(Received 14 October 2025, Accepted 27 October 2025, Scheduled 28 October 2025)

Identifying Autistic Children Using Deep Learning Based on the Temporal and Spatial Information of Eye-Tracking

Deyu Guo[1, †], Yan Zhang[2, †], Tengfei Ma[1], Xinhua Zhu[3, 4], and Sailing He[1, 3, *]

[1] *National Engineering Research Center for Optical Instruments*
Center for Optical and Electromagnetic Research, Zhejiang University, Hangzhou 310058, China
[2] *Taizhou Hospital of Zhejiang Province Affiliated to Wenzhou Medical University, Taizhou 318050, China*
[3] *Shanghai Institute for Advanced Study, Zhejiang University, Shanghai 201203, China*
[4] *Taizhou Angility Smart Technology Ltd, China*

ABSTRACT: This study addresses the challenge of detecting Autism Spectrum Disorder (ASD) in children, where clinical diagnostic scales used in practice suffer from subjectivity and high costs. Eye tracking (ET), as a non-contact sensing technology, offers the potential for objective ASD recognition. However, existing studies often use specially crafted visual stimuli, making them less reproducible, or rely on the construction of handcrafted features. Deep learning methods allow us to build more efficient models, but only a few studies simultaneously focused on visual behaviors of ASD in both temporal and spatial dimensions, and many studies compressed the temporal dimension, potentially losing valuable information. To address these limitations, this study employed a more flexible visual-stimulus selection criterion to collect ET data of ASD in social scenes, enabling analyses to be conducted both temporally and spatially. Findings indicate that the spatial attention distribution of ASD is more dispersed, and gaze trajectories are more unstable in the temporal dimension. We also observed that children with ASD exhibit slower responses in gaze-following scenarios. Additionally, data loss emerges as an effective feature for ASD identification. We proposed an SP-Inception-Transformer network based on CNN and Transformer encoder architecture, which can simultaneously learn temporal and spatial features. It utilized raw eye-tracking data to prevent information loss, and employed Inception and Embedding to enhance the performance. Compared to benchmark methods, our model demonstrated superior results in accuracy (0.886), AUC (0.8972), recall (0.82), precision (0.95), and F1 score (0.8719).

1. INTRODUCTION

Autism Spectrum Disorder (ASD) is a common neurodevelopmental disorder in the pediatric population characterized by social difficulties, restricted interests, and repetitive stereotyped behaviors [1]. According to a study in 2021, the worldwide median prevalence of ASD is estimated to be 1% [2] and is gradually increasing. The gold standard diagnostic instruments are standardized validated assessments that measure the presence of autistic social disability through both behavioral observation and parent interview [3]. However, such approach is subjective, expensive and time-consuming, running the risk of causing patients to miss the optimal intervention period [3, 4].

The development of an objective ASD identification system is crucial for improving the quality of life for individuals with autism, as well as reducing societal and healthcare costs. Recently, eye-tracking (ET) technology has been explored for studying the visual attention characteristics of ASD [5] and developing objective diagnostic methods based on these findings. As a non-invasive and convenient technique, ET can help avoid diagnostic biases caused by information asymmetry between physicians and patients [6]. Individuals with ASD (ASDIs) usually show atypical visual attention on different kinds of visual stimuli [7], because eye movements can reflect information about individuals' attention, eye control, and psychological factors [7]. A study suggests that ASDIs, when viewing images that include people, tend to show less attention to the eyes and faces compared to Typically Developing (TD) children [8, 9]. When observing dynamic visual content, ASDIs tend to look more at the background rather than main subject [9]. Regarding the pattern of visual attention distribution, studies have found that ASDIs exhibit a more pronounced center bias [10]. The unique visual attention characteristics observed in ASD, possibly associated with their social impairments, offer a theoretical foundation for utilizing ET to identify ASD.

A main category of ET-based ASD recognition research is ROI (Region of Interest) analysis. This approach uses prior knowledge to divide regions that can distinguish ASD and extracts spatial features for subsequent classification algorithms. A recognition study based on ROI divided the visual attention of ASDI on the human body into small areas such as the face, eyes, and mouth. It was found that ASDI spent less time looking at these regions [6]. Other studies have presented multiple images simultaneously to ASDI, such as displaying both social and non-social content, and extracted the duration of ASDIs' gaze on both types of stimuli for classification [11, 12]. Recently, research indicates that the temporal distribution of visual attention in ASD differs from TD, and combining temporal information can enhance the accuracy of the algorithm [13]. However, studies based on ROI struggle to obtain rich temporal

[*] Corresponding author: Sailing He (sailing@zju.edu.cn).
[†] These authors contributed equally to this work.

doi:10.2528/PIERM25101403

features, and current research often requires carefully designed visual stimuli, making them complex to use and may not reflect the attention characteristics of ASD in real-life scenarios.

Building upon existing research, the development of deep learning technology has provided more powerful tools for ASD identification based on eye-tracking. Deep learning algorithms can handle more complex data and require less manual feature engineering to learn distribution patterns among samples. Some work in this area is centered around the Saliency4ASD competition [14], which used images close to natural scenes as visual stimuli and provided participants' fixation for classification. SP-ASDNet employs CNN and LSTM to process information in both spatial and temporal dimensions [15]. Another CNN-based approach involves feeding both stimulus images and sequences representing gaze positions into a CNN to obtain prediction results [16]. A research based on prior knowledge introduces a gaze-following prior map to assist the learning of CNN-LSTM model [7], which leverages the fact that ASDIs exhibit weaker gaze-following behavior. In a study using a different dataset, fixation maps were used as inputs to a CNN, which focuses solely on the spatial distribution of visual attention in ASD [17]. Other studies extracted Scan Path images [18], representing the temporal dynamics and spatial distribution of eye-tracking trajectories [19–21], which have shown promising results on their dataset. However, a key limitation of these approaches is that fixation points are compressed in the temporal dimension, and scan path images ignore visual information from raw stimuli, thus current methods are making insufficient use of the eye-tracking data.

The aim of this study is therefore twofold: to analyze the characteristics of ASD gaze behavior in both temporal and spatial dimensions, and to develop a deep network using popular Transformer architecture [22] that can better utilize temporal features for ASD identification. Our main innovation lies in:

- Unlike some previous studies that leveraged much medical prior knowledge, we employed a relatively lenient stimulus selection criterion during the ET data acquisition, only using the fact that ASD exhibit different visual attention patterns compared to TD in social scenes [7, 14];
- In the data processing stage, we allowed for data loss rather than excluding these samples, as this might be related to ASD's inherent characteristics [23];
- Qualitative and quantitative analyses were conducted simultaneously and differences between visual behaviors of ASD and TD were observed. We found that ASDIs' attention was more dispersed, lacking a consistent template, and they showed a slower response in gaze-following processes;
- We constructed a deep learning network based on raw ET data, which has not been fully investigated in ASD research yet, and incorporated visual information from the original stimuli. We utilized the Inception [24] CNN structure to model spatial attention and employed a Transformer encoder to study temporal attention over longer time sequences, realizing more effective identification of ASD. The proposed model could be used in other ET researches if raw data is provided.

2. DATA ACQUISITION AND PROCESSING

2.1. Data Acquisition Experiment

We recruited a total of 60 children with ASD and 63 typically developing (TD) children as participants. Five ASD and one TD were excluded due to incomplete calibration processes, while five ASD and six TD were excluded during subsequent data analysis procedure due to low quality data. Therefore, the effective sample size for this study was 50 ASD and 56 TD. Participants in the ASD group were recruited from two local autism rehabilitation institutions (namely, Youai Autism Rehabilitation Institution and Xingyu Autism Rehabilitation Institution in the county of Wenling) through the help of the Children's Rehabilitation Department of Taizhou Rehabilitation Hospital. Inclusion criteria were:

- Age between 4 and 10 years;
- Normal or corrected-to-normal vision;
- A professionally issued diagnosis of ASD;
- The absence of other neurodevelopmental disorders such as mental retardation, cerebral palsy, no history of epilepsy, and no chronic illnesses.

