SpFormer: Spatio-Temporal Modeling for Scanpaths with Transformer

Wenqi Zhong*, Linzhi Yu*, Chen Xia†, Junwei Han, Dingwen Zhang†

School of Automation, Northwestern Polytechnical University, China
wenqizhong@mail.nwpu.edu.cn, 15160557827@mail.nwpu.edu.cn, cxia@nwpu.edu.cn,
junweihan2010@gmail.com, zhangdingwen2006yyy@gmail.com

Abstract

Saccadic scanpath, a data representation of human visual behavior, has received broad interest in multiple domains. Scanpath is a complex eye-tracking data modality that includes the sequences of fixation positions and fixation duration, coupled with image information. However, previous methods usually face the spatial misalignment problem of fixation features and loss of critical temporal data (including temporal correlation and fixation duration). In this study, we propose a Transformer-based scanpath model, SpFormer, to alleviate these problems. First, we propose a fixation-centric paradigm to extract the aligned spatial fixation features and tokenize the scanpaths. Then, according to the visual working memory mechanism, we design a local meta attention to reduce the semantic redundancy of fixations and guide the model to focus on the meta scanpath. Finally, we progressively integrate the duration information and fuse it with the fixation features to solve the problem of ambiguous location with the Transformer block increasing. We conduct extensive experiments on four databases under three tasks. The SpFormer establishes new state-of-the-art results in distinct settings, verifying its flexibility and versatility in practical applications. The code can be obtained from https://github.com/wenqizhong/SpFormer.

Introduction

The human visual system (HVS) plays an essential role in human perception, which receives and processes the majority of information perceived by humans. Human visual behaviors provide valuable insights into the underlying mechanisms of the HVS. A comprehensive understanding of human vision can greatly benefit various downstream tasks, *e.g.*, saliency prediction (Liu et al. 2015; Huang et al. 2015; Wang et al. 2019), salient object detection (Han et al. 2018; Fan et al. 2021), scanpath prediction (Xia et al. 2019), segmentation (Lang et al. 2022), onfocus detection (Zhang et al. 2022), and auxiliary diagnosis (Liu, Li, and Yi 2016; Xia et al. 2022). Two primary data types used to represent human visual behaviors are saliency maps and saccadic scanpaths (abbreviated as scanpaths). Saliency maps usually represent the static spatial probability distribution of attention for a group, while scanpaths typically depict the spatial-temporal attention distribution for an individual. Therefore, scanpaths are well-suited for individual-level analysis and prediction in various domains (Xia et al. 2022; Dalrymple et al. 2019; Mohammadhasani et al. 2020).

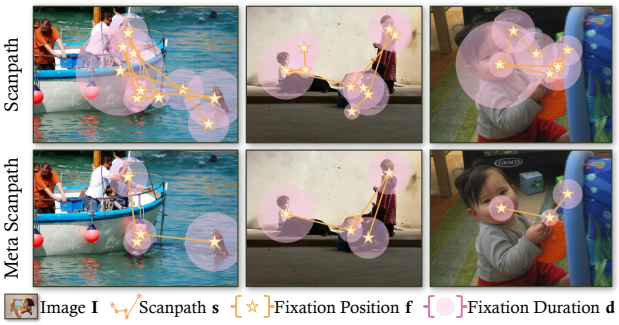

Figure 1: Illustration of the unique and intricate data modality for scanpath and *meta scanpath*. The scanpath contains a sequence of fixation positions and fixation duration, coupled with image information. The meta scanpath stores only a few local fixations (approximately 3-4 visual items).

Scanpath is a unique and intricate data modality, but existing methods for scanpaths often neglect their complexity. The scanpath is a multivariate time series composed of the fixation position and the corresponding duration (see Fig. 1). Moreover, the scanpath is intimately related to the image stimuli that excite the attention behavior. Overall, the intricate properties of the scanpath can be summarized in three key aspects: 1) The scanpath represents a *multivariate time series*. 2) Each time step of the scanpath comprises a pair of *fixation position and fixation duration*. 3) The scanpath exhibits a *coupling* with the corresponding image stimuli.

