Multi-Modal Progressive Fusion for ASD Screening Using Smartphone Video

Wenqi Zhong[1], Bohan Li[1,2], Chen Xia[1]✉, Kuan Li[3], and Dingwen Zhang[1]

[1]Northwestern Polytechnical University, Xi'an, China
cxia@nwpu.edu.cn
[2]Tongji University, Shanghai, China
[3]The Affiliated Hospital of Northwest University, Xi'an, China

Abstract. Screening for Autism Spectrum Disorder (ASD) is an important yet challenging task. Traditional screening tools, such as questionnaires and other technical methods, face difficulties in large-scale implementation, such as primary healthcare and home monitoring settings. To address this issue, we develop a smartphone application to highlight atypical eye movement behaviors in children with ASD and extract multi-modal features, including eye movements, head pose, and emotional expressions, from smartphone videos to characterize the subjects' viewing behavior. Additionally, we propose a multi-modal progressive fusion framework to comprehensively integrate the relationships between different modalities. The progressive fusion strategy combines multi-modal features at multiple scales to achieve attention-based deep fusion. Moreover, we develop a global intra- and inter-modality interaction (GIIMI) module to enhance competition and interaction within and between modalities. In the experiment, we constructed a smartphone video dataset of 124 children aged 3 to 6 years and validated the performance advantages of the proposed algorithm.

Keywords: Screening of Autism Spectrum Disorder (ASD) · Multi-Modal Fusion · Smartphones · Gaze Estimation · Transformer

1 Introduction

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition characterized by deficits in social communication, repetitive behaviors, and restricted interests [16]. Research indicates that the prevalence of ASD has been steadily increasing, placing greater pressure on healthcare, education, and social systems, while also adding emotional and financial burdens to families [17]. Early screening and intervention are essential for mitigating the symptoms of ASD and reducing its long-term impacts. Currently, ASD screening primarily relies on scales that assess social interaction, language communication, and related abilities through clinician-to-child observations. However, these traditional methods typically require trained professionals, which makes it challenging to implement widespread and effective screenings, particularly in rural or under-resourced areas with limited access to healthcare facilities [30, 31].

Recently, other advanced technologies such as functional magnetic resonance imaging (fMRI) [1, 2, 7, 11], electroencephalography (EEG) [8, 21], virtual reality [24], eye movement [10, 33], and skin conductance [25] have increasingly been employed as complementary tools for the objective identification of ASD. Despite more objective and accurate results, their widespread adoption in large-scale primary healthcare or home monitoring remains limited due to high costs or operational complexity. Furthermore, previous multi-modal approaches [8,14] often required specialized devices to collect data from multiple modalities, which further escalated screening costs. In recent years, the widespread use of mobile devices, particularly smartphones, has prompted research into mobile-based ASD screening [27]. Mobile screening can not only improve coverage in rural or underdeveloped areas but also address the growing need for remote ASD research, especially in contexts where offline interactions are limited.

Therefore, researchers have explored various approaches for mobile-based ASD screening. For instance, Dow et al. [6] proposed a remote method that allows parents or non-professionals to interact with children, while clinicians assess ASD symptoms by observing social, communicative, and repetitive behaviors in these interactions. Tariq et al. [27] approached ASD screening through children's behavioral videos, where three non-professional evaluators scored the videos on multiple items as the features for classifier training. Deveau et al. [5] introduced a mobile game-based ASD identification method, in which children imitate expressions based on smartphone prompts, and classification is performed based on parents' guesses. Perochon et al. [22] presented a tablet-based early screening model for children under 3 years old, where a parent held their child on their lap to watch movies and participate in a bubble-popping game. In this study, we propose a smarthphone-based, multi-modal ASD recognition paradigm that employs a simple eye-tracking experiment to create a child-friendly, low-cost, and convenient screening method. Specifically, we first design an application that highlights abnormal behaviors in children with ASD. We employ social-geometric contrast scenes to emphasize the atypical eye movements for non-social stimulus preferences of children with ASD. At the same time, we integrate facial expressions and head pose to provide a more comprehensive encoding of the subject's free-viewing behavior.

