RNN	0.73	0.59	0.87	0.81	0.71	0.55	0.86	0.77
Attention	0.78	0.73	0.83	0.82	0.69	0.55	0.83	0.72
CNN-Proto	**0.87**	**0.77**	**0.96**	**0.95**	**0.87**	**0.82**	**0.91**	**0.95**
RNN-Proto	0.80	0.73	0.87	0.85	0.78	0.68	0.87	0.85
Attention-Proto	0.82	0.77	0.83	0.90	0.72	0.78	0.73	0.83

classification labels does not generalize well to data from different sources. Differently, our prototypical DNNs (*e.g.*, CNN-Proto, and RNN-Proto) achieve similar performances across photos from different domains, despite the discrepancies between training and test data. Among the three different model architectures, CNN-Proto shows the best results. It suggests that, for freely taken photos without a specific order, a model operated independently on each image can better capture the differentiating characteristics of autism. Nevertheless, methods with sequential modeling may have the potential to prevail in data with temporal structure (e.g., RNNs working on videos) or with the integration of pretrained foundation models (*e.g.,* large-scale vision models [65], [66], [67] with attention mechanism).

These observations demonstrate the usefulness of our prototypical method for improving the generalizability of ASD classifiers. By projecting raw photos onto discriminative attention patterns (*i.e.,* the prototypes), it is less prone to overfitting to smaller datasets and has stronger applicability to real-world scenarios. Our method is also lightweight and practical for deployment, *e.g.,* with a negligible 0.1% increase in parameters for the CNN baseline.

C. Interpreting DNN Models' Decision-making Process

A key advantage of the proposed method resides in its ability to interpret the decision-making process. We demonstrate the effectiveness of our method in understanding the rationales behind decisions, which plays an important role in promoting the trustworthiness of the computational models.

We first interpret a DNN's prediction by quantifying the contribution of each semantic prototype to the classification. The contribution is measured either with the predicted classification probability of an individual photo (CNN) or based on the gradient-based importance [68] of each photo when multiple photos are used together (RNN or Attention). In particular, the contribution of a prototype is determined by computing the average contribution of photos with the nearest distance to the prototype in the feature space. The aforementioned method enables the discovery of discriminative patterns that characterize the attentional preferences of photographers. As shown in Fig. 3B, prototypical DNNs can identify a diverse set of distinguishable attentional preferences for ASD classification: For people with ASD, they consider prototypes related to social deficits (*e.g.*, occluded or expressionless face, photos taken from the back) and non-social objects (*e.g.,* devices, electronics, and other inanimate objects) to be important; For healthy controls, all prototypical DNNs consider objects (*i.e.*, decoration, plant, text) to be important. RNN-Proto and Attention-Proto also take into account positive expressions (*e.g.,* smiling faces or faces with natural eye contact) when classifying controls (prototypes selected 17.2% and 14.1% of the times by the two DNNs, respectively). These findings are consistent with the previous analyses on impaired attention in ASD [6], [7], [9], [10], [12], [22], suggesting that our prototypical DNNs capture the contributing factors for attention-based ASD classification. They also show that attention towards non-social cues (such as decorations, plants, and text) can be just as important as previously identified attention patterns related to social deficits. The developed data-driven and interpretable methods report a comprehensive range of social and non-social preferences in the ecologically-relevant first-person perspective.

Besides identifying the key attention patterns for classifying people with ASD and controls, our method also allows visualization of the regions of interest (ROIs) associated with the discovery of patterns. Such a unique feature provides opportunities for human experts to make second opinions based on not only which discriminative attention patterns are identified but also where they are discovered. In Fig. 4, we use CNN-Proto and CNN as examples and visualize their ROIs together with the prototypes activated for classification. For more accurate visualization and higher computational efficiency, we localize the ROIs with Grad-CAM [68], which directly measures the relative contributions of each region on predicting different classes with a single inference process (*i.e.,* in our case the classification of ASD or health control) and shows