Due to the unavailability (license issue) of the Autism Diagnostic Observation Schedule (ADOS) [25] in China, all children in the ASD group were diagnosed by professional physicians using the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) [26]. The TD group consisted of children recruited from the community as controls, sharing the same criteria as the ASD group except for the absence of autism. The statistical characteristics of the participants are as shown in Table 1. Both groups of participants are relatively close in terms of age and gender distribution.

TABLE 1. Statistical characteristics of the valid participants.

Group	Count	Age(mean ± SD)	Gender(boys/girls)
ASD	50	7.21 ± 1.47	39/11
TD	56	7.08 ± 1.36	42/14

To enhance the practical applicability of our approach, unlike some previous ET-based autism studies that used specially filmed or created visual stimuli, our research has loose criteria for the stimuli selection. We opted for screenshots from publicly available film clips (www.iqiyi.com) to reduce the accessibility challenges of recognition method. Specifically, images contained social stimuli with characters (single or multiple) interacting with environment or objects (such as looking at billboards, eating, taking photos) or engaging in communication with others. This approach ensured that these images include complex scenes that are closer to real-life situations and exclude any content that could cause dis-comfort for children.

We prepared a total of 60 different social images (RGB, 1920 × 1080) as visual stimuli for this study. The data collection experiment took place in a quiet room, where participating children sat approximately 60 cm away from a 24-inch monitor with a resolution of 1920 × 1080, as shown in Figure 1.

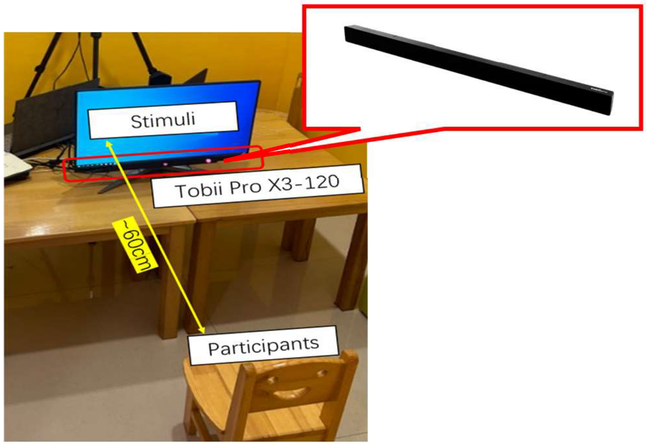

FIGURE 1. Data acquisition environment. A Tobii eye-tracker was fixed on the bottom of the screen to detect the gaze positions of participants. A zoomed-in view of the Tobii Pro X3-120 eye-tracker is provided in the inset. The device uses near-infrared illumination to create reflection patterns on the corneas of both eyes, which are captured by image sensors. These patterns are used to compute the gaze point on the screen based on calibration data.

The Tobii Pro X3-120 eye-tracking equipment was used to collect eye movement data at a frequency of 120 Hz. Before the experiment began, all participants completed a 5-point eye tracker calibration, followed by a validation process. An animation of a small cat appeared successively at different positions on the screen, and participants were instructed to focus on it to calibrate the eye tracker. This process was repeated necessarily until the calibration results met the requirements. To prevent the children from losing patience, we divided the 60 images into three sets and presented them sequentially to the participants. Each image was displayed on the screen for 3 seconds, with a 3-second gray background between consecutive ones. We used Tobii Pro Lab (www.tobii.com) software for data collection and exported raw data for subsequent processing.

All children in this study were accompanied by their parents or teachers, and written informed consent was obtained before the experiment. *The study was approved by the Ethics Committee of the Affiliated Hospital of Hangzhou Normal University (2023 (E2) -KS- 128).*

2.2. Qualitative Analysis on Visual Attention

We aggregated the data from all participants, obtaining visual attention distribution maps for each group across different stimuli, as shown in Figure 2. We marked the locations of faces in the stimuli with white ellipses and important objects being used or observed by characters with white triangles.

We observed that, for most stimuli, ASDIs' visual attention distribution was more dispersed, while TDIs' attention tended to be more focused with a few noticeable attention areas. This indicates that TD children exhibit stronger spatial attention consistency when observing social scenes, while ASD children have more diverse attention, which is similar to the point made in previous research that ASD adults exhibit within-group heterogeneity [27]. In Figure 2, TDIs' visual focus is on faces and corresponding objects, while ASDIs, although looking at these areas as well, also directs attention to regions outside. This phenomenon becomes more pronounced when the stimulus contains a complex background as interference or includes many characters, as shown in Figure 2(b). When the main character is positioned to one side of the image, such as Figure 2(c), ASD not only looks at the characters but also has a considerable amount of attention distributed in the center which lacks actual semantic information, therefore ASD may exhibit a more pronounced center bias in certain scenes.

However, spatial distribution information alone may be not sufficient for identifying ASD when the stimulus is relatively simple. As shown in Figure 2(d), the stimulus includes a character looking at an object in his hand, resembling a joint attention scenario. Research suggests that ASD exhibits behaviors distinct from TD when observing such scenes, such as reduced attention to the character's eyes [7]. Some studies even indicate differences in brain activity during joint attention tasks for individuals with ASD [28]. The spatial attention distribution for ASD is quite similar to that for TD in Figure 2(d), focusing on the face and the object. The previously mentioned characteristic of ASD having more dispersed attention is not evident here. However, when we unfold the data across the time dimension and analyze participants' viewing behavior within $0 \sim 1$ s, significant differences emerge during gaze following actions, as depicted in Figure 3.

In the initial (0.0, 0.25 s) phase, participants from both groups exhibited relatively random attention. In the (0.25, 0.5 s) phase, almost all participants from both groups noticed the face and directed their attention towards it. In the (0.5, 0.75 s) phase, TD children identified the gaze of the character and follow it to focus on the corresponding object, making it the center of attention, with less focus on the face. At the meantime, ASD children have not effectively completed the gaze following action, and the majority of their attention remained on the face. In the (0.75, 1.0 s) phase, TD children started to re-focus on the face, resulting in a more evenly distributed pattern. In contrast, ASD children exhibited behaviors similar to TD's previous phase. This suggests that ASDIs' gaze following procedure is slower than TD group, possibly related to their social challenges in real life.

In summary, there are spatial differences in visual attention distribution between ASD and TD. However, sometimes relying solely on spatial differences may not be sufficient to distinguish ASD, and introducing the time dimension is essential. Therefore, building algorithms that simultaneously consider spatial and temporal information is necessary for ASD identification.

