Smartphone-Based ASD Screening via Instance-Adaptive Cross-Modal Fusion and Contrastive Learning *

Jie Xu[1], Xinran Guo[1], Longbin Tang[1], Kuan Li[2], and Chen Xia[1]**

[1] Northwestern Polytechnical University, Xi'an, China
[2] The Affiliated Hospital of Northwest University, Xi'an, China

Abstract. Smartphone-based screening for Autism Spectrum Disorder (ASD) provides a scalable and low-cost alternative to traditional clinical assessments. However, unconstrained mobile data are often noisy and exhibit high inter-subject behavioral variability, which undermines the effectiveness of previous unimodal and static fusion methods. To address these challenges, we propose an adaptive multimodal fusion network that jointly models eye-movement patterns and temporal head-pose dynamics to capture instance-wise complementary behavioral cues. Specifically, we first extract high-level eye-movement features using a dual-branch network that encodes both static fixation and dynamic scanpath information. We then introduce a joint representation gating module for instance-level adaptive fusion, which dynamically reweights modalities to suppress unreliable signals and enhance discriminative cues. Additionally, we incorporate supervised contrastive learning as an auxiliary objective to structure the latent space, thereby improving intra-class compactness and inter-class separability, as well as overall generalization in limited and heterogeneous medical datasets. Experiments on our collected real-world mobile dataset of 124 children demonstrate state-of-the-art performance, achieving 91.93% accuracy and an F1-score of 0.9206.

Keywords: Autism Spectrum Disorder Screening · Mobile Health · Multimodal Fusion · Eye Movement · Supervised Contrastive Learning

1 Introduction

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterized by persistent deficits in social communication and restricted, repetitive behaviors [18]. Early diagnosis and intervention are crucial for improving long-term outcomes [11]. Conventional screening relies on standardized questionnaires and clinician-administered assessments [6]. To mitigate subjective bias, automated screening methods based on behavioral biomarkers, such as eye-tracking, electroencephalography (EEG), and facial analysis, are increasingly studied [23, 15].

* This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 62572402 and 62336001.
** Corresponding author: cxia@nwpu.edu.cn

Among these biomarkers, eye-tracking exhibits distinct advantages and has been widely studied due to its ease of use, suitability for children, and short duration. However, traditional eye-tracking studies require controlled laboratory settings and specialized equipment [26], limiting scalability. Recently, with advancements in gaze estimation, mobile-based eye-tracking has emerged as a research direction for ASD screening [25]. For instance, Chang et al. [3] estimated fixation positions and extracted left–right fixation ratios as features for classification.

Despite recent advances, mobile-based eye-tracking for ASD screening remains challenging. Mobile scenarios are inherently susceptible to noise caused by illumination variation, device motion, calibration errors, and head movement, degrading gaze estimation accuracy [13, 29]. Consequently, methods relying on precise fixation localization, such as object- or ROI-based features, may experience performance degradation. Moreover, ASD exhibits substantial behavioral heterogeneity. While some children show atypical gaze patterns, others may display subtle gaze differences but abnormal head movements [2]. In mobile settings, with less constrained head motion and noisier gaze estimation, relying solely on eye-tracking signals may be insufficient to capture ASD-related behaviors.

These limitations have motivated multimodal approaches for ASD screening to address unimodal constraints. Some methods combined name-calling responses with head pose, shoulder rotation, and response latency [33, 27], while others extracted diverse behavioral features from multiple tasks [5]. Eye-tracking has also been integrated with EEG using graph convolutional networks [28] or deep autoencoders [7]. Similarly, Perochon et al. [21] developed a tablet-based model for toddlers using statistical features and XGBoost, and Zhong et al. [30] introduced a multi-stream architecture to progressively combine information across modalities. Nevertheless, most existing multimodal ASD screening methods rely on static fusion strategies (e.g., concatenation or fixed-weight summation) [9], assuming uniform modality contributions across subjects. However, ASD behavioral patterns are highly heterogeneous across individuals [18], and eye-tracking reliability may fluctuate due to gaze loss or motion artifacts [19].