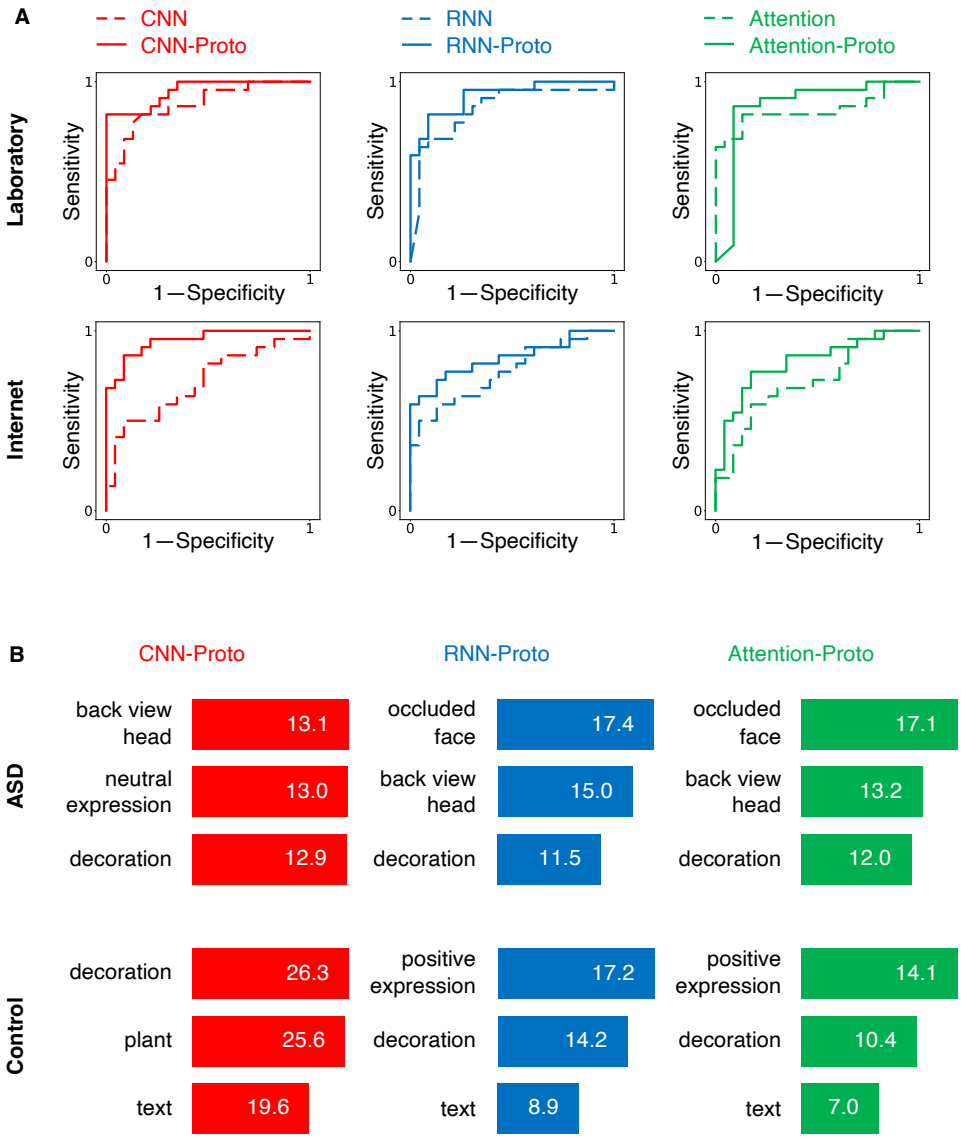

Fig. 3: **A.** The ROC curves represent the DNNs' classification performance. Each point on the curve denotes a true positive rate (sensitivity) and the false positive rate (1−specificity) of the DNNs. **B.** The most important semantic prototypes and their contributions to ASD classification. Bars indicate the average contribution weights of the semantic prototypes.

promise in various applications. The prototypes are selected based on their pairwise distance with observations (*i.e.*, D in Equation 3). Results show that CNN-Proto accurately captures the ROIs for distinct attention patterns (*e.g.,* occluded faces for ASD in Fig. 4A, smiling faces for control in Fig. 4H), whereas CNN has widespread (*e.g.,* Fig. 4A, F, K) or random (*e.g.,* Fig. 4C, E, I) focuses. The highlighted ROIs also align with the matched prototypes. By jointly considering the localized ROIs and the attention patterns encoded in matched prototypes, our method has the potential to be integrated into real-world clinical procedures to enhance their effectiveness and efficiency.

D. Ablation Study on Model Design

To obtain a comprehensive view of our model design, in this section, we perform ablation experiments on three key components.

First, to validate the usefulness of semantic prototypes (see Fig. 1B), we compare our method with three types of prototypes: (1) A naive baseline (*i.e.*, -random) that initializes prototypes with random features, and optimizes the prototypes together with other model parameters; (2) Our proposed prototypes determined with automatic clustering but without involving manual cleaning (*i.e.*, -w/o cleaning); and (3) Our full method (*i.e.,* -Proto). As shown in Table II, the randomly initialized prototypes achieve reasonable performance on in-domain photos through training with the domain-specific data, but fail to generalize well to the out-of-domain data with the absence of semantics prototypes. It also lacks the capability to elucidate the rationales of DNNs, as the randomly initialized prototypes are not associated with clear interpretation. Moreover, discarding the optional cleaning step results in comparable performance as our full method, which demonstrates the generalization of our approach without manual efforts.