However, previous scanpath-based models have not comprehensively considered the above-mentioned properties. Generally speaking, existing methods in the medical and psychology fields typically conduct statistic analysis with scanpath representation based on hand-craft features, *e.g.*, the fixation ratio of different regions (Jones and Klin 2013). In the computer community, learning-based models for scanpaths have gradually emerged in recent years (Jiang and Zhao 2017; Dalrymple et al. 2019; Rahman et al. 2021).

*These authors contributed equally.
†corresponding author

However, these methods are inadequate and underpowered to completely model the above-mentioned properties of scanpaths. As a result, these models face the issues such as the spatial misalignment of fixation features and the negligence of temporal correlation and duration of fixations, which cannot provide effective scanpath representation for downstream tasks. For more powerful representation, the model should incorporate specific inductive biases to effectively capture the spatio-temporal structure of scanpaths.

To this end, we propose a novel model, SpFormer, which follows a new regime: extracting spatial fixation features, modeling the temporal correlation of fixations, and integrating fixation duration. Specifically, we first introduce a fixation-centric paradigm that crops the image region around each fixation to tokenize the scanpaths and extract the spatially aligned fixation features without semantic deviation. Then, for fixation-to-fixation temporal correlation modeling, we introduce the global temporal correlation with a temporal mask to reconstruct the causality of fixations and eliminate pseudo correlation. More importantly, we construct a local meta attention to reduce the semantic redundancy of fixations. The generation of scanpaths is controlled by the visual working memory (VWM) mechanism (Epelboim and Suppes 2001), indicating only a few local fixations can be stored at each time, which is what we call a *meta scanpath* (see Fig. 1). The VWM mechanism reduces semantic redundancy within the stored fixations due to limited memory capacity. However, the typical global self-attention may ignore this local characteristic, leading to a slow training process and potential degradation of performance. Therefore, we develop a local meta attention that captures the correlation among stored fixation at each time. Inspired by the VWM mechanism, the local meta attention filters redundant fixations and visual noise, enabling the model to concentrate more effectively on the meta scanpath. We also introduce a consistency loss to ensure that the meta attention of different fixations remains consistent with the meta scanpath.

On the other hand, we integrate the cues of fixation duration into the model for a comprehensive scanpath representation. We observe that the fixation duration is often ignored in previous methods, resulting in incomplete information fusion (Liu, Li, and Yi 2016; Jiang and Zhao 2017; Xia et al. 2022). The fixation duration tends to provide additional cues for the visual allocation of fixations and helps filter the background noise. Based on this observation, we further leverage the fixation duration to adjust the weight of the fixation features. Unfortunately, the temporal location becomes ambiguous with the Transformer block increasing. To address this, we propose a progressive decay mechanism that transitions the weights from distinct to ambiguous, adapting to the progressively ambiguous location.

We conduct comprehensive experiments to evaluate the performance of SpFormer. We also explore the feasibility and generalization of the proposed model on four databases from three tasks, including recognition for autism spectrum disorder (ASD), toddler age prediction, and visual perceptual task prediction with four datasets. Our primary contributions can be summarized as follows:

- We conclude intricate properties of the scanpath modality and propose a novel scanpath-aware Transformer for capturing the spatio-temporal properties of scanpaths.
- We propose a local meta attention to guide the model to focus on local fixation and reduce semantic redundancy according to the VWM mechanism. Moreover, we progressively aggregate the fixation duration into fixation features.
- We design a fixation-centric paradigm to tokenize the scanpaths and address the spatial misalignment problem between fixations and extracted fixation features.
- We present comprehensive experiments across three domains using four datasets. The results highlight that SpFormer achieves new state-of-the-art performance on four real-world scanpath-based tasks.

Related Work

Scanpath-based Application

Scanpath is a type of data representation that offers insights into human visual behavior, which records the eye movements captured by an eye tracker. Scanpaths has wide application across various domains, including healthcare ((Xia et al. 2022; Marsh and Williams 2006; Mohammadhasani et al. 2020)), medical education (Kok and Jarodzka 2017), human-computer interaction (Piumsomboon et al. 2017), education, assisted driving, choice modeling, consumer psychology, and marketing (Klaib et al. 2021).