Furthermore, we propose a multi-modal progressive deep fusion framework to integrate the multi-modal correlations at different time scales. Traditional multi-modal ASD screening methods typically rely on post-processing techniques [8, 14, 22], where features are concatenated or directly input into a classifier, without leveraging the potential of multi-scale feature interaction and fusion. To address this, we first introduce a hybrid CNN-Transformer (CNN-Trans) encoder to extract both local transient changes and global behavioral patterns from long time-series data, such as eye and head pose. Then, we design a progressive fusion module for deep cross-modality integration, using a multi-scale approach that captures both fine-grained and coarse-grained interactions. At each scale, we propose a global intra- and inter-modality interaction (GI-IMI) module to integrate modality interactions. This mechanism concatenates

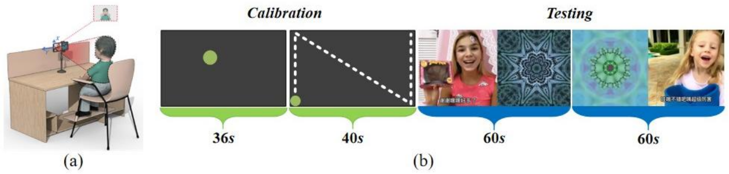

Fig. 1. Data collection via smartphone. (a) Experimental setup for video data collection. (b) Stimuli and time allocation in the smartphone-based test.

keys and values from two modalities to calculate intra- and inter-modality interactions in a unified manner, enabling adaptive fusion of information within and between modalities. Additionally, we introduce an emotion-enhanced fusion (EEF) module to better utilize the correlation between emotional features and eye movements. Finally, the three fused features are aggregated using learnable weights, and cross-entropy loss is employed for classification.

Our main contributions are as follows: (1) We introduce a low-cost, scalable, smartphone-based ASD screening framework that leverages a child-friendly eye-tracking experiment to collect video data and extracts multi-modal features to characterize viewing behavior. (2) We propose a progressive fusion framework that employs a multi-scale fusion strategy and incorporates the GIIMI and EEF modules to adaptively integrate intra- and inter-modality interactions at multiple scales. (3) We construct a mobile-based ASD identification dataset using an Android application and achieve an average accuracy of 86.96%, outperforming state-of-the-art classification models.

2 Video Data Acquisition

Participants and Experimental Scenario. We constructed the dataset using a smartphone-based platform in collaboration with multiple hospitals and rehabilitation centers. A total of 63 children with ASD and 61 typical development (TD) children, aged between 3 and 6, were recruited for data collection. Children with hearing or visual impairments were excluded from the study. All ASD participants were diagnosed by experienced clinicians and met the diagnostic criteria for ASD according to the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-V). The experimental scenario is shown in Fig. 1 (a). During data collection, the mobile phone was placed horizontally, with the subjects seated approximately 0.5 meters away from the device. While the video played, the mobile phone's camera recorded the subjects' head and facial data.

Stimuli. To achieve more accurate eye movement estimation, we designed a calibration phase, as shown in Fig. 1 (b). The calibration consists of two parts to ensure uniform screen coverage and to collect smooth calibration data [13,28]. In the first part, 20 points randomly flash on the screen, each displayed for 1.8 seconds, with sizes ranging from 18 to 50 pixels. The second part involves a green dot moving across the screen in a predefined zigzag pattern.

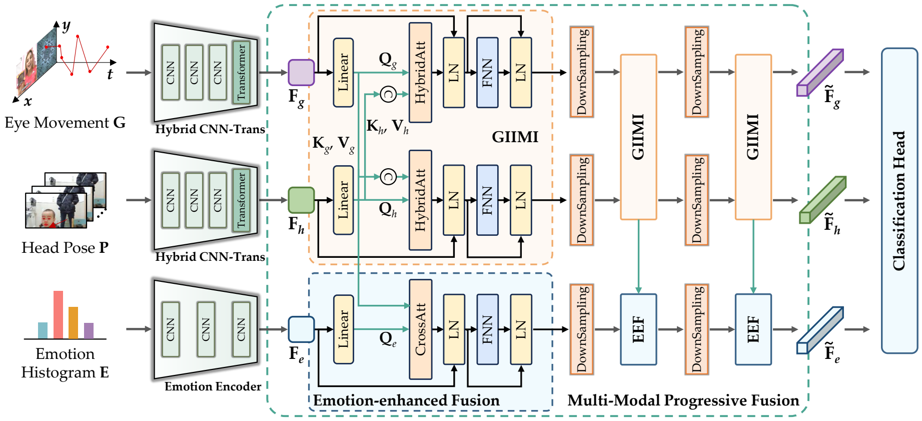

Fig. 2. Framework of the proposed method. First, we extract eye movements, head pose, and emotion data from the recorded video. A hybrid CNN-Trans encoder and an emotion encoder are then used to extract multi-modal features. Finally, the progressive fusion framework employs the GIIMI and EEF modules to deeply fuse these modalities at multiple scales for classification.