To address these challenges, we propose the Adaptive Multimodal Fusion Network (AMF-Net), a robust framework for mobile ASD screening. Instead of relying on low-level eye-movement features, we first construct high-level visual representations, including fixation heatmaps and scanpath maps, to obtain more robust mobile eye-tracking representations, and integrate them with temporal head-pose sequences. To handle multimodal heterogeneity, we design a Joint Representation Gating (JRG) module, which models cross-modal interactions to generate instance-level weights, suppressing noisy modalities while emphasizing discriminative cues. Furthermore, we introduce a joint optimization strategy combining cross-entropy and supervised contrastive learning [14], which imposes structural constraints on the latent space to promote intra-class compactness and inter-class separability. The main contributions are summarized as follows:

– We propose **AMF-Net**, an end-to-end multimodal framework that integrates eye-movement and head-pose cues to achieve robust visual behavior encoding for ASD screening in unconstrained mobile scenarios.

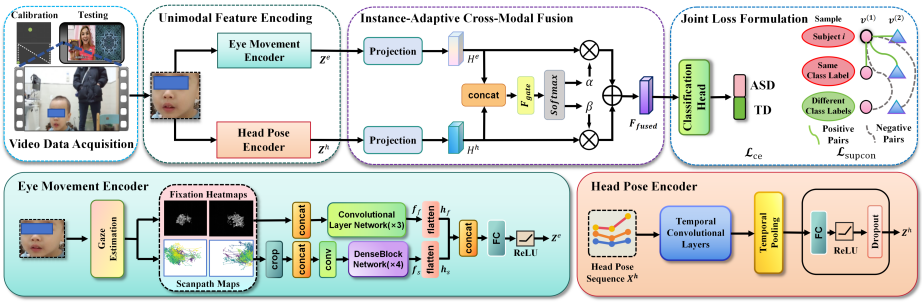

Fig. 1: Framework of the AMF-Net for smartphone-based ASD screening.

– We introduce instance-level adaptive cross-modal fusion via the **JRG module** to dynamically adjust modality contributions, and incorporate cross-modal contrastive learning to enhance feature representation.
– Our method achieves state-of-the-art classification performance on a real-world dataset of 124 subjects, demonstrating the superiority of instance-wise adaptive fusion over existing unimodal and multimodal approaches.

2 Methodology

2.1 Video Data Acquisition

We recruited 63 children with ASD and 61 typically developing (TD) children aged 2–8 years. ASD and TD status were established based on ADOS-2 assessments and clinical diagnoses by qualified clinicians. All participants had normal or corrected-to-normal vision, and those unable to complete the task were excluded. We developed an Android application to present visual stimuli and record facial videos. Each participant completed a 2-minute calibration procedure consisting of 20 randomly flashing points and a moving dot, followed by a 2-minute test video presenting social scenes and geometric patterns simultaneously, as described in [22]. To control screen-side bias, the test had two 1-minute sessions with left-right stimulus positions swapped. Social videos depicted a young girl interacting with her sister or two children playing a balloon game. Participants viewed the stimuli on a horizontally placed smartphone (1280×720 px) at 0.5 m, while the device's camera recorded head and facial movements in a quiet room with minimal personnel present. This study was approved by the Medical Ethics Committee of Northwestern Polytechnical University, China (Protocol No. 202502045). Written informed consent was obtained from the legal guardians of all participants in accordance with the Declaration of Helsinki.