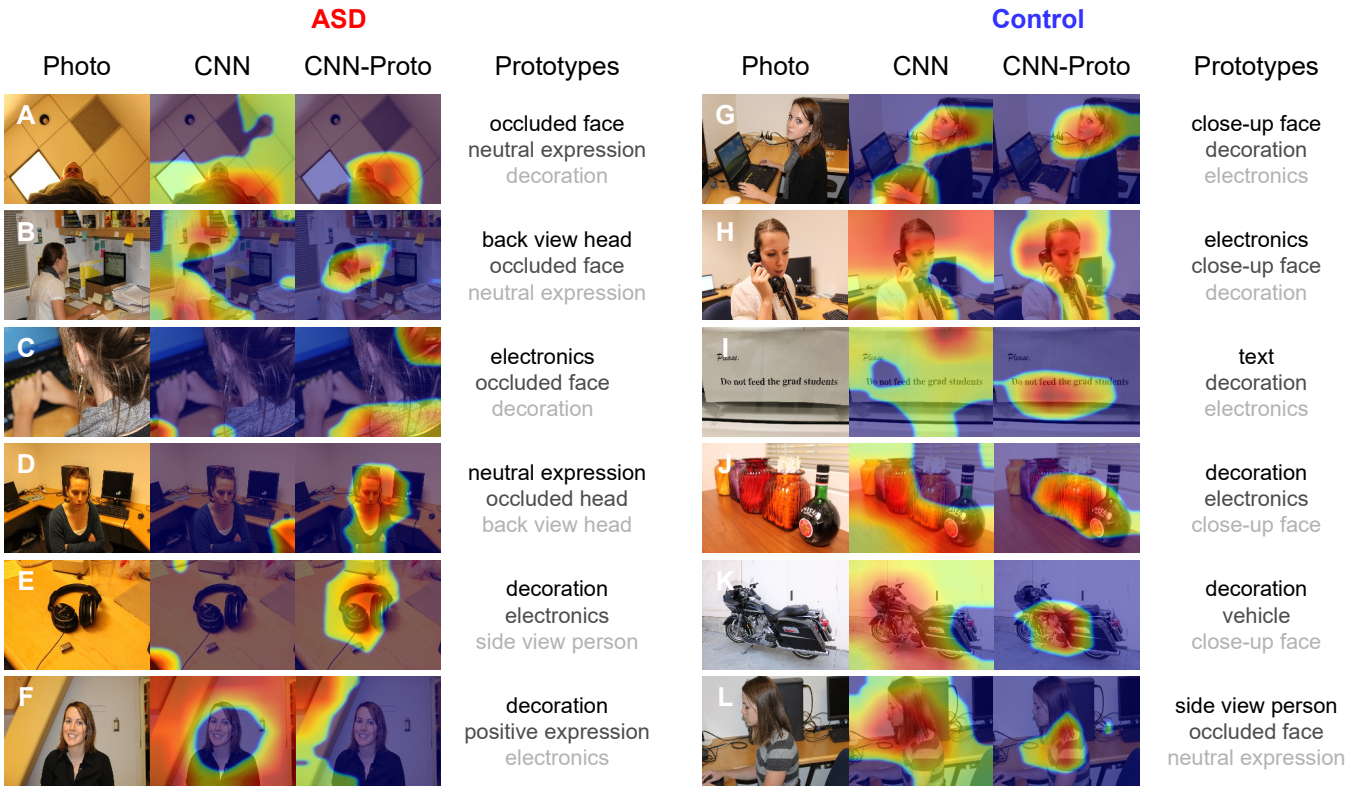

Fig. 4: Visualization of CNNs' attended regions of interest (heatmaps, CNN and CNN-Proto), and top-3 nearest semantic prototypes (text, CNN-Proto).

TABLE II: Comparison between our prototypes and their alternatives with random initialization (-random) and those without manual cleaning (-w/o cleaning). Best results for each baseline are highlighted in bold text.

	In-domain		Out-of-domain	
	Acc.	AUC	Acc.	AUC
CNN-Proto	0.87	0.94	**0.87**	**0.95**
-random	**0.89**	**0.95**	0.76	0.89
-w/o cleaning	0.89	0.91	0.85	0.91
RNN-Proto	**0.80**	0.85	**0.78**	0.85
-random	0.80	0.82	0.73	0.71
-w/o cleaning	0.80	**0.87**	0.73	**0.86**
Attention-Proto	**0.80**	**0.90**	**0.78**	**0.84**
-random	0.76	0.81	0.71	0.73
-w/o cleaning	0.78	0.83	0.78	0.80

TABLE III: Ablation results for different components. Best results for each baseline are highlighted in bold text.