In the testing phase, each subject watches a 2-minute video. Children with ASD and TD children exhibit significant differences in attention to geometric patterns and social scenes [23]. To highlight these attention differences, we selected the contrast scenes as stimuli. Specifically, in the first minute, clips of children playing with balloons are shown on the left side, while the right side displays a geometric transformation scene. In the second minute, we swap the positions of the scenes to control for any directional bias.

3 Methodology

The overall architecture of the proposed method is shown in Fig. 2. We first extract multi-modal cues from the collected video and then employ a hybrid CNN-Trans encoder and an emotion encoder to extract multi-modal features (Section 3.1). Subsequently, a progressive fusion framework integrates these features through two key components: the GIIMI module, which fuses eye movements and head pose data, and the EEF module, which enhances emotion features at multiple scales (Section 3.2).

3.1 Multi-modal Data Collection and Feature Extraction

Gaze Estimation. Children with ASD display atypical attention patterns, typically showing a preference for non-social stimuli when compared to TD children [12]. To characterize these attention behaviors, it is crucial to first capture eye movement data from mobile devices. We employ the gaze estimation method [32] on the video $\mathbf{V} \in \mathbb{R}^{H\times W\times T}$ to estimate the subject's eye movement. Specifically, we first perform landmark detection to crop eye images $\mathbf{I}_{\mathrm{e}}$ and face images $\mathbf{I}_{\mathrm{f}}$. These cropped images, along with the corresponding landmark features L, are then used to estimate gaze and represent eye movement behavior. This process can be summarized as follows:

$$\mathbf{G} = \mathcal{G}(\mathbf{I}_{\mathrm{e}}, \mathbf{I}_{\mathrm{f}}, L) \in \mathbb{R}^{T\times 2}, \tag{1}$$

where $\mathbf{G}$ represents the gaze coordinates for the stimuli, and $\mathcal{G}(\cdot)$ denotes the gaze estimation model.

Head Pose Estimation. In social interactions, individuals with ASD may display atypical head pose, such as involuntary head tilting or increased displacement and velocity [18]. These behaviors can lead to distinct patterns in yaw and pitch angles in response to external stimuli. Therefore, we also analyze head poses from video recordings for behavior encoding, as they capture the key cues of head pose. We perform head pose estimation $\mathcal{H}(\cdot)$ using the landmark-based model to analyze these movement differences as follows:

$$\mathbf{P} = \mathcal{H}(\mathbf{V}) \in \mathbb{R}^{T\times 3}, \tag{2}$$

where $\mathbf{P}$ represents the head posture angles, including pitch, yaw, and roll.

Emotion Recognition. Facial expressions are a crucial form of non-verbal communication in social interactions, and individuals with ASD often face difficulties in emotional recognition and expression [3, 19]. Therefore, we perform emotion recognition $\mathcal{E}(\cdot)$ on the video $\mathbf{V}$ to obtain the emotion results for subjects, performed frame by frame, encompassing the emotions: anger, disgust, fear, happiness, sadness, surprise, and neutrality. Emotion recognition is performed using a ResNet-18 model trained on the FER2013 dataset. We further compress the emotion recognition results into an emotion histogram $\mathbf{E} = Histogram(\mathcal{E}(\mathbf{V})) \in \mathbb{R}^7$ to obtain a more comprehensive and robust representation.

Hybrid CNN-Trans Encoder. Eye movement and head pose represent long time series that contain rich information [26], both in local transient changes and global behavioral patterns. Therefore, we propose a hybrid CNN-Trans encoder HybridCNNTrans$(\cdot)$ to capture the local transient changes and global behavior patterns of eye movement and head pose, which can be summarized as follows:

$$\mathbf{F}_g = \mathrm{HybridCNNTrans}(\mathbf{G}), \quad \mathbf{F}_h = \mathrm{HybridCNNTrans}(\mathbf{H}), \tag{3}$$

where $\mathbf{F}_g$ and $\mathbf{F}_h$ denote the eye movement features and head pose features, respectively. Furthermore, we utilize an emotion encoder, which combines linear and convolutional operations to extract emotion feature $\mathbf{F}_e$.