The application paradigm of scanpaths can be broadly categorized into two aspects. Firstly, many factors like age, gender, neurodevelopment, and visual tasks have been examined to comprehend inter-group variances (Xia et al. 2022; Mastergeorge, Kahathuduwa, and Blume 2021). Therefore, many studies have concentrated on classifying distinct groups, such as individuals with ASD and typical development subjects. Secondly, scanpath analysis has been applied to study the visual behavior of individuals within a group for downstream applications. For instance, scanpaths were utilized to analyze the visual expertise of medical professionals and develop intelligent diagnostic systems in medical imaging (Brunyé et al. 2019).

Transformer

Transformer (Vaswani et al. 2017) was proposed to use self-attention to cover the long-distance dependencies. It has quickly achieved state-of-the-art performance in almost all the natural language processing (NLP) tasks (Devlin et al. 2018; Clark et al. 2020). For example, the Transformer has been successfully employed in the GPT series models (Radford et al. 2018) like Chat-GPT. More recently, Transformer architecture has been further extended in the image and video domain and displayed advanced performance in various tasks, including image recognition (Dosovitskiy et al. 2020), object detection (Carion et al. 2020), semantic segmentation (Strudel et al. 2021; Zheng et al. 2021), video recognition (Bertasius, Wang, and Torresani 2021; Arnab et al. 2021), and super-resolution (Yang et al. 2020). For example, Dosovitskiy *et al.* introduced the vision Transformer (ViT), which adopted a convolution-free architecture to replace the traditional CNN with self-attention mechanisms (Dosovitskiy et al. 2020). Their model can capture global

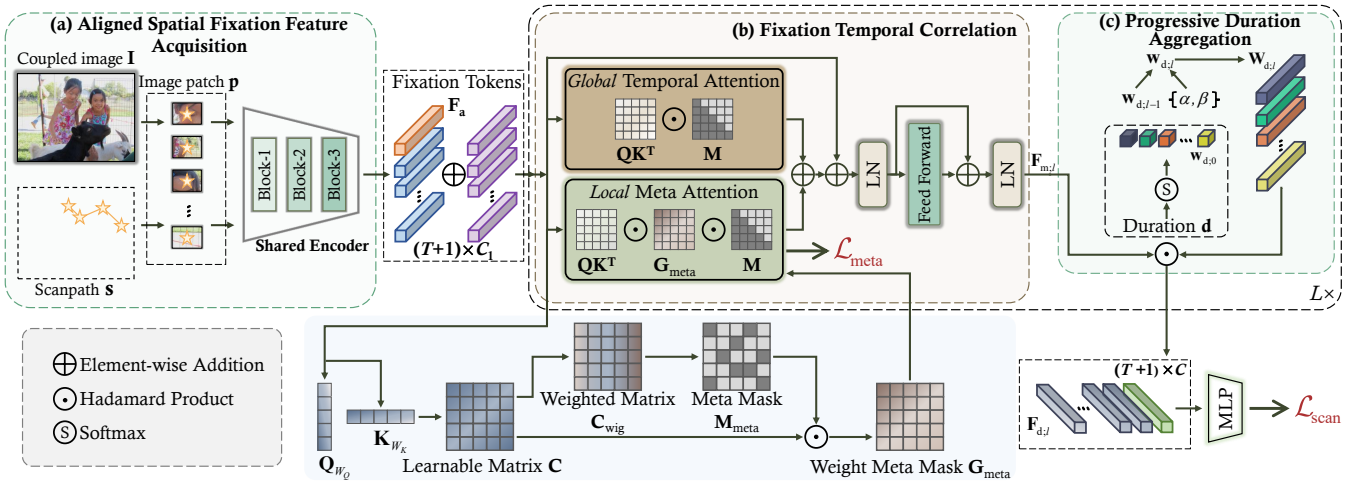

Figure 2: Overall architecture of the proposed SpFormer. First, the SpFormer tokenizes the scanpaths by directly cropping the central patches of fixations to acquire aligned spatial fixation features. Then, we utilize masked global attention and local meta attention to explore fixation temporal correlation. We also introduce a consistency loss to guide the meta of different fixations consistent with the meta scanpath. Finally, we progressively integrate the fixation duration into the SpFormer.

features and relationships for image classification, which achieved strong performance on several benchmark datasets.

The transformer is naturally well-suited for modeling the temporal data (Kim et al. 2022; Wang et al. 2022). In this work, we use the Transformer architecture to capture the spatio-temporal correlation for the intricate data modality.