	In-domain		Out-of-domain	
	Acc.	AUC	Acc.	AUC
CNN-Proto	**0.87**	**0.94**	**0.87**	**0.95**
w/ only matching	0.84	0.92	0.84	0.91
w/ only proto-cls	0.84	0.91	0.84	0.87
RNN-Proto	**0.80**	**0.85**	**0.78**	**0.85**
w/ only matching	0.80	0.82	0.76	0.83
w/ only proto-cls	0.76	0.81	0.78	0.81
Attention-Proto	**0.80**	**0.90**	**0.78**	**0.84**
w/ only matching	0.80	0.83	0.73	0.80
w/ only proto-cls	0.76	0.82	0.71	0.77

Second, we validate the contribution of different architectural designs. Our prototypical DNNs leverage two key designs to support reasoning with prototypes, *i.e.,* the prototype matching loss and the prototypical classifier. Table III reports results on DNNs with either design. Results show that dropping either component leads to a visible loss of performance, which highlights the integral design of our method. Compared to the prototypical classifier, the prototype matching loss tends to have a larger impact on the AUC scores, suggesting the higher importance of explicitly matching the visual features with interpretable prototypes.

Third, we investigate the impacts of data scarcity on our model. In particular, we study the effects of different numbers of evaluation photos on classification performance. While our full method takes into all photos (*i.e.,* around 50 photos) taken by a participant for autism screening, as reported in Figure 5, it is relatively robust to limited data and can achieve competitive

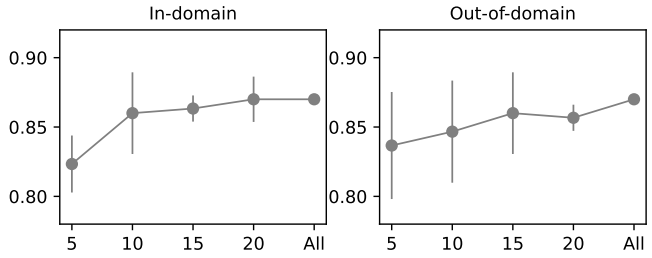

Fig. 5: Classification accuracy for CNN-Proto using different numbers of photos for evaluation on in-domain and out-of-domain data. The photos are uniformly sampled, and each experiment is repeated three times to derive the mean and standard deviation of scores.

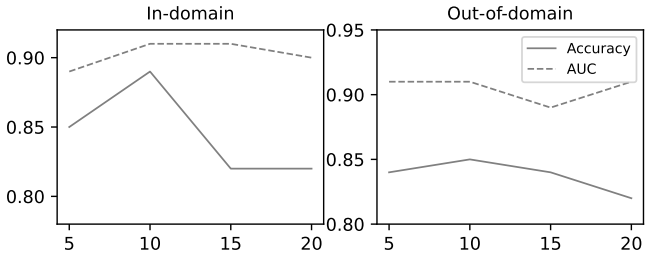

Fig. 6: Classification accuracy for CNN-Proto with different numbers of prototypes. Note that the prototypes are automatically derived without manual cleaning.

TABLE IV: Comparative results of the proposed method using different settings of hyperparameters. The best results for each baseline are highlighted in bold text.

	In-domain		Out-of-domain	
	Acc.	AUC	Acc.	AUC
$\lambda = 0.001$	0.84	0.91	0.87	0.91
$\lambda = 0.005$	0.82	0.90	0.8	0.89
$\lambda = 0.01$	**0.87**	**0.94**	**0.87**	**0.95**
$\lambda = 0.02$	0.84	0.90	0.76	0.81
$\lambda = 0.05$	0.84	0.92	0.78	0.82
$\xi = 0$	0.82	0.92	0.87	0.92
$\xi = 40$	0.82	0.89	0.84	0.90
$\xi = 80$	**0.87**	**0.94**	**0.87**	**0.95**
$\xi = 160$	0.87	0.94	0.84	0.92
$\xi = 500$	0.82	0.86	0.82	0.85

performance (*e.g.,* outperforming all baselines) with a minimum of 5 photos.

E. Ablation Study on Hyperparameters

Two hyperparameters λ and ξ are used in our objective function (Equation 8) to adjust the contribution of the proposed prototype matching loss L_{proto}. Specifically, λ adjusts the relative weights of the prototype matching loss. ξ encourages the input photos to lie close/distant enough to the correct/incorrect semantic prototypes, and disregards easy negatives (*e.g.,* incorrect prototypes far from the input photos). To study the effects of these hyperparameters, we train and evaluate our best model (*i.e.*, CNN-Proto) under different hyperparameter settings. We change the value of each hyperparameter at a time, and keep the other one fixed. As reported in Table IV, setting $\lambda = 0.01$ and $\xi = 80$ provides the best performance on both datasets. Lower values of λ lead to inferior performance, suggesting the effectiveness of matching input photos with semantic prototypes of the correct classes. On the other hand, very large values of λ hurt the performance, which is likely caused by overfitting the training data. The effects of ξ are also significant. To achieve higher performance, it is important to select a reasonable value of ξ (neither too high nor too low) so that models can pay attention to hard negatives during optimization and align input photos with the correct prototypes.