3.2 Multi-Modal Progressive Fusion

Multi-Scale Progressive Fusion. The extracted data exhibit heterogeneous information across different modalities. Most previous works rely on late fusion [20], which ignores the temporal alignment of features. In contrast, we propose a progressive fusion module that extracts features at multiple temporal scales and enables deep fusion of multi-modal interactions at the corresponding scale. Specifically, we first extract multi-scale features from each modality along the temporal dimension. Then, multi-modal cross-attention calculations are performed at each temporal scale. This process can be described as follows:

$$\mathbf{F} = \mathrm{DownSampling}(\mathrm{Fusion}(\mathbf{F}, \mathbf{F}')), \tag{4}$$

where DownSampling$(\cdot)$ represents the downsampling operation using convolutional layers to reduce the resolution of features. This variation in feature resolution across blocks ensures the alignment of features at different scales, which is crucial for effective deep fusion. Fusion$(\cdot)$ refers to the proposed fusion module that fuses the feature $\mathbf{F}$ and $\mathbf{F}'$.

Global Intra- and Inter-Modality Interactions. Existing methods typically perform intra- and inter-modality interactions separately [8, 14], which ignores the trade-offs between these two types of interactions. In this study, we propose a novel approach that simultaneously models intra- and inter-modality interactions to adaptively balance between them. Taking the calculation under the eye movement modality as an example, we first map the eye movement features using a linear layer to obtain the query $\mathbf{Q}_g$, key $\mathbf{K}_g$, and value $\mathbf{V}_g$. Additionally, we extract the head pose features to generate the key $\mathbf{K}_h$ and value $\mathbf{V}_h$. Next, we concatenate the eye movement key $\mathbf{K}_g$ with the head pose key $\mathbf{K}_h$, and the eye movement value $\mathbf{V}_g$ with the head pose value $\mathbf{V}_h$. A hybrid attention mechanism, HybridAtt, is then applied to compute the attention map as follows:

$$\begin{aligned} \mathrm{HybridAtt} &= \mathrm{softmax}\left(\frac{\mathbf{Q}_g \cdot \mathrm{cat}(\mathbf{K}_g^\top, \mathbf{K}_h^\top)}{\sqrt{d_k}}\right) \mathrm{cat}(\mathbf{V}_g, \mathbf{V}_h) \\ &= \big[\mathrm{SelfAttMap}, \mathrm{CrossAttMap}\big] \begin{bmatrix} \mathbf{V}_g \\ \mathbf{V}_h \end{bmatrix}, \end{aligned} \tag{5}$$

where cat$(\cdot)$ represents the concatenation operation along the sequence length dimension. softmax$(\cdot)$ and d_k denote the softmax operation and the channel dimension, respectively. In this process, we compute the intra-similarity using $\mathbf{Q}_g \cdot \mathbf{K}_g^\top$ to obtain the self-attention score, and the inter-similarity using $\mathbf{Q}_g \cdot \mathbf{K}_h^\top$ to obtain the cross-attention score. Then, we proceed with the conventional calculation steps as follows:

$$\begin{aligned} \mathbf{F}'_g &= \mathrm{LayerNorm}(\mathbf{F}_g + \mathrm{HybridAtt}(\mathbf{F}_g, \mathbf{F}_h)) \\ \tilde{\mathbf{F}}_g &= \mathrm{LayerNorm}(\mathbf{F}'_g + \mathrm{FeedForward}(\mathbf{F}'_g)). \end{aligned} \tag{6}$$

where LayerNorm$(\cdot)$ and FeedForward$(\cdot)$ denote the layer normalization and the multi-layer feedforward network, respectively.

Furthermore, due to the inherent correlation between eye movements and emotion, eye movements can enhance emotion recognition more effectively. To leverage this relationship, we propose an emotion-enhanced fusion method that integrates the emotion histogram with eye movements. Specifically, we use the emotion query $\mathbf{Q}_e$ together with the eye movement key $\mathbf{K}_g$ to compute the attention map.

Table 1. Comparison of the proposed method with baselines.