2.3. Data Preprocessing

The raw eye-tracking data we collected cannot be directly used for classification because it contains some missing values (Null). Missing values may occur when participants experience significant head movements during the experiment, leading to the failure of the eye tracker. Alternatively, they may result

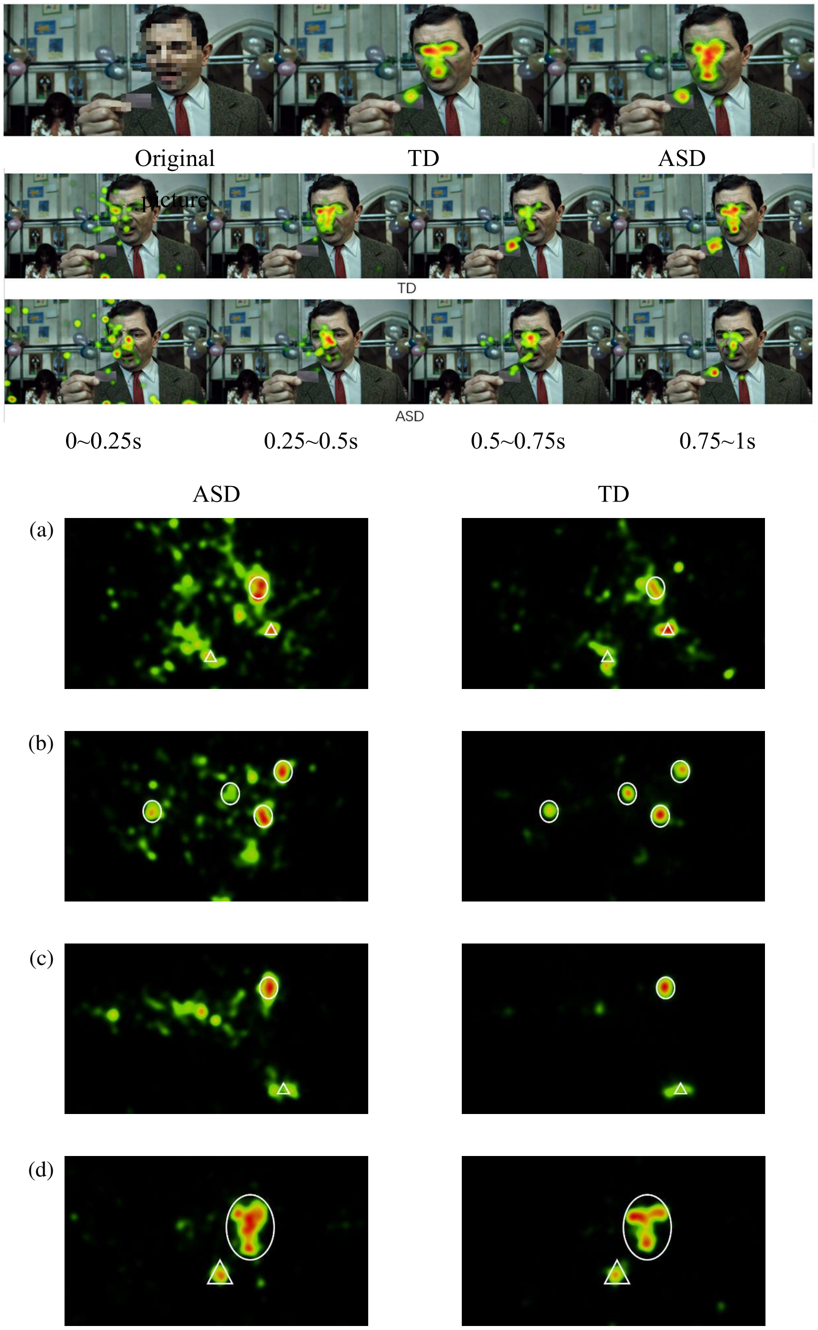

FIGURE 2. Accumulated visual attention distribution from eye-tracking (ET) for a child with Autism Spectrum Disorder (ASD) and a typically developing (TD) child in response to a scene from the movie *Mr. Bean's Holiday* (Universal Studios, 2007). The original stimulus image has been intentionally blurred in this article to mitigate copyright concerns, and its use falls under fair use for academic research and critique. The gaze trajectories over time (0–1 s) reveal differences in the temporal dynamics of attention to the card held by the character. Color indicates fixation density, with warmer colors (e.g., red) representing higher density and cooler colors (e.g., blue) representing lower density. For (a)–(d), the left column represents ASD, and the right column represents TD. The white ellipse in the image indicates the presence of a face in this area, while the white triangle indicates an object interacting with the character. Fixation density is color-coded, with warmer colors (e.g., red) representing higher density and cooler colors (e.g., blue) representing lower density. Each row corresponds to the stimulus content as follows: (a) single character with a complex background, (b) single character with a simple background, (c) multiple characters, (d) single character looking at an object in his hand.

from blinking or occasionally looking outside the screen, causing the eye tracker to miss capturing valid gaze positions. The first type of technical data loss should be discarded, while the second type of data loss is allowed in our study as it may be a crucial feature in distinguishing between ASD and TD [23]. The first step in data preprocessing is to fill null values with zeros, and the overall preprocessing flowchart is shown in Figure 4.

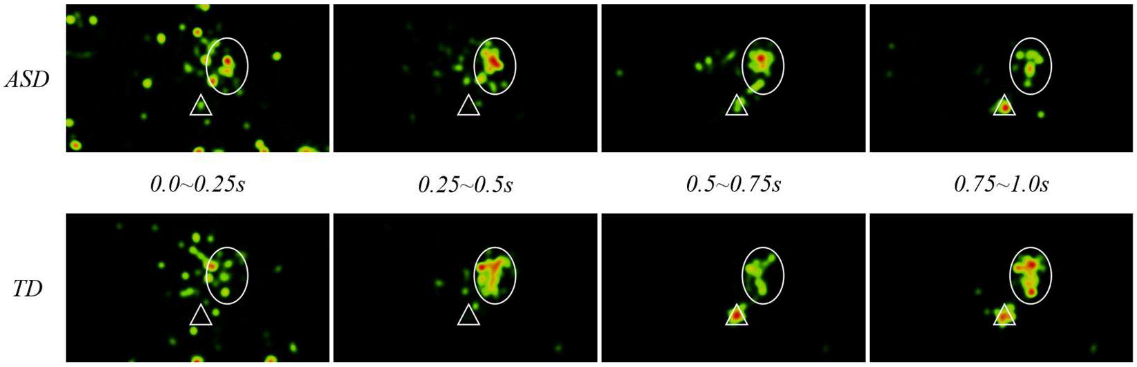

FIGURE 3. The temporal evolution of visual attention distribution of ASD and TD on the same stimulus. The top row represents ASD, and the bottom row represents TD. From left to right, each column corresponds to a 0.25 s time window. The white ellipse in the image indicates the presence of a face while the white triangle indicates an object interacting with the character.

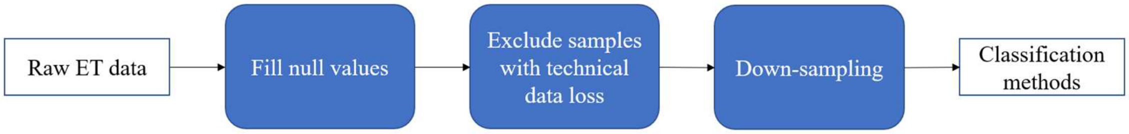

FIGURE 4. Flowchart of data preprocessing step.

To eliminate samples with severe technical data loss, which is typically characterized by the alternating occurrence of valid coordinates and null data, our approach is as follows: We define a segment of data as abnormal when its effective length is less than 5. We then count the occurrences of abnormal data segments for all participants, considering participants with more than 2000 abnormal segments to have severe technical loss, and we remove these samples (5 ASD and 6 TD) from the dataset.