Next, we study the effects of different numbers of prototypes on the model performance. A smaller number of prototypes can increase the difficulty of interpretation due to the diverse semantics in images, while a larger number of prototypes makes it more challenging for the model to correlate ASD with a variety of semantics. As shown in Figure 6, while the proposed method is relatively robust to the choice of the number of prototypes, selecting 10 prototypes per category (*i.e.,* ASD or control) achieves a reasonable balance between interpretability (*i.e.,* Section IV-C) and performance.

V. Conclusion

This paper presents an interpretable, and generalizable method for ASD classification based on freely taken photos. Through the use of prototypes encoding attentional preferences and highlighting important regions, our method offers explainable predictions, enabling human experts to review the results. This enhances objectivity and trustworthiness in ASD classification. Extensive validation on photos captured in diverse environments demonstrates that our method surpasses human experts and existing computational methods. It also provides insights into the decision-making process of DNNs. Our method also proves robust in different data domains, making it highly applicable to real-world clinical settings.

Our work serves as a solid step toward connecting research in cognitive development and artificial intelligence, and can generate significant societal impacts from two distinct perspectives: First, our framework provides a practical solution for regions with scarce clinical resources, enabling the democratization of access and timely assistance without reliance on specialized clinicians or high-end diagnostic instruments. Second, by elucidating the evidence used for model decisions, our approach fosters trust in computer-aided systems and aids clinical experts in achieving objective diagnoses. Third, our emphasis on generalizability to real-world scenarios contributes to the development of ASD classification systems for practical clinical applications.

Our work also has room for improvement. While our proposed method exhibits effectiveness, the current focus

on ASD necessitates further research on other types of neurodevelopmental disorders (NDDs) such as Attention Deficit Hyperactivity Disorder and Obsessive Compulsive Disorder. The proposed prototype-based method offers a solid and principled tool for studying these NDDs. With the high degree of heterogeneity and comorbidity among them, we envision that the proposed method has the potential to provide valuable insights into understanding and addressing the complexities associated with these conditions. In addition, as we partially demonstrated with the out-of-domain evaluation, exploration of data from broader scenarios, including participants with diverse backgrounds (*e.g.,* cultures and regions) and data collected in different studies, would play a critical role in facilitating the utilization of AIs in clinical applications. It would also be an important direction to automatically discover prototypes in these scenarios, *e.g.,* starting with a large set of potential behavioral patterns as prototypes and encouraging the model to adaptively select important ones for classification.

REFERENCES

[1] M. J. Maenner, K. A. Shaw, J. Baio, and et al., "Prevalence of autism spectrum disorder among children aged 8 years — autism and developmental disabilities monitoring network, 11 sites, united states, 2016." *MMWR Surveill Summ 2020*, vol. 69 (4):1–12, 2020.

[2] C. Lord, "How common is autism?" *Nature*, vol. 474, no. 7350, pp. 166–167, Jun 2011.

[3] J. Bradshaw, A. M. Steiner, G. Gengoux, and L. K. Koegel, "Feasibility and effectiveness of very early intervention for infants at-risk for autism spectrum disorder: A systematic review," *Journal of Autism and Developmental Disorders*, vol. 45, no. 3, pp. 778–794, 2015.

[4] R. Landa, "Diagnosis of autism spectrum disorders in the first 3 years of life," *Nature Clinical Practice Neurology*, vol. 4, no. 3, pp. 138–147, 2008.

[5] Z. Wang, J. Liu, W. Zhang, W. Nie, and H. Liu, "Diagnosis and intervention for children with autism spectrum disorder: A survey," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 3, pp. 819–832, 2022.

[6] G. Dawson, A. N. Meltzoff, J. Osterling, J. Rinaldi, and E. Brown, "Children with autism fail to orient to naturally occurring social stimuli," *Journal of Autism and Developmental Disorders*, vol. 28, no. 6, pp. 479–485, 1998.

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

[8] G. Dawson, S. J. Webb, and J. McPartland, "Understanding the nature of face processing impairment in autism: insights from behavioral and electrophysiological studies," *Developmental neuropsychology*, vol. 27, no. 6, pp. 403–424, 2005.