Methodology

The objective of scanpath representation is to comprehensively capture the inherent characteristics of scanpaths and generate discriminative features for subsequent tasks. In particular, for a given scanpath $\mathbf{s}=\{\mathbf{f},\mathbf{d}\}\in\mathcal{S}$, each of which contains a sequence of fixation positions $\mathbf{f}=\{(x_1,y_1),(x_2,y_2),\cdots,(x_T,y_T)\}\in\mathbb{R}^{2\times T}$ and a sequence of fixation duration $\mathbf{d}=\{d_1,d_2,\cdots,d_T\}\in\mathbb{R}^T$, where (x_t,y_t) and d_t denote the fixation coordinate and fixation duration for the t-th fixation, respectively. In addition, the scanpath also depends on the viewed image $\mathbf{I}\in\mathcal{I}$ with the size $H\times W\times 3$, which influences the intrinsic cues of scanapth. Therefore, the input sample can be represented by a coupled variable $\mathbf{x}=(\mathbf{s},\mathbf{I})\in\mathcal{X}$, where $\mathcal{X}=\mathcal{S}\times\mathcal{I}$ is the Cartesian product of scanpath set $\mathcal{S}$ and the image set $\mathcal{I}$.

To provide complete spatio-temporal modeling for scanpaths, we build a novel model, SpFormer, to cover the special modality. The proposed SpFormer consists of three major components (see Fig. 2), *i.e.*, aligned spatial fixation feature acquisition, fixation temporal correlation modeling, and progressive duration aggregation (PDA). We will elaborate on those modules in this section.

Aligned Spatial Fixation Feature Acquisition

We first tokenize the scanpath using the spatial fixation features. However, previous methods typically extract the entire image features and select the fixation features on the downsampling features (Jiang and Zhao 2017). More specifically, given an image $\mathbf{I}$, utilizing the CNN as the backbone network to obtain the image feature map $\mathbf{F}_\mathrm{I}$ as follows:

$$\mathbf{F_I}=\mathcal{F}_{\text{backbone}}(\mathbf{I})\in\mathbb{R}^{C\times\frac{W}{\psi}\times\frac{H}{\psi}}, \tag{1}$$

where C, H, W, and ψ are the channel, width, height, and downsampling ratio, respectively. The fixation feature sequence $\mathbf{s}_\mathrm{f}$ according to the corresponding fixation position (x_t,y_t) in feature map $\mathbf{F_I}$ is can be derived as:

$$\mathbf{s}_\mathrm{f}^t=\mathbf{F_I}\left[:,\left\lfloor\frac{x_t}{\psi}\right\rfloor,\left\lfloor\frac{y_t}{\psi}\right\rfloor\right]\in\mathbb{R}^C, \tag{2}$$

$$\mathbf{s}_\mathrm{f}=\left\{\mathbf{s}_\mathrm{f}^1,\mathbf{s}_\mathrm{f}^2,\cdots,\mathbf{s}_\mathrm{f}^T\right\}\in\mathbb{R}^{T\times C}, \tag{3}$$

where $t\in\{1,2,\cdots,T\}$ indexes the time step of scanpath. $\lfloor\cdot\rfloor$ denotes rounding toward negative infinity. $\mathbf{s}_\mathrm{f}$ denotes the fixation feature sequence. However, the above diagram unavoidably introduces a spatial misalignment problem between the fixation $\mathbf{f}$ and the extracted scanpath features $\mathbf{s}_\mathrm{f}$. Specifically, each spatial position in feature map $\mathbf{F_I}$ corresponds to an $\psi\times\psi$-dimensional region of the original image $\mathbf{I}$. In other words, any fixation falling within the $\psi\times\psi$-dimensional region will select the same fixation feature that represents the feature of the region center rather than fixation, thereby leading to the issue of spatial misalignment.