2.2 Unimodal Feature Encoding

Eye-Movement Encoder The overall architecture of AMF-Net is illustrated in Fig. 1. To model visual behavioral patterns from mobile videos, we first employ an appearance-based gaze estimaton model [31] to obtain frame-wise gaze

coordinates. The model is pre-trained on GazeCapture [16] and fine-tuned using calibration data to improve robustness in unconstrained settings. Instead of using low-level features, the eye-movement encoder adopts a dual-stream CNN to jointly capture static and dynamic cues, learning discriminative eye-movement representations. Specifically, we construct two structured inputs: (1) fixation heatmaps, obtained by accumulating fixation points with Gaussian smoothing to encode spatial attention, and (2) scanpath maps, generated by connecting consecutive fixations with temporal color to preserve sequential dynamics.

The first branch processes cumulative fixation heatmaps from two sessions, $\mathcal{M}_f^1$ and $\mathcal{M}_f^2$, concatenated along the channel dimension to improve robustness to session variability, and encoded by the fixation heatmap encoder $\mathcal{E}_f$ to extract global spatial features:

$$\boldsymbol{f}_f = \mathcal{E}_f(\mathcal{F}_{concat}(\mathcal{M}_f^1, \mathcal{M}_f^2)), \tag{1}$$

where $\mathcal{E}_f$ comprises three convolutional layers, each followed by max pooling to progressively aggregate spatial context and encode global attention patterns.

The second branch captures fine-grained spatiotemporal dynamics from scanpath maps $\mathcal{M}_s^1$ and $\mathcal{M}_s^2$. Unlike the first branch, which encodes global fixation information, we apply a fixed center-cropping operation $\mathcal{F}_{crop}$ to reduce peripheral noise in detailed representation. The cropped maps are then concatenated and fed into a scanpath encoder $\mathcal{E}_s$ to extract detailed dynamic features:

$$\boldsymbol{f}_s = \mathcal{E}_s\big(\mathcal{F}_{concat}(\mathcal{F}_{crop}(\mathcal{M}_s^1), \mathcal{F}_{crop}(\mathcal{M}_s^2))\big), \tag{2}$$

where $\mathcal{E}_s$ consists of an initial convolutional layer followed by a four-layer DenseBlock [10] with dense connectivity to enhance feature reuse and gradient flow.

The resulting feature maps, $\boldsymbol{f}_h$ and $\boldsymbol{f}_s$, are first flattened into one-dimensional vectors, $\boldsymbol{h}_f = \mathcal{F}_{flatten}(\boldsymbol{f}_f)$ and $\boldsymbol{h}_s = \mathcal{F}_{flatten}(\boldsymbol{f}_s)$, and then concatenated and passed through a fully connected layer with ReLU activation to obtain the final eye-movement representation, $\boldsymbol{Z}^e$:

$$\boldsymbol{Z}^e = \mathrm{ReLU}(\mathcal{F}_{concat}(\boldsymbol{h}_f, \boldsymbol{h}_s) \cdot \boldsymbol{W}_1 + \boldsymbol{b}_1), \tag{3}$$

where $\boldsymbol{W}_1$ and $\boldsymbol{b}_1$ denote the learnable parameters of the fully connected layer.

Head-Pose Encoder Head pose data, represented as Euler angle sequences (yaw, pitch, roll), encodes subjects' motor control patterns and long-range dependencies. To extract discriminative features from the temporal input $\boldsymbol{X}^h$, we employ a temporal TCN [1], which integrates temporal feature extraction, aggregation, and semantic projection in an end-to-end manner:

$$\boldsymbol{Z}^h = \mathrm{Dropout}\left(\mathrm{ReLU}\big(\boldsymbol{W}_2 \cdot \mathcal{P}(\Phi_{TCN}(\boldsymbol{X}^h)) + \boldsymbol{b}_2\big)\right), \tag{4}$$

where $\Phi_{TCN}(\cdot)$ stacks causal dilated convolutions with varying dilation rates to enlarge the receptive field while preserving temporal causality. The resulting sequence features are aggregated into a fixed-length representation via temporal

pooling $\mathcal{P}(\cdot)$. The pooled vector is then projected to the semantic space through a fully connected layer parameterized by $\boldsymbol{W}_2$ and $\boldsymbol{b}_2$, followed by ReLU activation and Dropout regularization. The output $\boldsymbol{Z}^h$ serves as a compact high-level head-pose representation for subsequent multimodal fusion.