[9] N. J. Sasson, J. T. Elison, L. M. Turner-Brown, G. S. Dichter, and J. W. Bodfish, "Brief report: Circumscribed attention in young children with autism," *Journal of Autism and Developmental Disorders*, vol. 41, no. 2, pp. 242–247, 2011.

[10] J. W. Tanaka and A. Sung, "The "eye avoidance" hypothesis of autism face processing," *Journal of Autism and Developmental Disorders*, vol. 46, no. 5, pp. 1538–1552, 2016.

[11] K. M. Dalton, B. M. Nacewicz, T. Johnstone, H. S. Schaefer, M. A. Gernsbacher, H. H. Goldsmith, A. L. Alexander, and R. J. Davidson, "Gaze fixation and the neural circuitry of face processing in autism," *Nature Neuroscience*, vol. 8, no. 4, pp. 519–526, 2005.

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

[13] M. Jiang and Q. Zhao, "Learning visual attention to identify people with autism spectrum disorder," in *2017 IEEE International Conference on Computer Vision*, 2017, pp. 3287–3296.

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

[15] Y. Fang, H. Duan, F. Shi, X. Min, and G. Zhai, "Identifying children with autism spectrum disorder based on gaze-following," in *2020 IEEE International Conference on Image Processing*, 2020, pp. 423–427.

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

[17] S. Chen and Q. Zhao, "Attention-based autism spectrum disorder screening with privileged modality," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 1181–1190.

[18] H. Egger, G. Dawson, J. Hashemi, K. Carpenter, S. Espinosa, K. Campbell, S. Brotkin, J. Schaich-Borg, Q. Qiu, M. Tepper, J. Baker, R. Bloomfield, and G. Sapiro, "Automatic emotion and attention analysis of young children at home: a researchkit autism feasibility study," *npj Digital Medicine*, vol. 1, 12 2018.

[19] G. Tan, K. Xu, J. Liu, and H. Liu, "A trend on autism spectrum disorder research: Eye tracking-eeg correlative analytics," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 3, pp. 1232–1244, 2022.

[20] V. Jyoti, S. Gupta, and U. Lahiri, "Understanding the role of objects in joint attention task framework for children with autism," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 13, no. 3, pp. 524–534, 2021.

[21] A. Klin, D. J. Lin, P. Gorrindo, G. Ramsay, and W. Jones, "Two-year-olds with autism orient to non-social contingencies rather than biological motion," *Nature*, vol. 459, pp. 257–261, 2009.

[22] S. Wang, J. Xu, M. Jiang, Q. Zhao, R. Hurlemann, and R. Adolphs, "Autism spectrum disorder, but not amygdala lesions, impairs social attention in visual search," *Neuropsychologia*, vol. 63, pp. 259–274, 2014.

[23] C. Ames and S. Fletcher-Watson, "A review of methods in the study of attention in autism," *Developmental Review*, vol. 30, no. 1, pp. 52–73, 2010.

[24] E. Birmingham, M. Cerf, and R. Adolphs, "Comparing social attention in autism and amygdala lesions: Effects of stimulus and task condition," *Social Neuroscience*, vol. 6, pp. 420 – 435, 2011.

[25] K. Chawarska, S. Macari, and F. Shic, "Decreased spontaneous attention to social scenes in 6-month-old infants later diagnosed with autism spectrum disorders," *Biological Psychiatry*, vol. 74, pp. 195–203, 2013.

[26] A. Holzinger, C. Biemann, C. Pattichis, and D. Kell, "What do we need to build explainable ai systems for the medical domain?" *ArXiv*, vol. abs/1712.09923, 2017.

[27] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." *Queue*, vol. 16, no. 3, p. 31–57, 2018.

[28] S. Wang, S. Fan, B. Chen, S. Habimi, L. K. Paul, Q. Zhao, and R. Adolphs, "Revealing the world of autism through the lens of a camera," *Current Biology*, vol. 26, no. 20, pp. 909–910, 2016.

[29] J. Megerian, S. Dey, R. Melmed, D. Coury, M. Lerner, C. Nicholls, K. Sohl, R. Rouhbakhsh, A. Narasimhan, J. Romain, S. Golla, S. Shareef, A. Ostrovsky, J. Shannon, C. Kraft, S. Liu-Mayo, H. Abbas, D. Gal-Szabo, D. Wall, and S. Taraman, "Evaluation of an artificial intelligence-based medical device for diagnosis of autism spectrum disorder," *npj Digital Medicine*, vol. 5, p. 57, 05 2022.

[30] H. Drimalla, T. Scheffer, N. Landwehr, I. Baskow, S. Roepke, B. Behnia, and I. Dziobek, "Towards the automatic detection of social biomarkers in autism spectrum disorder: introducing the simulated interaction task (sit)," *npj Digital Medicine*, vol. 3, no. 1, p. 25, 2020.