How to extract aligned spatial fixation features? A typical way to extract aligned fixation features relies on feature-ware interpolation operation (He et al. 2017). However, the fixation represents the center of the local gaze region, as the fovea is a local region according to the human visual mechanism. Therefore, we propose to clip the original image which does not have many overlapping regions. We clip the fixation-centric region to tokenize the spatial cues of fixation, as shown in Fig. 2 (a). Specifically, we first crop the image $\mathbf{I}$ based on the fixation $\mathbf{f}$ to obtain the fixation-centric image patch sequence $\mathbf{p}=\left\{\mathbf{p}^1,\mathbf{p}^2,\cdots,\mathbf{p}^T\right\}$, which can be formulated as:

$$\mathbf{p}^t\triangleq\mathbf{I}[\mathbf{f}^t+o,:]\in\mathbb{R}^{(2s+1)\times(2s+1)\times 3}, \tag{4}$$

where $o \in [-s, s] \times [-s, s]$ is a 2D *integer offset* with the patch window of size $(2s+1)\times(2s+1)$. $2s+1$ represents the width and height of the image patch $\mathbf{p}^t$. Note that zero padding is applied when the original spatial size $H\times W$ is exceeded. Then, we utilize the encoder network $\mathcal{E}$ and convolutional block to generate the fixation tokens $\mathbf{F}_{\mathrm{a}}$ as follows:

$$\mathbf{F}_{\mathrm{a}} = \mathcal{F}_{\mathrm{conv}}(\mathcal{E}(\mathbf{p})) \in \mathbb{R}^{T\times C_1}, \tag{5}$$

where $\mathcal{F}_{\mathrm{conv}}$ denotes the sequential convolution operations, which capture the aligned features of size C_1. Note that the pixel size of the image patch $\mathbf{p}^t$ can carry the changing size of distance by adjusting s to the predefined size.

Fixation Temporal Correlation

Research has shown the significant role of temporal cues in scanpaths for modeling visual attention and subsequent tasks (Sun, Chen, and Wu 2019). However, previous studies have usually ignored the temporal correlation, concentrating solely on fixation position cues. Therefore, integrating the temporal correlation of scanpaths into the model is an important topic that has not yet been thoroughly discussed.

Global Temporal Attention Traditional Transformers employ a self-attention mechanism, which calculates all paired token correlations to capture the global relationships among the current features. However, the current fixation token is only influenced by the preceding fixation tokens and is not affected by the subsequent fixation tokens. This temporal causality is distinct from typical tasks, such as image classification and detection, where different tokens lack temporal causality and can model correlations between arbitrary tokens. Therefore, we add a simple temporal mask $\mathbf{M}$ to model the temporal causality relationships, which can be formulated as:

$$\mathbf{A}_{\mathrm{global}} = \frac{(\mathbf{Q}_{\mathrm{global}}\mathbf{K}^{\mathrm{T}}_{\mathrm{global}}) \odot \mathbf{M}}{\sqrt{d}}, \tag{6}$$

where $\mathbf{Q}_{\mathrm{global}} \in \mathbb{R}^{T\times \frac{C}{h}}$ and $\mathbf{K}_{\mathrm{global}} \in \mathbb{R}^{T\times \frac{C}{h}}$ denote the global query and global key, respectively. $d = \frac{C}{h}$ is the size of the embedding feature as a scaling factor and h is the number of heads. The temporal mask $\mathbf{M}$ is a lower triangular unit matrix in which the lower triangular elements are 1 with the remaining elements 0.

Local Meta Attention The generation of scanpaths is controlled by the VWM mechanism, responsible for temporarily storing and manipulating visual information in the cognitive system (Ungerleider, Courtney, and Haxby 1998). However, the capacity of VWM is limited, and it can typically store only a small number of local fixations (approximately 3-4 visual items) at each time (Luck and Vogel 1997), which we called a *meta scanpath*. Moreover, scanpaths are often long sequences, which may be influenced by the randomness of visual behavior, leading to the production of noisy fixations. Therefore, we propose to find the meta scanpath that is discriminative fixation and to achieve a compressed representation. Specifically, we first embed the query vector and key vector to obtain the learnable matrix as follows:

$$\mathbf{C} = \mathbf{Q}_{W_Q}\mathbf{K}^{\mathrm{T}}_{W_K} \in \mathbb{R}^{T\times T}, \tag{7}$$

where $\mathbf{Q}_{W_Q}$ and $\mathbf{K}_{W_K}$ denote the two matrices generated with learnable parameters W_Q and W_K. Then, we calculate the local meta attention based on the matrix $\mathbf{C}$ to distill the meta scanpath. For the i-th fixation, we first calculate the index matrix of the max value along the second axis of $\mathbf{C}$ as:

$$\mathbf{D}_{\max} = [\arg\max_j(\mathbf{C}_{ij})] \otimes \mathbb{I}^{\mathrm{T}} \in \mathbb{R}^{T\times T}, \tag{8}$$

where $\mathbf{D}_{\max}$ denotes an index matrix with each element representing the index of the maximum value along the second axis of $\mathbf{C}$. $\mathbb{I}^{\mathrm{T}}$ denotes the vector of ones, and the superscript of $\mathbb{I}^{\mathrm{T}}$ denotes the transpose operation. $\otimes$ denotes the Kronecker product. Then, we calculate the mean value along the second axis of $\mathbf{C}$ as:

$$\mathbf{m} = [\frac{1}{T}\sum_{j=1}^{T}\mathbf{C}_{ij}] \in \mathbb{R}^{T}, \tag{9}$$

where $\mathbf{m}$ is the mean value vector. We argue that the small value for $\mathbf{C}_i$ denotes the corresponding noise, which does not need to be stored. In addition, to reduce the semantic redundancy, we also consider the distance to the maximum value due to the semantic redundancy at close fixation, which can be summarized as:

$$\mathbf{C}_{\mathrm{wig}} = (\mathbf{C} \odot |\mathbf{D}_{\max} - \mathbf{D}|) \in \mathbb{R}^{T\times T}, \tag{10}$$

where $\mathbf{C}_{\mathrm{wig}}$ represents the weighted matrix. $\mathbf{D}$ is the index matrix in which the element of each row is the index for the second axis. Then, we calculate the meta mask $\mathbf{M}_{\mathrm{meta}}$ as:

$$\mathbf{M}_{\mathrm{meta}} = \mathbb{1}(\mathbf{C}_{\mathrm{wig}} > (\mathbf{m}^{\mathrm{T}} \otimes \mathbb{I})) \in \mathbb{R}^{T\times T}, \tag{11}$$

$$\mathbf{G}_{\mathrm{meta}} = \mathbf{C} \odot \mathbf{M}_{\mathrm{meta}} \in \mathbb{R}^{T\times T}, \tag{12}$$

where $\mathbf{G}_{\mathrm{meta}}$ and $\mathbf{M}_{\mathrm{meta}}$ represent the weighted meta mask and meta mask, respectively. $\mathbb{1}$ denotes the indicator function. When the input is true, the output is 1, and when the input is false, the output is 0. Finally, we calculate the local meta attention as:

$$\mathbf{A}_{\mathrm{local}} = \frac{(\mathbf{Q}_{\mathrm{local}}\mathbf{K}^{\mathrm{T}}_{\mathrm{local}}) \odot \mathbf{G}_{\mathrm{meta}} \odot \mathbf{M}}{\sqrt{d}}, \tag{13}$$

where $\mathbf{A}_{\mathrm{local}}$ is the local meta attention. After that, we fuse the global temporal attention and local meta attention as:

$$\mathbf{A}_{\mathrm{f}} = \mathcal{F}_{\mathrm{softmax}}\left(\mathbf{A}_{\mathrm{glb}} + \mathbf{A}_{\mathrm{local}}\right)\mathbf{V}, \tag{14}$$

where $\mathbf{A}_{\mathrm{f}}$ denotes the fused attention. $\mathbf{V}$ denotes the value. $\mathcal{F}_{\mathrm{softmax}}$ denotes the softmax function. Then, we calculate the feature $\mathbf{F}_{\mathrm{m};l}$ through the transformer block $\mathcal{F}_{\mathrm{block}}$ with the fused attention $\mathbf{A}_{\mathrm{f}}$, which can be summarized as:

$$\mathbf{F}_{\mathrm{m};l} = \mathcal{F}_{\mathrm{block}}(\mathbf{F}_{\mathrm{m};l-1}; \mathbf{A}_{\mathrm{f};l}) \in \mathbb{R}^{T\times C}, \tag{15}$$

where $l \in \{1, 2, \ldots, L\}$ is the index of transformer block.