Methods	**Accuracy**	**Precision**	**Recall**	**F1-score**
Percentage of Different Scenes	0.6297	0.6306	0.6231	0.6232
X and Y Coordinates	0.6440	0.6405	0.6385	0.6337
Distance from the Center	0.5557	0.5451	0.5923	0.5653
Saccadic Amplitude	0.6373	0.5988	0.8192	0.6887
Changes in Saccadic Direction	0.6207	0.6837	0.6654	0.5674
Percentage of Saccadic Direction	0.6850	0.7456	0.5564	0.6352
APM [4]	0.7747	0.8833	0.6215	0.7257
DVP [30]	0.7743	0.8440	0.6703	0.7457
GPF [34]	0.7340	0.7422	0.7333	0.7311
SOFT [9]	0.7391	0.7143	0.8333	0.7692
iTransformer [15]	0.7826	**1.0000**	0.6429	0.7826
Proposed	**0.8696**	0.7857	**1.0000**	**0.8800**

Classification Head. After the intra- and inter-modality fusion, we apply a learnable weight to combine the features from three modalities. We then use a multilayer perceptron (MLP) as a classification head to classify the subjects. Additionally, we employ cross-entropy loss to optimize the model and obtain the classification results.

4 Experiment

Implementation Details. The smartphone screen has a resolution of 1280×720 pixels, while the collected video V has a resolution of 1080×1920. For gaze estimation, we resized eye images to 112×112 pixels and face images to 224×224. The hybrid CNN-Trans encoder consists of three convolutional layers followed by a Transformer block. The emotion encoder includes a linear layer and three convolutional layers. In the multi-modal progressive fusion strategy, the feature dimensions for the three stages are 256, 128, and 64, respectively. We used a batch size of 64 and train the model for 10 epochs. The initial learning rate is set to 1×10^{-3}, with a weight decay of 5×10^{-3}. We employed stratified five-fold cross-validation to evaluate performance. All experiments were conducted in PyTorch, using the Adam optimizer.

Evaluation Metrics. To comprehensively evaluate the models, we used four metrics: accuracy, precision, recall, and F1-score. Precision measures the proportion of true positives among all predicted positives, while recall reflects the model's ability to identify all actual positive instances. To balance the trade-off between precision and recall, we compute the F1-score, which provides a harmonized metric that captures both aspects of model performance.

Performance Comparison. We compare the proposed method with several baseline models, as shown in Table 1. Specifically, we evaluate it against a range of statistical methods [12,29] and state-of-the-art deep learning models, including

Table 2. Ablation studies on the impact of modality utilization.

id	Eye Movement	Head Pose	Emotion	**Accuracy**	**Precision**	**Recall**	**F1-score**
(a)	✓	✗	✗	0.6761	0.7297	0.6426	0.6686
(b)	✗	✓	✗	0.7181	0.6936	0.8233	0.7490
(c)	✗	✗	✓	0.7177	**0.7872**	0.6247	0.6922
(d)	✓	✓	✓	**0.8696**	0.7857	**1.000**	**0.8800**

Table 3. Ablation studies on the proposed module. "PF" represents the progressive fusion strategy. "Fusion" represents the GIIMI and EEF modules.

id	CNN-Trans	Fusion	PF	**Accuracy**	**Precision**	**Recall**	**F1-score**
(a)	✗	✗	✗	0.7391	0.7273	0.7273	0.7273
(b)	✗	✓	✓	0.8261	**0.8182**	0.8182	0.8182
(c)	✓	✓	✗	0.7826	0.8571	0.8000	0.8276
(d)	✓	✗	✓	0.7500	0.7059	0.9231	0.8000
(e)	✓	✓	✓	**0.8696**	0.7857	**1.000**	**0.8800**

APM [4], DVP [30], GPF [34], SOFT [9], and iTransformer [15]. The experimental results demonstrate that our method achieves an accuracy of 86.96% and the highest F1-score, surpassing existing approaches. The superior recall performance underscores the method's capability to reduce the likelihood of missed diagnoses. These findings highlight the potential of our framework as a robust and effective tool for ASD screening.

Ablation Study. First, as shown in Table 2, we evaluate the performance of individual signal modalities to validate the effectiveness of multi-modal fusion. The results indicate that combining data from different modalities can improve the encoding of viewing behavior and uncover hidden ASD features, thereby improving recognition accuracy. On the other hand, we evaluate the contribution of each module. As shown in Table 3, removing any module results in performance degradation, indicating that each module plays a crucial role in the overall improvement. For instance, the model without "Fusion" (d) reverts to traditional late fusion model, causing a 11.96% drop in accuracy. This underscores the importance of exploring and leveraging the correlations among different modalities.