[31] J. S. Oliveira, F. O. Franco, M. C. Revers, A. F. Silva, J. Portolese, H. Brentani, A. Machado-Lima, and F. L. S. Nunes, "Computer-aided autism diagnosis based on visual attention models using eye tracking," *Scientific Reports*, vol. 11, no. 1, p. 10131, 2021.

[32] D. Wall, J. Kosmicki, T. Deluca, E. Harstad, and V. Fusaro, "Use of machine learning to shorten observation-based screening and diagnosis of autism," *Translational psychiatry*, vol. 2, p. e100, 02 2012.

[33] C. Küpper, S. Stroth, N. Wolff, F. Hauck, N. Kliewer, T. Schad-Hansjosten, I. Kamp-Becker, L. Poustka, V. Roessner, K. Schultebraucks, and S. Roepke, "Identifying predictive features of autism spectrum disorders in a clinical sample of adolescents and adults using machine learning," *Scientific Reports*, vol. 10, no. 1, p. 4805, 2020.

[34] J. Osterling and G. Dawson, "Early recognition of children with autism: A study of first birthday home videotapes," *Journal of Autism and Developmental Disorders*, vol. 24, no. 3, pp. 479–485, 1994.

[35] N. J. Sasson, L. M. Turner-Brown, T. N. Holtzclaw, K. S. Lam, and J. W. Bodfish, "Children with autism demonstrate circumscribed attention during passive viewing of complex social and nonsocial picture arrays," *Autism Research*, vol. 1, no. 1, pp. 31–42, 2008.

[36] L. Byrge, J. Dubois, J. M. Tyszka, R. Adolphs, and D. P. Kennedy, "Idiosyncratic brain activation patterns are associated with poor social comprehension in autism," *Journal of Neuroscience*, vol. 35, no. 14, pp. 5837–5850, 2015.
[37] K. Warnell, J. Moriuchi, W. Jones, and A. Klin, "Parsing heterogeneity in autism spectrum disorders: Visual scanning of dynamic social scenes in school-aged children," *Journal of the American Academy of Child and Adolescent Psychiatry*, vol. 51, pp. 238–48, 03 2012.
[38] A. Santos, T. Chaminade, D. da Fonseca, C. Silva, D. Rosset, and C. Deruelle, "Just another social scene: Evidence for decreased attention to negative social scenes in high-functioning autism," *Journal of Autism and Developmental Disorders*, vol. 42, pp. 1790–1798, 2012.
[39] M. Freeth, P. Chapman, D. Ropar, and P. Mitchell, "Do gaze cues in complex scenes capture and direct the attention of high functioning adolescents with asd? evidence from eye-tracking," *Journal of autism and developmental disorders*, vol. 40, pp. 534–47, 11 2009.
[40] A. Klin, W. Jones, R. Schultz, F. Volkmar, and D. Cohen, "Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism," *Archives of General Psychiatry*, vol. 59, no. 9, pp. 809–816, 2002.
[41] J. Swettenham, S. Baron-Cohen, T. Charman, A. Cox, G. Baird, A. Drew, L. Rees, and S. Wheelwright, "The frequency and distribution of spontaneous attention shifts between social and nonsocial stimuli in autistic, typically developing, and nonautistic developmentally delayed infants," *Journal of Child Psychology and Psychiatry*, vol. 39, no. 5, pp. 747–753, 1998.
[42] D. Bone, M. Goodwin, M. P. Black, C.-C. Lee, K. Audhkhasi, and S. S. Narayanan, "Applying machine learning to facilitate autism diagnostics: Pitfalls and promises," *Journal of Autism and Developmental Disorders*, vol. 45, no. 5, pp. 1121–1136, 2015.
[43] D. K. Oller, P. Niyogi, S. Gray, J. A. Richards, J. Gilkerson, D. Xu, U. Yapanel, and S. F. Warren, "Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development," *Proceedings of the National Academy of Sciences*, vol. 107, no. 30, pp. 13 354–13 359, 2010.
[44] C. Ecker, A. Marquand, J. Mourão-Miranda, P. Johnston, E. M. Daly, M. J. Brammer, S. Maltezos, C. M. Murphy, D. Robertson, S. C. Williams, and D. G. M. Murphy, "Describing the brain in autism in five dimensions—magnetic resonance imaging-assisted diagnosis of autism spectrum disorder using a multiparameter classification approach," *Journal of Neuroscience*, vol. 30, no. 32, pp. 10 612–10 623, 2010.
[45] A. Anzulewicz, K. Sobota, and J. T. Delafield-Butt, "Toward the autism motor signature : gesture patterns during smart tablet gameplay identify children with autism," *Scientific Reports*, vol. 6, 2016.
[46] A. Crippa, C. Salvatore, P. Perego, S. Forti, M. Nobile, M. Molteni, and I. Castiglioni, "Use of machine learning to identify children with autism and their motor abnormalities," *Journal of autism and developmental disorders*, vol. 45, no. 7, pp. 2146–2156, 2015.
[47] H. Duan, G. Zhai, X. Min, Z. Che, Y. Fang, X. Yang, J. Gutiérrez, and P. L. Callet, "A dataset of eye movements for the children with autism spectrum disorder," in *Proceedings of the 10th ACM Multimedia Systems Conference*, ser. MMSys '19, 2019, pp. 255–260.
[48] W. Liu, M. Li, and L. Yi, "Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework," *Autism Research*, vol. 9, no. 8, pp. 888–898, 2016.
[49] H. Duan, X. Min, Y. Fang, L. Fan, X. Yang, and G. Zhai, "Visual attention analysis and prediction on human faces for children with autism spectrum disorder," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 15, no. 3s, 2019.
[50] M. Jiang, S. M. Francis, D. Tseng, D. Srishyla, M. DuBois, B. K, C. Conelea, Q. Zhao, and S. Jacob, "Predicting core characteristics of asd through facial emotion recognition and eye tracking in youth," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2020, pp. 871–875.
[51] B. Kim, M. Wattenberg, J. Gilmer, C. J. Cai, J. Wexler, F. B. Viégas, and R. Sayres, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)." in *International Conference on Machine Learning*, vol. 80, 2018, pp. 2673–2682.
[52] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su, "This looks like that: Deep learning for interpretable image recognition," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.
[53] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in Neural Information Processing Systems*, vol. 27, 2014.
[54] S. P. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, vol. 28, pp. 129–137, 1982.
[55] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.
[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
[57] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[58] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.
[59] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.
[60] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 2015.
[61] V. Hus and C. Lord, "The autism diagnostic observation schedule, module 4: Revised algorithm and standardized severity scores," *Journal of Autism and Developmental Disorders*, vol. 44, pp. 1996–2012, 2014.
[62] A. L. Couteur, M. Rutter, C. Lord, P. Rios, S. Robertson, M. Holdgrafer, and J. McLennan, "Autism diagnostic interview: A standardized investigator-based instrument," *Journal of Autism and Developmental Disorders*, vol. 19, pp. 363–387, 1989.
[63] C. Lord, M. Rutter, and A. L. Couteur, "Autism diagnostic interview-revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders," *Journal of Autism and Developmental Disorders*, vol. 24, pp. 659–685, 1994.
[64] V. N. Vapnik, "An overview of statistical learning theory," *Trans. Neur. Netw.*, vol. 10, no. 5, pp. 988–999, 1999.
[65] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9630–9640.
[66] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139, 2021, pp. 8748–8763.
[67] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 15 979–15 988.
[68] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