Moreover, inspired by (Lin et al. 2022), we propose a loss to guide the local meta attention $\mathbf{A}_{\mathrm{local}}$ of each fixation consistent with the meta scanpath $\mathbf{s}_{\mathrm{m}}$, which is the mean along the first axis of $\mathbf{M}_{\mathrm{meta}}$. First, we calculate the attention-weighted features $\mathbf{e}^t$ using only the local meta attention, which can be summarized as:

$$\mathbf{e}^t = [\mathbf{F}_{\mathrm{m}}]_{([1],2)} \star [[\mathbf{C} \odot \mathbf{M}_{\mathrm{meta}} \odot \mathbf{M}]_t]_{([1])} \in \mathbb{R}^{T}, \tag{16}$$

where the t of $[\mathbf{C} \odot \mathbf{M}_{\text{meta}} \odot \mathbf{M}]_t$ denotes the row index, $[\cdot]$ at the subscript of $[\mathbf{F}_{\text{m}}]_{([1],2)}$ is the selected dimension of tensor contraction operator $\star$ (Comon 2014). To maintain consistency, we adopt the distance metric to regularize the meta local attention as:

$$\mathcal{L}_{\text{meta}} = \frac{1}{T}\sum_{t=1}^{T} -\mathcal{D}(\mathbf{e}^t, \frac{1}{T}\sum_{t=1}^{T}\mathbf{e}^t), \tag{17}$$

where $\mathcal{D}$ denotes a distance metric. We adopt cosine similarity for the specific implementation.

Progressive Duration Aggregation

Previous scanpath-based methods usually discard the fixation duration and only consider the fixation position (Jiang and Zhao 2017; Arru, Mazumdar, and Battisti 2019; Wu et al. 2019; Tao and Shyu 2019). However, the fixation duration is an important attribute that indicates visual perceived behavior and attention distribution, which is usually used to analyze visual behavior (Jones and Klin 2013). Therefore, we propose to progressively integrate the duration information to the fixation features $\mathbf{F}_{\text{m};l}$. We first leverage the fixation duration $\mathbf{d}$ to obtain the initial weight $\mathbf{w}_{\text{d};0}$ as follows:

$$\omega_j = \frac{e^{d_j}}{\sum_{j=1}^{|\mathcal{I}|} e^{d_j}}, d_j \in \mathbf{d}, \tag{18}$$

$$\mathbf{w}_{\text{d};0} = \{\omega_1, \omega_2, \cdots, \omega_T\} \in \mathbb{R}^T, \tag{19}$$

where $\mathbf{w}_{\text{d};0}$ denotes the initial aggregation weight. However, the temporal location of fixations can become ambiguous as the number of blocks increases, similar to CNNs, where shallow layers emphasize temporal location while high-level features contain more semantic features with ambiguous temporal location. The ambiguous location makes it challenging to distinguish the relationship between fixation duration and fixation. Therefore, we propose a progressive decay mechanism to more reasonably combine the duration into the fixation features, which can be summarized as:

$$\mathbf{w}_{\text{d};l} = \alpha\mathbb{I} + \beta\mathbf{w}_{\text{d};l-1}, \tag{20}$$

where $\alpha = \sigma$ and $\beta = 1 - \sigma$ are progressive decay coefficients. After that, the decay duration weight results are integrated into the extract features $\mathbf{F}_{\text{m};l}$ to generate the fusion features $\mathbf{F}_{\text{d};l}$ of the *l-th* layer:

$$\mathbf{W}_{\text{d};l} = \mathcal{F}_{\text{repeat}}(\mathbf{w}_{\text{d};l}) \in \mathbb{R}^{T\times C}, \tag{21}$$

$$\mathbf{F}_{\text{d};l} = \mathbf{W}_{\text{d};l} \odot \mathbf{F}_{\text{m};l} \in \mathbb{R}^{T\times C}, \tag{22}$$

where $\mathcal{F}_{\text{repeat}}$ represents the repeat operation that repeats the $\mathbf{w}_{\text{d};l}$ at channel dimension to the size of $\mathbb{R}^{T\times C}$.