2.3 Instance-Adaptive Cross-Modal Fusion

In unconstrained mobile screening scenarios, the quality and stability of behavioral signals may vary substantially across subjects and recording conditions. Eye-movement estimation can be affected by gaze loss, illumination variation, or calibration drift, while head-pose measurements may fluctuate due to spontaneous movements. Such variability renders fixed or globally learned fusion weights suboptimal. To address this issue, we introduce an instance-level Joint Representation Gating (JRG) module that adaptively weights modality contributions for each sample. Given high-level unimodal representations $\boldsymbol{Z}^e$ and $\boldsymbol{Z}^h$, we first project them into a shared latent space:

$$\boldsymbol{H}^e = \phi_e(\boldsymbol{Z}^e), \quad \boldsymbol{H}^h = \phi_h(\boldsymbol{Z}^h), \tag{5}$$

where $\phi_e(\cdot)$ and $\phi_h(\cdot)$ denote modality-specific nonlinear projections that mitigate cross-modal heterogeneity and enable direct aggregation.

Then, we estimate modality weights conditioned on the joint representation:

$$[\alpha, \beta] = \mathrm{Softmax}\big(\mathcal{F}_{\mathrm{gate}}\big(\mathcal{F}_{\mathrm{concat}}(\boldsymbol{H}^e, \boldsymbol{H}^h)\big)\big), \tag{6}$$

where $\mathcal{F}_{\mathrm{gate}}$ is a lightweight multilayer perceptron that produces modality-specific logits normalized by $\mathrm{Softmax}(\cdot)$. Conditioned on the joint representation, the learned weights adaptively modulate modality contributions for each instance. For example, when gaze signals are unstable, the eye-movement weight is automatically reduced, allowing the model to rely more on head-pose cues. By dynamically adjusting modality contributions, the model becomes more robust to signal instability in mobile ASD screening.

The final fused representation is computed as:

$$\boldsymbol{F}_{\mathrm{fused}} = \alpha \boldsymbol{H}^e + \beta \boldsymbol{H}^h. \tag{7}$$

2.4 Joint Loss Formulation

To improve generalization under limited samples and inter-subject variability, we adopt a joint objective combining cross-entropy loss with supervised contrastive regularization [14]. In the contrastive branch, the eye-movement and head-pose embeddings of the same sample are treated as two aligned views in a shared embedding space. Given a mini-batch of B samples, the unimodal embeddings are denoted as $\mathbf{z}_i^e$ and $\mathbf{z}_i^h$ ($i \in [1, B]$). After aligning the eye-movement embedding via a linear mapping $\varphi(\cdot)$, both embeddings are processed by a shared projection head $\psi(\cdot)$ followed by ℓ_2 normalization $\mathrm{Norm}(\cdot)$:

$$\mathbf{v}_i^{(1)} = \mathrm{Norm}\left(\psi(\varphi(\mathbf{z}_i^e))\right), \quad \mathbf{v}_i^{(2)} = \mathrm{Norm}\left(\psi(\mathbf{z}_i^h)\right). \tag{8}$$

For an anchor view $\mathbf{v}_i^{(m)}$, the positive set $P(i,m)$ includes all mini-batch views sharing the same class label as sample i (including the other-modality view of the same sample and views from other same-class samples), excluding the anchor itself. The anchor-wise supervised contrastive loss is defined as:

$$\mathcal{L}_{i,m} = -\frac{1}{|P(i,m)|} \sum_{\mathbf{v}_p \in P(i,m)} \log \frac{\exp(\text{sim}(\mathbf{v}_i^{(m)}, \mathbf{v}_p))}{\sum_{\mathbf{v}_n \in D(i,m)} \exp(\text{sim}(\mathbf{v}_i^{(m)}, \mathbf{v}_n))}, \tag{9}$$

where $D(i,m)$ denotes all views in the batch excluding the anchor and $\text{sim}(\cdot,\cdot)$ is the temperature-scaled similarity. The numerator promotes intra-class alignment, while the denominator performs batch-wise normalization, which separates the anchor from hard negatives and enlarges inter-class margins. Averaging $\mathcal{L}_{i,m}$ over all anchors in the batch yields the final supervised contrastive loss $\mathcal{L}_{\text{supcon}}$.