To reduce the data volume for easier processing and training, we down-sampled the data to 60 Hz, which is sufficient for studying ASD eye gaze behavior. For each participant, the length of the down-sampled data is $3\,\mathrm{s} * 60\,\mathrm{stimuli} * 60\,\mathrm{Hz} = 10800$ data points, with each data point representing an (x, y) coordinate, where both x and y are floating point numbers between 0 and 1. The down-sampling operation is done as:

$$\mathrm{Data}_{60,i} = \begin{cases} \mathrm{Data}_{120,i} & \text{if } \mathrm{Data}_{120,i} \text{ or } \mathrm{Data}_{120,i+1} = 0 \\ \frac{\mathrm{Data}_{120,i}+\mathrm{Data}_{120,i+1}}{2} & \text{else} \end{cases} \tag{1}$$

The shape of the entire eye-tracking dataset is (106, 60, 180, 2). Differing from previous studies, we will use the ET sequences to build the classification network, rather than fixations. Extracting fixations can aggregate eye-tracking data into several gaze points, reducing the length in temporal dimension, but may result in the loss of valuable information. Suitable deep learning algorithms allow us to input raw time-series data and extract useful features directly.

3. CLASSIFICATION METHODS

To validate the validity of our data and demonstrate the superiority of the proposed model, we selected two existing deep learning-based methods and several traditional machine learning methods as benchmarks.

3.1. Benchmark 1: Scan Path Method

The Scan Path method [18] is a data processing technique widely used in ASD Eye-Tracking research recently. It involves converting raw ET data into RGB images, which can represent the movement trajectory of gaze points and provide information about the speed of gaze movement over time. The steps to obtain Scan Path images are as follows:

- Calculate the speed, acceleration, and jerk (rate of change of acceleration) between every two data points and scale them by dividing one-fourth of the diagonal length of the stimulus image;

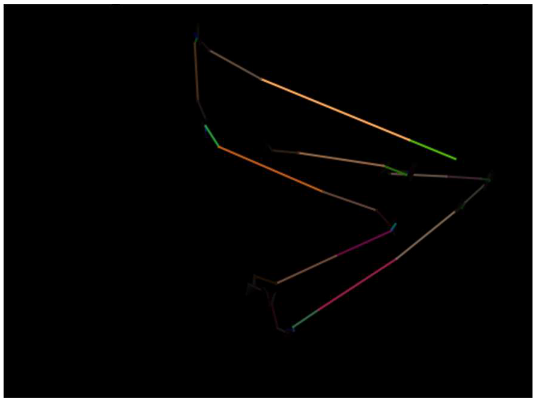

FIGURE 5. An example of Scan Path images generated from our dataset.

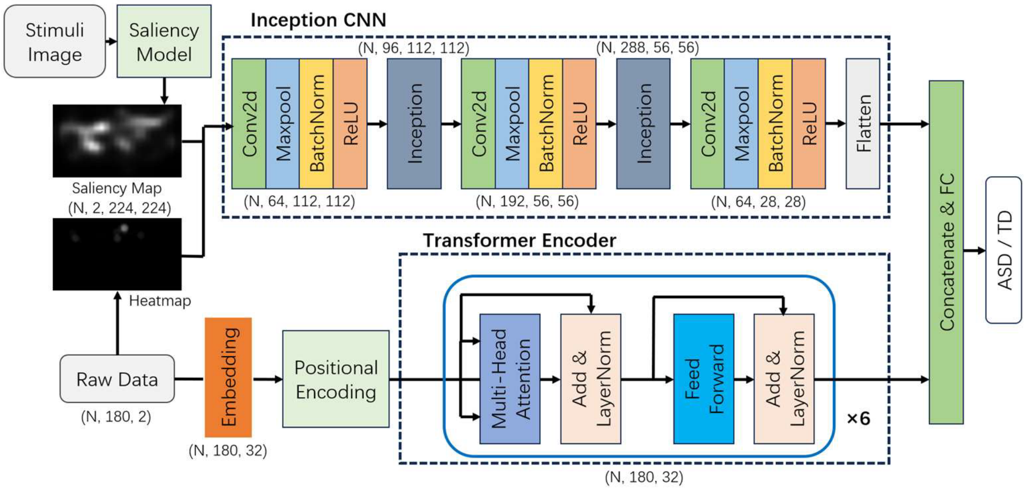

FIGURE 6. The overall framework of SP-Inception-Transformer network.

- Draw a line between two data points, and the color (R, G, B) of the line is determined by the corresponding speed, acceleration, and jerk values.

An example of extracted Scan Path image is shown in Figure 5.

We adopted a Convolutional Neural Network (CNN) structure proposed in a previous study [19] that demonstrated high classification performance for classification. Concretely, it is a simple four-layer CNN where the feature map's scale is reduced by half, and the channel count is doubled in each layer. The output of the final convolutional layer is flattened and fed into a fully connected network for classification. Additionally, this method preprocesses the input Scan Path image by converting it to Grayscale and resizing it to 100×100 shape.

3.2. Benchmark 2: SP-ASDNet

SP-ASDNet (SA) [15] is a CNN-LSTM model that considers both spatial and temporal dimensions, although the temporal dimension is compressed.

In the spatial feature extraction part, this method initially employs a pre-trained Saliency prediction model to extract a saliency map from the original stimuli images, describing the degree of attraction for each pixel to the observer. Recently, saliency prediction models based on deep neural networks have seen rapid development, demonstrating superior performance compared to traditional methods [29]. In the temporal dimension, the raw data undergoes filtering and compression to obtain a series of fixation coordinates along with the duration. Based on fixations, patches are cropped from the saliency map and we followed the original study's setup to set the patch size as 225. Patches are then input into a CNN to extract 1-D spatial features. The paper did not specify the activation function used after the Convolutional layers and we chose the commonly used ReLU (Rectified Linear Unit). As the used LSTM (Long Short-Term Memory) [30] network only accepts fixed-length inputs, the next step involves concatenating all features for each stimulus and padding them with zeros to a uniform length L. The duration is concatenated with spatial features to ensure that the data input to LSTM includes both spatial and temporal information.

3.3. Traditional Machine Learning Methods

To quantitatively compare the differences in gaze behavior between ASD and TD, we extracted a series of features from raw data and performed classification tasks using traditional ML methods. Feature extraction is conducted on a stimulus-by-stimulus basis, and the total features for each participant are the averages or sums of the data on each stimulus. Specifically, the features extracted were:

- Mean gaze position (on x and y direction).
- Distance of gaze positions to the screen center (Center Bias).
- Standard deviation of gaze coordinates (on x and y direction).
- Longest fixation duration.
- Number of fixations.
- Amount of data loss.
- Average eye movement speed.
- Average eye movement acceleration.
- Average eye movement jerk.

Distance of gaze positions to the screen center reflects the participants' tendency of center bias, while Standard deviation of gaze coordinates describes the dispersion of visual attention. Amount of data loss encompassing the quantity of zero data resulting from participants blinking or not looking at the screen. All mentioned features are extracted from the down-sampled 60 Hz data, so the unit of duration is 1/60 s. Traditional ML methods used for classification were:

- Support Vector Machine (SVM).
- Random Forest (RF).
- Logistic Regression (LR).
- K-nearest neighbors (KNN).

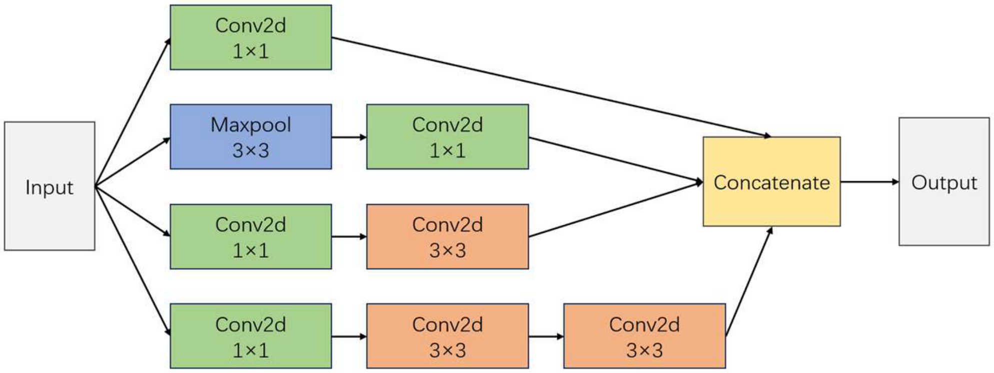

FIGURE 7. The structure of an Inception block. Batch Normalization and ReLU are not shown in this picture but are added after convolution in our model.