5 Conclusion

In this study, we propose a scalable, multi-modal, and at-home ASD screening framework that captures participant videos using smartphone cameras. Based on the core characteristics of ASD, we extract three distinct modalities: eye movements, head pose, and emotion. Furthermore, we develop a multi-modal progressive fusion model with hierarchical feature integration across different scales. The proposed GIIMI and EEF modules leverage intra- and inter-modality interactions to enhance modal complementarity. Experimental results demonstrate significant improvements in recognition accuracy compared to previous methods, offering a promising foundation for more accessible and efficient ASD detection in real-world settings.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grants 62036005, 62336001, 62172334. It was also sponsored by the Innovation Foundation for Doctoral Dissertations of Northwestern Polytechnical University CX2024015.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bannadabhavi, A., Lee, S., Deng, W., Ying, R., Li, X.: Community-aware transformer for autism prediction in fmri connectome. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 287–297. Springer (2023)
2. Berto, S., Treacher, A.H., Caglayan, E., Luo, D., Haney, J.R., Gandal, M.J., Geschwind, D.H., Montillo, A.A., Konopka, G.: Association between resting-state functional brain connectivity and gene expression is altered in autism spectrum disorder. Nat. Commun. **13**(1), 3328 (2022)
3. Chaidi, I., Drigas, A.: Autism, expression, and understanding of emotions: literature review. International Journal of Online and Biomedical Engineering (2020)
4. Chen, S., Zhao, Q.: Attention-based autism spectrum disorder screening with privileged modality. In: Proc. IEEE Int. Conf. Comput. Vis. pp. 1181–1190 (2019)
5. Deveau, N., Washington, P., Leblanc, E., Husic, A., Dunlap, K., Penev, Y., Kline, A., Mutlu, O.C., Wall, D.P.: Machine learning models using mobile game play accurately classify children with autism. Intelligence-Based Medicine **6**, 100057 (2022)
6. Dow, D., Holbrook, A., Toolan, C., McDonald, N., Sterrett, K., Rosen, N., Kim, S.H., Lord, C.: The brief observation of symptoms of autism (bosa): development of a new adapted assessment measure for remote telehealth administration through covid-19 and beyond. Journal of Autism and Developmental Disorders **52**(12), 5383–5394 (2022)
7. Duan, K., Eyler, L., Pierce, K., Lombardo, M.V., Datko, M., Hagler, D.J., Taluja, V., Zahiri, J., Campbell, K., Barnes, C.C., et al.: Differences in regional brain structure in toddlers with autism are related to future language outcomes. Nat. Commun. **15**(1), 5075 (2024)
8. Han, J., Jiang, G., Ouyang, G., Li, X.: A multimodal approach for identifying autism spectrum disorders in children. IEEE Trans. Neur. Sys. Reh. **30**, 2003–2011 (2022)
9. Han, L., Chen, X.Y., Ye, H.J., Zhan, D.C.: Softs: Efficient multivariate time series forecasting with series-core fusion. arXiv preprint arXiv:2404.14197 (2024)
10. Her, P., Manderle, L., Dias, P.A., Medeiros, H., Odone, F.: Uncertainty-aware gaze tracking for assisted living environments. IEEE Trans. on Image Process. (2023)
11. Jeong, A.Y., Heo, D.W., Kang, E., Suk, H.I.: BrainWaveNet: Wavelet-based Transformer for Autism Spectrum Disorder Diagnosis . In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. vol. LNCS 15002, pp. 56 – 66. Springer Nature Switzerland (October 2024)