For classification, the fused representation $\boldsymbol{F}_{\text{fused}}$ is fed into a classifier to produce logits, which are normalized via Softmax to obtain predicted probabilities for computing the cross-entropy loss $\mathcal{L}_{\text{ce}}$. The overall optimization objective is thus formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ce}} + \lambda_{\text{con}} \mathcal{L}_{\text{supcon}}, \tag{10}$$

where λ_{con} balances the contribution of contrastive regularization. In this formulation, $\mathcal{L}_{\text{ce}}$ encourages learning discriminative decision boundaries for classification, while $\mathcal{L}_{\text{supcon}}$ enforces cross-modal semantic consistency within each class and enlarges inter-class margins, thereby improving robustness to individual differences and noisy modality measurements.

3 Experiments

3.1 Implementation Details

Experiments were conducted in `PyTorch` using Adam optimizer, with a learning rate and weight decay of 1×10^{-4}, batch size 16, and 30 training epochs. The gaze branch used three convolutional layers (32/64/128 filters) for 128×128 inputs, while the scanpath branch comprised a 32-filter convolution, a four-layer DenseBlock (growth rate 16), a 1×1 transition layer, and three max-pooling layers. Performance was evaluated via subject-independent 5-fold cross-validation with Accuracy, Precision, Recall, and F1-score.

3.2 Overall Preformance

As shown in Table 1, we compare our framework with representative baselines: APM [4], DVP [26], GPF [32], SOFT [8], iTransformer [17], and MMPF [30] for a comprehensive comparison in mobile ASD screening. Most models exhibit moderate performance, with the mobile multimodal competitor MMPF achieving high recall but lower precision, likely due to environmental sensitivity. In contrast, our method achieves strong and balanced performance across all metrics, demonstrating the effective integration of cues for mobile ASD screening.

Table 1: Comparison of the Proposed Method with State-of-the-Art Models

Method	Accuracy	Precision	Recall	F1-Score
APM [4]	0.7747	0.8833	0.6215	0.7257
DVP [26]	0.7743	0.8440	0.6703	0.7457
GPF [32]	0.7340	0.7422	0.7333	0.7311
SOFT [8]	0.7391	0.7143	0.8333	0.7692
iTransformer [17]	0.7826	**1.0000**	0.6429	0.7826
MMPF [30]	0.8696	0.7857	1.0000	0.8800
AMF-Net	**0.9193**	0.9218	**0.9246**	**0.9206**

Table 2: Ablation Study on Modalities, Fusion Strategies, and Loss Functions

Settings	Accuracy	Precision	Recall	F1-Score
1. Impact of Input Modalities				
Eye-Movement Only	0.8307	**0.8803**	0.7831	0.8243
Head-Pose Only	0.7907	0.7509	**0.9192**	0.8206
Multimodal (Concatenation)	**0.8710**	0.8710	0.8792	**0.8722**
2. Impact of Fusion Strategies				
Concatenation	0.8710	0.8710	0.8792	0.8722
Equal Weights	0.8873	**0.9170**	0.8679	0.8827
Global Learned Weights	0.8793	0.8908	0.8846	0.8790
Cross-Attention [24]	0.8873	0.8733	0.9023	0.8848
FiLM [20]	0.8797	0.9083	0.8403	0.8703
Uncertainty-Aware Fusion [12]	0.8870	0.8560	**0.9398**	0.8894
JRG (Proposed w/o $\mathcal{L}_{\text{supcon}}$)	**0.9037**	0.8899	0.9367	**0.9058**
3. Impact of Supervised Contrastive Learning				
JRG + $\mathcal{L}_{\text{ce}}$ (Baseline)	0.9037	0.8899	**0.9367**	0.9058
JRG + $\mathcal{L}_{\text{ce}} + 0.2\mathcal{L}_{\text{supcon}}$	0.9033	0.8929	0.9226	0.9057
JRG + $\mathcal{L}_{\text{ce}} + 0.5\mathcal{L}_{\text{supcon}}$	0.9113	0.9081	0.9226	0.9130
JRG + $\mathcal{L}_{\text{ce}} + 0.1\mathcal{L}_{\text{supcon}}$ (Proposed)	**0.9193**	**0.9218**	0.9246	**0.9206**