3.4. SP-Inception-Transformer Network

We proposed a deep network called SP-Inception-Transformer (SIT), which combined a CNN using Inception blocks and a Transformer encoder structure. A pretrained deep saliency prediction model SALICON was included to extract spatial information from the raw stimuli. The overall framework is shown in Figure 6. As mentioned above, the input raw coordinates data has shape $(N, T, 2)$, where N denotes batch size, and T represents the number of frames, which in our case is 180 after down-sampling.

3.4.1. Generating Saliency Maps and Heatmaps

We utilized the well-established SALICON model [31] to generate saliency maps, which has been pre-trained on a large dataset and is relatively easy to use. Saliency map can be viewed as highlighting the visual interest information in the original image while reducing interference from other details.

To compare participants' spatial visual attention distribution with saliency map, we extracted heatmaps from ET data as the following procedure:

- Initialize an array of zeros with the same size (h, w) as the original stimulus.
- Iterate through the ET data. For each fixation point (x, y), increment the pixel values within a radius $r = 0.05 * h$ around that point by 1.
- Scale all pixel values to the range (0, 255):

$$\text{Pixel}_{0\sim255} = 255 \times \frac{\text{Pixel}}{180} \tag{2}$$

In heatmaps, higher pixel values indicate stronger attention from participants to the corresponding positions.

3.4.2. Spatial Modeling with Inception CNN

The saliency map and the heatmap obtained in the previous step were concatenated channel-wise and fed into the CNN, ultimately producing a 1D spatial feature.

We used a five-layer structure CNN, including three standard 2D convolution operations and two inception blocks. Each standard convolution had a kernel size of 3, a stride of 1, and padding of 1, meaning the convolution did not change the size of the feature map. After each convolution, we applied a maxpool2d layer with a kernel size of 2, a stride of 2, reducing the size of the feature map by half to extract dominant features. The number of channels for each convolution layer is annotated in Figure 8, being 64, 192, and 64, respectively. Additionally, we added Batch Normalization after each convolution to suppress overfitting and enhance the stability of the network [32]. A dropout layer with rate 0.3 is added after each convolution block.

To allow the network to consider spatial information at multiple scales simultaneously, we introduced inception blocks, whose structure is shown in Figure 7. The original Inception structure included convolution kernels of sizes 1×1, 3×3, and 5×5, concatenating the features extracted by these different-scale convolution kernels and returning them to the next layer [24]. The 1×1 convolution in the figure serves to reduce dimensionality along the channel dimension, reducing model complexity. Inception V2 improved this structure by introducing Batch Normalization and replacing the original 5×5 convolution with two 3×3 convolutions, thereby reducing the number of parameters [33].

3.4.3. Temporal Modeling with Transformer Encoder

Since the introduction of the Transformer model, its derived structures have shown excellent performance in both the NLP and CV fields [34]. The original Transformer model consists of two parts, Encoder and Decoder. Since our task was encoding time series, we only used the Encoder part. The structure of each layer is depicted in Figure 6 and mainly includes two operations: Multi-head self-attention and feedforward (fully connected), along with corresponding Layer Normalization. The encoder includes several blocks of this type, and the number of layers is 6 in our model.

Self-attention mechanism is the key part of Transformer, enabling it to handle longer sequences, while multi-head self-attention allows the model to flexibly focus on information at different timestamps. Specifically, given an input sequence $X \in R \wedge (n \times d)$, where n is the sequence length, and d is

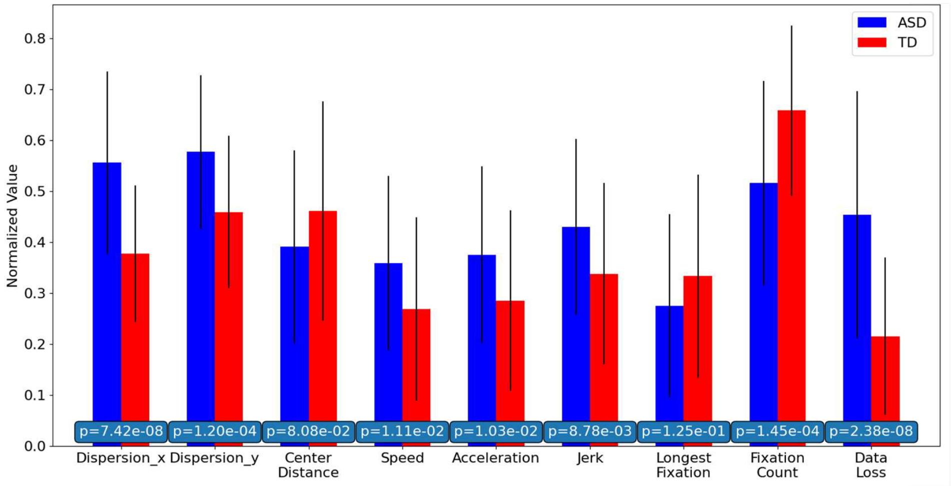

FIGURE 8. Comparison of spatial and temporal statistical features between ASD and TD. Values were normalized to $0 \sim 1$ for better visualization.

the dimensionality of each element, for the ith attention head, multi-head self-attention first calculates corresponding queries q_i keys k_i, and values v_i for each element z_i as:

$$Q_i = X \cdot W_i^Q \tag{3}$$

$$K_i = X \cdot W_i^K \tag{4}$$

$$V_i = X \cdot W_i^V \tag{5}$$

where W_i^Q, W_i^K and W_i^V are the linear transformation weight matrixes associated with the ith attention head. For each position j, the ith head's attention score is calculated as:

$$\text{Attention}_i(x_i, x_j) = \frac{\left(Q_i \cdot K_i^T\right)_{ij}}{\sqrt{d_h}} \tag{6}$$

where $(Q_i \cdot K_i^T)_{ij}$ is the (i, j)th element of the matrix, and d_h is the dimension of attention heads. The attention weight at each position is normalized attention score using softmax as:

$$\alpha_{ij}^i = \frac{\exp\left(\text{Attention}_i(x_i, x_j)\right)}{\sum_{k=1}^{n} \exp\left(\text{Attention}_i(x_i, x_k)\right)} \tag{7}$$

And the output of the ith head is:

$$H_i = \sum_{j=1}^{n} \alpha_{ij}^i \cdot V_{ij} \tag{8}$$

The outputs of all heads are concatenated to obtain information about all the elements in the input sequence. However, there is no positional information included in self-attention mechanism, so the distance between every two elements is always the same. Transformer addressed this issue by adding a positional encoding operation, which is done as:

$$\text{PE}_{\text{pos},2i} = \sin\left(\frac{\text{pos}}{5000^{2i/\text{dim}}}\right) \tag{9}$$

$$\text{PE}_{\text{pos},2i+1} = \cos\left(\frac{\text{pos}}{5000^{2i/\text{dim}}}\right) \tag{10}$$

where pos is the positional index in a sequence, and $2i$ indicates the index of positional encoded vector, and dim is the dimension of this vector. Following the original Transformer paper [22], PE has the same dimension as input and is directly added on it.