12. Jones, W., Klin, A.: Attention to eyes is present but in decline in 2–6-month-old infants later diagnosed with autism. Nature **504**(7480), 427–431 (2013)
13. Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., Torralba, A.: Eye tracking for everyone. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 2176–2184 (2016)
14. Li, J., Bhat, A., Barmaki, R.: A two-stage multi-modal affect analysis framework for children with autism spectrum disorder. arXiv preprint arXiv:2106.09199 (2021)
15. Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., Long, M.: itransformer: Inverted transformers are effective for time series forecasting. arXiv preprint arXiv:2310.06625 (2023)
16. Lord, C., Charman, T., Havdahl, A., Carbone, P., Anagnostou, E., Boyd, B., Carr, T., De Vries, P.J., Dissanayake, C., Divan, G., et al.: The lancet commission on the future of care and clinical research in autism. The Lancet **399**(10321), 271–334 (2022)
17. Maenner, M.J.: Prevalence and characteristics of autism spectrum disorder among children aged 8 years¡ªautism and developmental disabilities monitoring network, 11 sites, united states, 2020. MMWR. Surveillance Summaries **72** (2023)
18. Martin, K.B., Hammal, Z., Ren, G., Cohn, J.F., Cassell, J., Ogihara, M., Britton, J.C., Gutierrez, A., Messinger, D.S.: Objective measurement of head movement differences in children with and without autism spectrum disorder. Molecular autism **9**, 1–10 (2018)
19. Mazefsky, C.A., Herrington, J., Siegel, M., Scarpa, A., Maddox, B.B., Scahill, L., White, S.W.: The role of emotion regulation in autism spectrum disorder. Journal of the American Academy of Child & Adolescent Psychiatry **52**(7), 679–688 (2013)
20. Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., Sun, C.: Attention bottlenecks for multimodal fusion. Advances in neural information processing systems **34**, 14200–14213 (2021)
21. Peng, Z., He, Z., Jiang, Y., Wang, P., Yuan, Y.: GBT: Geometric-oriented Brain Transformer for Autism Diagnosis . In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. vol. LNCS 15012, pp. 142 – 152. Springer Nature Switzerland (October 2024)
22. Perochon, S., Di Martino, J.M., Carpenter, K.L., Compton, S., Davis, N., Eichner, B., Espinosa, S., Franz, L., Krishnappa Babu, P.R., Sapiro, G., et al.: Early detection of autism using digital behavioral phenotyping. Nature Medicine **29**(10), 2489–2497 (2023)
23. Pierce, K., Marinero, S., Hazin, R., McKenna, B., Barnes, C.C., Malige, A.: Eye tracking reveals abnormal visual preference for geometric images as an early biomarker of an autism spectrum disorder subtype associated with increased symptom severity. Biol. Psychiat. **79**(8), 657–666 (2016)
24. Robles, M., Namdarian, N., Otto, J., Wassiljew, E., Navab, N., Falter-Wagner, C.M., Roth, D.: A virtual reality based system for the screening and classification of autism. IEEE transactions on visualization and computer graphics **28**(5), 2168–2178 (2022)
25. Sharma, A., Khosla, A., Khosla, M.: Skin conductance response patterns of face processing in children with autism spectrum disorder. Advances in autism **3**(2), 76–86 (2017)
26. Song, Y., Wang, X., Yao, J., Liu, W., Zhang, J., Xu, X.: Vitgaze: gaze following with interaction features in vision transformers. Visual Intelligence **2**(1), 1–15 (2024)
27. Tariq, Q., Daniels, J., Schwartz, J.N., Washington, P., Kalantarian, H., Wall, D.P.: Mobile detection of autism through machine learning on home video: A development and prospective validation study. PLoS Medicine **15**(11), e1002705 (2018)

28. Valliappan, N., Dai, N., Steinberg, E., He, J., Rogers, K., Ramachandran, V., Xu, P., Shojaeizadeh, M., Guo, L., Kohlhoff, K., et al.: Accelerating eye movement research via accurate and affordable smartphone eye tracking. Nat. Commun. **11**(1), 1–12 (2020)
29. Wang, S., Jiang, M., Duchesne, X.M., Laugeson, E.A., Kennedy, D.P., Adolphs, R., Zhao, Q.: Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking. Neuron **88**(3), 604–616 (2015)
30. Xia, C., Zhang, D., Li, K., Li, H., Chen, J., Min, W., Han, J.: Dynamic viewing pattern analysis: towards large-scale screening of children with ASD in remote areas. IEEE Trans. Biomed. Eng. **70**(5), 1622–1633 (2023)
31. Zhong, W., Xia, C., Yu, L., Li, K., Li, Z., Zhang, D., Han, J.: A learning paradigm for selecting few discriminative stimuli in eye-tracking research. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
32. Zhong, W., Xia, C., Zhang, D., Han, J.: Uncertainty modeling for gaze estimation. IEEE Trans. on Image Process. (2024)
33. Zhong, W., Yu, L., Xia, C., Han, J., Zhang, D.: Spformer: Spatio-temporal modeling for scanpaths with transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 7605–7613 (2024)
34. Zhou, W., Yang, M., Tang, J., Wang, J., Hu, B.: Gaze patterns in children with autism spectrum disorder to emotional faces: Scanpath and similarity. IEEE Transactions on Neural Systems and Rehabilitation Engineering (2024)