3.3 Ablation Study

Impact of Input Modalities As shown in Table 2, we first assess the contribution of individual behavioral modalities. Eye-movement alone achieves high precision but lower recall, potentially missing some ASD cases, whereas head-pose alone attains higher recall at the expense of precision, indicating more false positives. Simple concatenation of both modalities enhances overall performance, demonstrating their complementary role in ASD characterization.

Impact of Fusion Strategies We then analyze different fusion mechanisms. *Concatenation* combines eye-movement and head-pose features without considering their relative importance. *Equal Weights* assigns fixed, equal contributions, improving over concatenation but assuming uniform influence across modalities. *Global Learned Weights* introduces learnable weights but applies them globally,

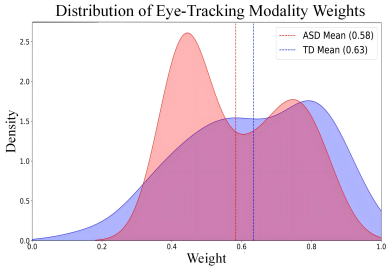

Fig. 2: Eye-tracking weights in adaptive multimodal fusion.

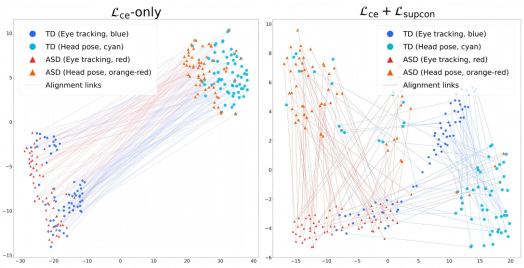

Fig. 3: Cross-modal feature alignment under different loss functions.

limiting adaptation to individual behaviors. In addition, representative multimodal fusion methods such as Cross-Attention [24], FiLM [20], and Uncertainty-Aware Fusion [12] also fail to achieve competitive performance in our setting. In contrast, *JRG* dynamically adjusts modality weights at the instance level, capturing individual differences more accurately and outperforming all other strategies under cross-validation, particularly in recall, while maintaining high precision for ASD recognition in variable scenarios.

Impact of Supervised Contrastive Learning Finally, we analyze the effect of loss functions. Introducing the Supervised Contrastive loss $\mathcal{L}_{\text{supcon}}$ boosts performance, with $\lambda_{\text{con}} = 0.1$ yielding the best results across all metrics. This suggests that $\mathcal{L}_{\text{supcon}}$ effectively regularizes the model, enhancing intra-class compactness and inter-class separability. However, larger weights (e.g., 0.2 or 0.5) degrade performance, likely due to the model over-prioritizing geometric structure at the expense of classification.

3.4 Visualization Analysis

We examine the interpretability of the proposed Joint Representation Gating (JRG) module. As shown in Fig. 2, the distribution of eye-tracking modality weights reveals distinct patterns between groups. TD samples are right-shifted (mean = 0.63), reflecting relatively stable gaze behavior, whereas ASD samples are left-shifted with a lower mean (0.58) and broader dispersion, indicating greater variability and increased reliance on head-pose cues. These findings support the instance-level adaptive weighting mechanism of JRG. Perturbation experiments provide complementary evidence for the robustness of JRG. As one modality is progressively degraded, its corresponding weight consistently decreases while the other increases, leading to only a mild performance drop. This behavior is consistent with the adaptive weighting pattern observed in Fig. 2, further supporting the reliability of the proposed mechanism.