In our case, the dimension of input is 2, which is relatively small, limiting the Transformer model's capability in representing the temporal dimension. To increase the dimension of our input as well as the hidden size in Transformer encoder, an embedding layer is added before positional encoding procedure, which is a fully connected layer without activation that extends the dimension to 32 while keep the original feature distribution. Several different embedding sizes have been tested in the final ablation experiment to find the best architecture.

The temporal output is flattened and concatenated with the spatial feature. A fully connected network with 1 hidden layer and dropout rate 0.3 is used for the final prediction. Batch size was set to 64 to accelerate training process. We chose different learning rate for spatial and temporal branch, because CNNs tend to converge faster when training data is sufficient, while Transformer models are usually difficult to train. Learning rate was set to 1e-5 for spatial branch and 3e-4 for temporal branch, which in our experiments got the best result.

4. RESULTS AND DISCUSSION

4.1. Comparison of Gaze Behavior between ASD and TD

To quantitatively validate our qualitative observations, we plotted the mean and standard deviation of each manually extracted feature in Section 3.3, along with the p-values between the two groups. The comparison is shown in Figure 8.

The most notable finding is that the dispersion of gaze points in both x and y directions for ASD is significantly higher than TD group, which is consistent with our observations in Section 2.2. We hypothesize that this occurs because individuals with ASD are more likely to be attracted by background elements when viewing social scenes and that stimuli we used have complex elements with potentially more points of interest in the background, suggesting that ASD individuals do not focus primarily on the characters. This may indicate that different ASD participants have distinct personal preferences for specific objects in the background, which may be related to the color, shape, semantic information of objects, and the individual's upbringing environment.

We also found that ASD individuals have a stronger center bias compared to TD (i.e., gaze points are closer to the center), but with $p = 0.08$, indicating that this difference is not significant. As observed in Section 2.2, a significantly stronger center bias in ASD was evident when the character in the image was on one side, but it was not apparent in other stimuli. Since our statistical results were averaged across all stimuli without weighting based on stimulus type, it led to a weaker center bias feature.

Furthermore, our analysis revealed that data loss in ASD is significantly higher than in TD, indicating that ASDIs blink and look away from the screen more frequently. This suggests that the attentional focus of ASD is poorer when viewing social scenes, possibly related to their social impairments, which aligns with existing research [23]. Additionally, we observed that fixation appearance in ASD is less than TD. This could be attributed to the fact that ASDIs spend less time viewing stimuli and may focus on only a few specific locations, thus not effectively capturing the information contained in the entire image within a 3-second time window. However, there were no significant differences in the longest fixation duration between two groups.

Regarding the motion attributes of gaze, ASDIs showed significantly higher speed, acceleration, and jerk compared to TDIs. Combining this with the fixation feature and spatial features mentioned earlier, we believe that the gaze behavior of ASDIs has such characteristics: there are more possible points of interest for ASDI in a social stimulus because of the complex visual elements and background, but individual will only focus on a few of them, and their gaze shifts between different positions faster and more unstable. This may be associated with the muscle control ability for eye movements in ASDIs, as higher acceleration and jerk values indicate greater variability in eye movement.

4.2. Classification Experiment Setup and Evaluation Metrics

We attach more importance to the system's recognition performance on participants (each with 60 stimuli), therefore, during the validation phase, we take the average of all outputs across 60 visual stimuli from the same participant to get the final prediction.

To objectively compare the performance of different methods, we employ a 5-fold cross-validation strategy, calculating the average performance to evaluate different methods. In addition, to prevent data leakage, we initially split the dataset into training and validation sets according to participants, and further divide these two parts based on visual stimulus, ensuring that data belonging to the same participant does not simultaneously appear in both training and validation sets.

Evaluation metrics used in our study are Accuracy, AUC (Area Under the Curve), Recall, Precision, and F1-score. The definitions are as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{11}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{12}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{13}$$

$$\text{F1 score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{14}$$

where TP is true positive, TN the true negative, FP the false positive, and FN the false negative. In addition, AUC is the area under the Receiver Operating Characteristic (ROC) curve, measuring the model's performance at different thresholds, ranging from 0.5 to 1, where 1 indicates a perfect classifier, and 0.5 represents random classification.

All programs are implemented with Python 3.8 and PyTorch. The deep learning algorithms are trained and validated with a Nvidia GeForce RTX 2080Ti.

4.3. Comparison between Different Classification Methods

The validation results of all methods are shown in Table 2, where the ranking is based on accuracy from highest to lowest.

TABLE 2. Performance of classification models (averaged over 5-fold cross-validation).

Methods	Accuracy	AUC	Recall	Precision	F1-score
SIT	0.887	0.8972	0.82	0.95	0.8719
Scan Path	0.8303	0.8812	0.74	0.9014	0.7959
SVM	0.7588	0.7474	0.64	0.7989	0.7054
SA	0.7541	0.8253	0.62	0.8833	0.6909
LR	0.7273	0.7539	0.64	0.7456	0.6822
RF	0.6701	0.7355	0.62	0.6638	0.6326
KNN	0.6619	0.6723	0.64	0.6514	0.6325

As summarized in Table 2, the SP-Inception-Transformer model achieved the best performance in terms of accuracy, AUC, recall, precision, and F1 score, indicating a better capacity in modeling spatial and temporal information. Notably, the second-ranking method is the Scan Path approach. We attribute its lower performance to the fact that it did not utilize any information from the stimuli when constructing the Scan Path map. This limitation is significant because knowing which parts of the stimuli are worth attention and observing whether ASDIs focus on these theoretically interesting areas could aid identification. Additionally, the Scan Path image primarily focuses on the movement trajectory information of gaze, such as speed and acceleration, but cannot represent duration. This limitation arises from the original construction of the Scan Path, where if a participant stares at a point without moving, the RGB values for that point should be very small, appearing close to black on the graph. Duration's length has almost no impact on the color.

SP-ASDNet showed lower score than the first two methods on our dataset, achieving about 75% classification accuracy and an AUC of 0.82. We attribute this to the fact that, although SP-ASDNet considers both spatial and temporal dimensions, extracting fixations from ET data is not an optimal choice for our dataset. Since we allow for data loss, the total number of fixations on a stimulus may be limited. Data loss may occur between two fixations, but the model can only infer loss based on durations less than 3 seconds without knowing where the loss occurred. The fixation extraction process itself ignores the motion details between two points, preventing the network from knowing the participants' eye movement speed and variability. As discussed in Section 4.1, these properties are essential for distinguishing ASD. Additionally, this model chose LSTM to encode the temporal dimension, but LSTM requires a fixed input length. Padding the input sequence with zeros is necessary because each stimulus has a different number of fixations. Since LSTM is based on recurrent neural network and needs to propagate in the time dimension [30], too many zeros may cause a loss of previously learned data, thereby disrupting the model's learning process and resulting in inferior result [35].

Among traditional machine learning methods, SVM achieved the best performance with an accuracy of 0.75 and an AUC of 0.74. This suggests that the selected features can effectively distinguish between ASD and TD to some extent. However, the AUC and precision of these methods are significantly lower than three deep learning-based approaches, indicating that current features may lead the model to more frequently misclassify TD as ASD. We believe this could be due to the insufficient richness in feature selection, where the behavior of TD and ASD appears similar in some features. However, constructing richer and more effective features might depend on additional prior knowledge, potentially involving the delineation of ROI to capture the gaze behavior in specific subregions such as the face, which undoubtedly requires substantial effort, while deep learning-based methods do not necessitate such steps.