Shi Chen received the B.E. degree from the School of Computer Science, Wuhan University, Wuhan, China, in 2015. He received the M.S. and Ph.D. degrees from the Department of Computer Science, University of Minnesota, in 2017 and 2023, respectively. His research interests broadly include computer vision, vision and language, human vision, and machine learning.

Ming Jiang received the B.E. and M.E. degrees from Zhejiang University, Hangzhou, China, and the Ph.D. degree in electrical and computer engineering from the National University of Singapore. He is currently a researcher at the Department of Computer Science and Engineering, University of Minnesota. His research interests include computer vision, cognitive vision, machine learning, psychophysics, neuroscience, and brain-machine interface.

Qi Zhao (Member, IEEE) received a Ph.D. degree in computer engineering from the University of California, Santa Cruz in 2009. She is currently an associate professor in the Department of Computer Science and Engineering at the University of Minnesota, Twin Cities. She was a postdoctoral researcher in the Computation & Neural Systems, and the Division of Biology at the California Institute of Technology from 2009 to 2011. She has published more than 100 journal and conference papers in computer vision, machine learning, and cognitive neuroscience venues, and edited a book with Springer, titled Computational and Cognitive Neuroscience of Vision, which provides a systematic and comprehensive overview of vision from various perspectives. She serves as an associate editor at the IEEE Transactions on Neural Networks and Learning Systems (TNNLS) and IEEE Transactions on Multimedia (TMM), as a program chair at IEEE Winter Conference on Applications of Computer Vision (WACV), and as an organizer and/or area chair at IEEE Conference on Computer Vision and Pattern Recognition (CVPR) and other major venues in computer vision and AI. Her main research interests include computer vision, machine learning, cognitive neuroscience, and healthcare.