To examine the structural impact of supervised contrastive learning ($\mathcal{L}_{\text{supcon}}$), Fig. 3 compares cross-modal feature alignment. Without contrastive regularization (i.e., $\mathcal{L}_{\text{ce}}$-only), embeddings exhibit a clear modality gap, forming clusters

organized by sensor type. After introducing $\mathcal{L}_{\text{supcon}}$, cross-modal features from the same subject align more closely, yielding more compact and separable ASD-TD clusters. This improvement is further reflected by the cosine similarity between eye-movement and head-pose features increasing from 0.68 to 0.98.

4 Conclusion

In this paper, we propose AMF-Net, a framework for mobile ASD screening under environmental noise and behavioral heterogeneity. By combining an instance-level Joint Representation Gating for adaptive eye-head fusion with a Supervised Contrastive Learning for structured feature alignment, the model effectively captures discriminative cues from unconstrained recordings. Experiments on a 124-subject cohort achieve state-of-the-art performance, reaching 91.93% accuracy and a 0.9206 F1-score. This work presents a novel mobile-oriented ASD screening framework that introduces subject-specific adaptive weighting for eye movements and head pose, achieving robust and accurate ASD identification in real-world noisy environments.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.
Code: https://github.com/JIESunshine/AMF-Net.

References

1. Bai, S.: An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271 (2018)
2. Campbell, K., Carpenter, K.L., Hashemi, J., Espinosa, S., Marsan, S., Borg, J.S., Chang, Z., Qiu, Q., Vermeer, S., Adler, E., et al.: Computer vision analysis captures atypical attention in toddlers with autism. Autism **23**(3), 619–628 (2019)
3. Chang, Z., Di Martino, J.M., Aiello, R., Baker, J., Carpenter, K., Compton, S., Davis, N., Eichner, B., Espinosa, S., Flowers, J., et al.: Computational methods to measure patterns of gaze in toddlers with autism spectrum disorder. JAMA Pediatr. **175**(8), 827–836 (2021)
4. Chen, S., Zhao, Q.: Attention-based autism spectrum disorder screening with privileged modality. In: Proc. IEEE Int. Conf. Comput. Vis. (ICCV). pp. 1181–1190. IEEE (2019)
5. Cheng, M., Zhang, Y., Xie, Y., Pan, Y., Li, X., Liu, W., Yu, C., Zhang, D., Xing, Y., Huang, X., et al.: Computer-aided autism spectrum disorder diagnosis with behavior signal processing. IEEE Trans. Affect. Comput. **14**(4), 2982–3000 (2023)
6. Curnow, E., Utley, I., Rutherford, M., Johnston, L., Maciver, D.: Diagnostic assessment of autism in adults–current considerations in neurodevelopmentally informed professional learning with reference to ados-2. Front. Psychiatry **14**, 1258204 (2023)
7. Han, J., Jiang, G., Ouyang, G., Li, X.: A multimodal approach for identifying autism spectrum disorders in children. IEEE Trans. Neural Syst. Rehabil. Eng. **30**, 2003–2011 (2022)