The confusion matrix for SP-Inception-Transformer model is illustrated in Figure 9. A particularly important observation is that among the individuals predicted as ASD, only 2 out of 43 were TD. This indicates that our model exhibits a relatively high precision, tending to be stringent in diagnosing ASD. However, 9 out of 50 ASD samples were misclassified as TD, resulting in a relatively lower recall. Examining the results in Table 2, it's evident that the recall performance of all models is not as ideal as precision. This pattern suggests that there are instances in our dataset where the visual behavior of some ASDIs closely resembles that of TDIs. Given that ASD might be an umbrella term for various subtypes, with some individuals exhibiting behaviors like TD in certain aspects. For instance, some high-functioning autistic children may possess unique talents in mathematics. To build a more advanced recognition system, future research should consider categorizing ASD into several subtypes in the medical field.

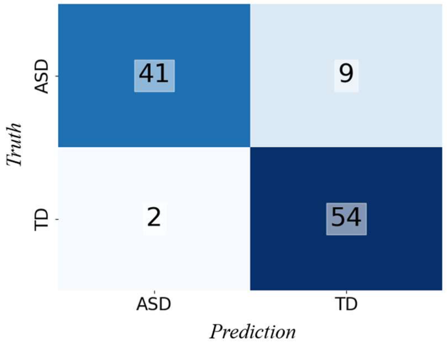

FIGURE 9. The confusion matrix of proposed network.

4.4. Ablation Experiment Results

To investigate the importance of utilizing both spatial and temporal information and prove the reliability of proposed SP-Inception-Transformer model, some ablation experiments were conducted on 5-fold cross-validation. The results are listed in Tables 3–5.

TABLE 3. Performance of branches (averaged over 5-fold cross-validation.

Methods	Accuracy	AUC	Recall	Precision	F1-score
Spatiotemporal	0.887	0.8972	0.82	0.95	0.8719
Temporal	0.8493	0.9015	0.76	0.9096	0.82
Spatial	0.8301	0.8769	0.72	0.9228	0.8048

TABLE 4. Results with different embedding size (averaged over 5-fold cross-validation).

Embedding	Accuracy	AUC	Recall	Precision	F1-score
-	0.6995	0.7621	0.48	0.6667	0.5392
16	0.8112	0.8619	0.72	0.8814	0.7798
32	0.8493	0.9015	0.76	0.9096	0.82
64	0.8402	0.8948	0.76	0.8825	0.8148
128	0.8489	0.8975	0.82	0.8595	0.8317

TABLE 5. Results with and without PE (averaged over 5-fold cross-validation).

PE	Accuracy	AUC	Recall	Precision	F1-score
Yes	0.8493	0.9015	0.76	0.9096	0.82
No	0.7536	0.7957	0.52	0.7355	0.6101

We first split the proposed network into two independent branches and trained them separately. As shown in Table 3, using only spatial or temporal branch got lower scores than the spatiotemporal method. Spatial or temporal dimension individually can only provide part of the information, which can be explained by our previous analysis. However, the results were competitive compared to benchmarks, proving that our proposed network had the ability of extract useful temporal and spatial features.

Embedding is important for Transformer based model. The temporal branch with no embedding showed poor results on our dataset, with accuracy of 69.95% and recall of 0.48. Adding an embedding layer before encoder significantly improved model's performance, as shown in Table 5. When embedding size increased from 16 to 32, the performance of the Temporal-only model got obvious improvement, while larger embedding sizes including 64 and 128 had brought no significant changes. We believe that the bottleneck of increasing embedding size lies in the limited information in 2-dimensional raw data. The embedding operation we utilize is a fully connected layer with no nonlinear activation, which is a linear combining operation of the original data. For a 2-d (x, y) coordinate, the meaningful combination is finite.

Positional encoding is another important factor that determines model's performance. Transformer with positional encoding outperformed the one with no positional information by about 10% accuracy. ASD showed different gaze behavior, as discussed before, and might have special gaze orders, which were missed in no PE model. Besides, applying PE enables Transformer to capture more temporal features, such as speed and acceleration between two gaze points, thus learning the instability of visual behavior, while model with no PE can only know where participants look.

5. CONCLUSIONS

In this study, we collected ET data from individuals with ASD and TD in social stimuli. Qualitative and quantitative analyses of their visual behaviors were conducted in both temporal and spatial dimensions. An SP-Inception-Transformer classification network was proposed that took both these two aspects into consideration more effectively.

In summary, this study demonstrates that even with a relatively lenient stimulus selection criterion, we were able to identify ASD, thereby reducing the application difficulty of ET-based systems. **The key findings reveal that** compared to TD, ASDIs exhibit a much more dispersed attention distribution, lacking consistency in visual patterns across different individuals. We believe that complex social images have more interest points for ASD children, which are often non-human elements in the background, and the primary visual focus of ASD is not concentrated on the characters within a short viewing time. However, the degree of attention to these points of interest varies among ASDIs, which may be associated with individual preferences and living environments. Moreover, in certain specific scenarios, the spatial attention distribution of ASD and TD is quite similar, but the visual behavior differs temporally. We observed that in some cases of gaze-following, ASDIs showed a slower response to the gaze of character, which indicate that they may have a relatively low efficiency while understanding social cues, thus experience communication difficulties. The gaze movement of ASD is more unstable, which may be related to their control over eye muscles and nerves. Notably, data loss (blink and looking away from the screen) proves to be a useful feature in distinguishing between ASD and TD, as ASD often have poor concentration on social scenes.

Our proposed SP-Inception-Transformer network can simultaneously learn spatial and temporal information from raw eye-tracking data, minimizing the loss of useful information. The using of Inception block enables us to extract features that contain spatial information at multiple scales, so the network can learn whether the ASD gaze positions are significant in the saliency map and whether these positions are the main areas of interest. We utilized Transformer encoder to model temporal information instead of widely used LSTM, making the raw sequence data as input possible. Compared to benchmarks, our model achieved the best performance in accuracy (0.886), AUC (0.8972), recall (0.82), precision (0.95), and F1 score (0.8719). Single spatial or temporal branch achieved lower results, proving the importance of leveraging information from both two aspects. However, results of branches from our model are also acceptable, making them useful in some low computational resource conditions. Due to that the proposed model utilizes raw ET data, it can be easily applied in other eye-tracking scenarios as well and an improved result may be expected. More recent AI algorithms (see, e.g., [36, 37]) could be employed for better future results.

ACKNOWLEDGEMENT

This work was supported by "Pioneer" and "Leading Goose" R&D Program of Zhejiang Province (2023C03002, 2023C03083), the National Key Research and Development Program of China (Grant No. 2022YFC3601002) and the Ningbo Science and Technology Project Grant No. 2023Z179), and the Special Development Fund of Shanghai Zhangjiang Science City. We thank Dr. Julian Evans and Yilun Shen of Zhejiang University for helpful discussions.

REFERENCES

[1] Wei, Q., H. Cao, Y. Shi, X. Xu, and T. Li, "Machine learning based on eye-tracking data to identify Autism Spectrum Disorder: A systematic review and meta-analysis," *Journal of Biomedical Informatics*, Vol. 137, 104254, 2023.

[2] Zeidan, J., E. Fombonne, J. Scorah, A. Ibrahim, M. S. Durkin, S. Saxena, A. Yusuf, A. Shih, and M. Elsabbagh, "Global prevalence of autism: A systematic review update," *Autism Research*, Vol. 15, No. 5, 778–790, 2022.

[3] Jones, W., C. Klaiman, S. Richardson, M. Lambha, M. Reid, T. Hamner, C. Beacham, P. Lewis, J. Paredes, L. Edwards, *et al.*, "Development and replication of objective measurements of social visual engagement to aid in early diagnosis and assessment of autism," *JAMA Network Open*, Vol. 6, No. 9, e2330145, 2023.

[4] Jones, W., C. Klaiman, S. Richardson, C. Aoki, C. Smith, M. Minjarez, R. Bernier, E. Pedapati, S. Bishop, W. Ence, *et al.*, "Eye-tracking-based measurement of social visual engagement compared with expert clinical diagnosis of autism," *JAMA*, Vol. 330, No. 9, 854–865, 2023.

[5] Falck-Ytter, T., S. Bölte, and G. Gredebäck, "Eye tracking in early autism research," *Journal of Neurodevelopmental Disorders*, Vol. 5, No. 1, 28, 2013.

[6] Wan, G., X. Kong, B. Sun, S. Yu, Y. Tu, J. Park, C. Lang, M. Koh, Z. Wei, Z. Feng, Y. Lin, and J. Kong, "Applying eye tracking to identify Autism Spectrum Disorder in children," *Journal of Autism and Developmental Disorders*, Vol. 49, No. 1, 209–215, 2019.

[7] Fang, Y., 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 (ICIP)*, 423–427, Abu Dhabi, United Arab Emirates, 2020.

[8] Fujioka, T., K. Inohara, Y. Okamoto, Y. Masuya, M. Ishitobi, D. N. Saito, M. Jung, S. Arai, Y. Matsumura, T. X. Fujisawa, *et al.*, "Gazefinder as a clinical supplementary tool for discriminating between Autism Spectrum Disorder and typical develop-

ment in male adolescents and adults," *Molecular Autism*, Vol. 7, No. 1, 19, 2016.

[9] Klin, A., 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," *Arch Gen Psychiatry*, Vol. 59, No. 9, 809–816, 2002.

[10] Wang, S., 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, 604–616, 2015.

[11] Oliveira, J. S., 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, 10131, 2021.

[12] Hedger, N. and B. Chakrabarti, "Autistic differences in the temporal dynamics of social attention," *Autism*, Vol. 25, No. 6, 1615–1626, 2021.

[13] Atyabi, A., F. Shic, J. Jiang, C. E. Foster, E. Barney, M. Kim, B. Li, P. Ventola, and C. H. Chen, "Stratification of children with Autism Spectrum Disorder through fusion of temporal information in eye-gaze scan-paths," *ACM Transactions on Knowledge Discovery from Data*, Vol. 17, No. 2, 1–20, 2023.

[14] Gutiérrez, J., Z. Che, G. Zhai, and P. L. Callet, "Saliency4ASD: Challenge, dataset and tools for visual attention modeling for Autism Spectrum Disorder," *Signal Processing: Image Communication*, Vol. 92, 116092, 2021.

[15] Tao, Y. and M.-L. Shyu, "SP-ASDNet: CNN-LSTM based ASD classification model using observer scanpaths," in *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 641–646, Shanghai, China, 2019.

[16] Liaqat, S., C. Wu, P. R. Duggirala, S.-C. S. Cheung, C.-N. Chuah, S. Ozonoff, and G. Young, "Predicting ASD diagnosis in children with synthetic and image-based eye gaze data," *Signal Processing: Image Communication*, Vol. 94, 116198, 2021.

[17] Praveena, K. N. and R. Mahalakshmi, "Classification of Autism Spectrum Disorder and typically developed children for eye gaze image dataset using convolutional neural network," *International Journal of Advanced Computer Science and Applications*, Vol. 13, No. 3, 2022.

[18] Carette, R., M. Elbattah, G. Dequen, J.-L. Guérin, and F. Cilia, "Visualization of eye-tracking patterns in Autism Spectrum Disorder: Method and dataset," in *2018 Thirteenth International Conference on Digital Information Management (ICDIM)*, 248–253, Berlin, Germany, 2018.

[19] Kanhirakadavath, M. R. and M. S. M. Chandran, "Investigation of eye-tracking scan path as a biomarker for autism screening using machine learning algorithms," *Diagnostics*, Vol. 12, No. 2, 518, 2022.

[20] Ahmed, I. A., E. M. Senan, T. H. Rassem, M. A. H. Ali, H. S. A. Shatnawi, S. M. Alwazer, and M. Alshahrani, "Eye tracking-based diagnosis and early detection of Autism Spectrum Disorder using machine learning and deep learning techniques," *Electronics*, Vol. 11, No. 4, 530, 2022.

[21] Elbattah, M., J.-L. Guérin, R. Carette, F. Cilia, and G. Dequen, "Vision-based approach for autism diagnosis using transfer learning and eye-tracking," in *Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies*, Vol. 5, 256–263, Online, 2022.

[22] Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 2017.

[23] Wang, Y.-W., K. J. Dommer, S. J. Webb, and F. Shic, "On the value of data loss: A study of atypical attention in Autism Spectrum Disorder using eye tracking," in *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications*, 1–2, New York, NY, USA, 2023.

[24] Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–9, Boston, MA, USA, June 2015.

[25] Lord, C., S. Risi, L. Lambrecht, E. H. Cook, B. L. Leventhal, P. C. DiLavore, A. Pickles, and M. Rutter, "The autism diagnostic observation schedule-generic: A standard measure of social and communication deficits associated with the spectrum of autism," *Journal of Autism and Developmental Disorders*, Vol. 30, No. 3, 205–223, 2000.

[26] American Psychological Association, *Diagnostic and Statistical Manual of Mental Disorders*, 5th ed., American Psychiatric Publishing, 2013.

[27] Keles, U., D. Kliemann, L. Byrge, H. Saarimäki, L. K. Paul, D. P. Kennedy, and R. Adolphs, "Atypical gaze patterns in autistic adults are heterogeneous across but reliable within individuals," *Molecular Autism*, Vol. 13, No. 1, 39, 2022.

[28] Zhu, H., J. Li, Y. Fan, X. Li, D. Huang, and S. He, "Atypical prefrontal cortical responses to joint/non-joint attention in children with Autism Spectrum Disorder (ASD): A functional near-infrared spectroscopy study," *Biomedical Optics Express*, Vol. 6, No. 3, 690–701, 2015.

[29] Borji, A., "Saliency prediction in the deep learning era: Successes and limitations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 43, No. 2, 679–700, 2021.

[30] Hochreiter, S. and J. Schmidhuber, "Long short-term memory," *Neural Computation*, Vol. 9, No. 8, 1735–1780, 1997.

[31] Huang, X., C. Shen, X. Boix, and Q. Zhao, "SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 262–270, Santiago, Chile, 2015.

[32] Ioffe, S. and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, Vol. 37, 448–456, Lille, France, 2015.

[33] Szegedy, C., V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2818–2826, Las Vegas, NV, USA, 2016.

[34] Dosovitskiy, A., 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 16×16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, Vienna, Austria, 2021.

[35] Dwarampudi, M. and N. V. S. Reddy, "Effects of padding on LSTMs and CNNs," *arXiv:1903.07288*, 2019.

[36] Han, G.-R., A. Goncharov, M. Eryilmaz, S. Ye, B. Palanisamy, R. Ghosh, F. Lisi, E. Rogers, D. Guzman, D. Yigci, *et al.*, "Machine learning in point-of-care testing: Innovations, challenges, and opportunities," *Nature Communications*, Vol. 16, No. 1, 3165, 2025.

[37] Amin, Y., P. Cecere, T. Pomil, and P. P. Pompa, "Smartphone-integrated YOLOv4-CNN approach for rapid and accurate point-of-care colorimetric antioxidant testing in saliva," *Progress In Electromagnetics Research*, Vol. 181, 9–19, 2024.