8. Han, L., Chen, X.Y., Ye, H.J., Zhan, D.C.: Softs: Efficient multivariate time series forecasting with series-core fusion. Adv. Neural Inf. Process. Syst. (NIPS) **37**, 64145–64175 (2024)
9. Han, Z., Zhang, C., Fu, H., Zhou, J.T.: Trusted multi-view classification with dynamic evidential fusion. IEEE Trans. Pattern Anal. Mach. Intell. **45**(2), 2551–2566 (2022)
10. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 4700–4708 (2017)
11. Jiang, M., Zhao, Q.: Learning visual attention to identify people with autism spectrum disorder. In: Proc. IEEE Int. Conf. Comput. Vis. (ICCV). pp. 3267–3276. Venice, Italy (2017)
12. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 7482–7491 (2018)
13. Khamis, M., Alt, F., Bulling, A.: The past, present, and future of gaze-enabled handheld mobile devices: Survey and lessons learned. In: Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services. pp. 1–17 (2018)
14. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Adv. Neural Inf. Process. Syst. (NIPS) **33**, 18661–18673 (2020)
15. Kollias, K.F., Syriopoulou-Delli, C.K., Sarigiannidis, P., Fragulis, G.F.: The contribution of machine learning and eye-tracking technology in autism spectrum disorder research: A systematic review. Electronics **10**(23), 2982 (2021)
16. 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. (CVPR). pp. 2176–2184 (2016)
17. Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., Long, M.: itransformer: Inverted transformers are effective for time series forecasting. In: Proc. Int. Conf. Learn. Represent. (ICLR) (2024), spotlight
18. 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)
19. Ma, H., Han, Z., Zhang, C., Fu, H., Zhou, J.T., Hu, Q.: Trustworthy multimodal regression with mixture of normal-inverse gamma distributions. Adv. Neural Inf. Process. Syst. (NIPS) **34**, 6881–6893 (2021)
20. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). vol. 32 (2018)
21. Perochon, S., Di Martino, J.M., Carpenter, K.L.H., Compton, S., Davis, N., Eichner, B., Espinosa, S., Franz, L., Babu, P.R.K., Sapiro, G., et al.: Early detection of autism using digital behavioral phenotyping. Nat. Med. **29**(10), 2489–2497 (2023)
22. 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. Biological Psychiatry **79**(8), 657–666 (2016)
23. Tecar, C., Chiperi, L.E., Iftimie, B.E., Livint-Popa, L., Stefanescu, E., Lucia, S.M., Draghici, N.C., Muresanu, D.F.: Eye-tracking as a screening tool in the early diag-

nosis of autism spectrum disorder: A systematic review and meta-analysis. J. Clin. Med. **14**(24), 8801 (2025)
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Adv. Neural Inf. Process. Syst. (NIPS) **30** (2017)
25. Wei, Q., Cao, H., Shi, Y., Xu, X., Li, T.: Machine learning based on eye-tracking data to identify autism spectrum disorder: A systematic review and meta-analysis. Journal of Biomedical Informatics **137**, 104254 (2023)
26. 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 (2022)
27. Yang, Z., Zhang, Y., Ning, J., Wang, X., Wu, Z.: Early diagnosis of autism: A review of video-based motion analysis and deep learning techniques. IEEE Access **13**, 2903–2928 (2025)
28. Zhang, S., Chen, D., Tang, Y., Zhang, L.: Children ASD evaluation through joint analysis of EEG and eye-tracking recordings with graph convolution network. Front. Hum. Neurosci. **15**, 651349 (2021)
29. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Appearance-based gaze estimation in the wild. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 4511–4520 (2015)
30. Zhong, W., Li, B., Xia, C., Li, K., Zhang, D.: Multi-modal progressive fusion for ASD screening using smartphone video. In: Int. Conf. Med. Image Comput. Comput.-Assist. Interv. (MICCAI). pp. 393–403 (2025)
31. Zhong, W., Xia, C., Zhang, D., Han, J.: Uncertainty modeling for gaze estimation. IEEE Trans. Image Process. **33**, 2851–2866 (2024)
32. 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 Trans. Neural Syst. Rehabil. Eng. **32**, 865–874 (2024)
33. Zhu, F.L., Wang, S.H., Liu, W.B., Zhu, H.L., Li, M., Zou, X.B.: A multimodal machine learning system in early screening for toddlers with autism spectrum disorders based on the response to name. Front. Psychiatry **14**, 1039293 (2023)