Training and Inference

Following the ViT (Dosovitskiy et al. 2020), we adopt the extra token to aggregate the information of all fixations and input the MLP to generate prediction $\hat{\mathbf{y}}$. The cross-entropy (CE) loss is leveraged to evaluate the difference between the predicted results $\hat{\mathbf{y}}_{i,j}$ the and ground-truth $\mathbf{y}_i$ as:

$$\mathcal{L}_{\text{scan}} = \frac{1}{|\mathcal{U}_{\text{train}} \times \mathcal{I}|}\sum_{i=1}^{|\mathcal{U}_{\text{train}}|}\sum_{j=1}^{|\mathcal{I}|}\text{CE}(\hat{\mathbf{y}}_{i,j}, \mathbf{y}_i), \tag{23}$$

where $\mathcal{U}_{\text{train}}$ denotes the subject set for training. Finally, we combine the CE loss and the consistency loss with the coefficient λ as:

$$\mathcal{L} = \mathcal{L}_{\text{scan}} + \lambda\mathcal{L}_{\text{meta}}. \tag{24}$$

Experiment

In this section, we conduct the experiment under three different tasks, including ASD recognition, toddler age prediction, and visual perceptual task prediction, to verify the generalization and effectiveness of the SpFormer.

Autism Spectrum Disorder Recognition

Datasets ASD recognition is a critical application in eye-tracking, as it enables early detection in infants and offers an objective and efficient assessment. We apply two datasets, Saliency4ASD (Duan et al. 2019) dataset and our collected dataset, to evaluate the ASD recognition performance of the SpFormer. The Saliency4ASD dataset was collected from 14 children with ASD and 14 typically developing (TD) children. All subjects viewed 300 images, and each image played for 3 seconds. The 300 images were selected from the MIT1003 dataset (Judd et al. 2009). For our dataset, we recruited 58 subjects between 2 and 8 years of age from the hospital to collect eye-tracking data. The participants included 30 children with ASD and 28 TD children.

Baselines For a comprehensive comparison, we adopt the saliency-based models for comparisons. We also follow (Rahman et al. 2021) to report the performance of HoG, Gist, and VGG16, respectively.

Evaluation metrics Following the previous work (Chen and Zhao 2019), we report the scanpath-wise results that evaluate the classification performance based on a single scanpath. We also provide the subject-wise results since the ultimate objective of ASD recognition is to obtain a subject-specific evaluation. Consistent with prior work (Chen and Zhao 2019), we compute the subject-wise probability $\mathbf{p}(c)$ that equally sum the scanpath-wise results across all images as $\mathbf{p}(c) = \frac{1}{|\mathcal{I}|}\sum_{j=1}^{|\mathcal{I}|}\mathbf{p}_j(c)$, where $\mathbf{p}_j(c)$ denotes the subject-wise probability as the class c for subject i.

Main Results Tables 1 and 2 show the experimental results of different approaches under evaluation metrics. Part results follow (Wei et al. 2021) and (Rahman et al. 2021). It can be found that our SpFormer outperforms the advanced models by a considerable margin and sets a new state-of-the-art. With the subject-wise results, the SpFormer receives 100% performance for AUC, sensitivity, specificity, BA, and accuracy at a threshold of 0.5. We achieve 0.0714 AUC and 10.67% accuracy (0.5 thresholds) improvements over the previous best results on the Saliency4ASD. In addition, we observe that the APM and CETS, which capture the temporal cues, have a noticeable performance superiority compared to the model without temporal modeling. Our method has balanced results between sensitivity and specificity compared to other methods. As for our dataset, the SpFormer receives the best performance under most metrics for scanpath-wise results and significantly surpasses

Result	*Method*	AUC ↑	*Sen.* ↑	*Spe.* ↑	BA ↑	Accuracy ↑			
						0.4	0.5	0.6	*avg.*
Scanpath-Wise	DoF (Jiang and Zhao 2017)	0.6070	0.1707	**0.9199**	0.5453	0.5728	0.5492	0.5136	0.5452
	APM (Chen and Zhao 2019)	0.6099	0.5568	0.6113	0.5841	0.5788	0.5849	0.5778	0.5805
	RM3ASD (Arru et al. 2019)	0.5930	0.6843	0.5056	0.5950	-	0.5950	-	-
	SSM (Startsev and Dorr 2019)	0.5984	0.7171	0.4843	0.6007	-	**0.6439**	-	-
	IBM (Wu et al. 2019)	0.5513	0.6350	0.4711	0.5531	-	0.6130	-	-
	SySM* (Wu et al. 2019)	0.5415	0.7407	0.3506	0.5457	-	0.5746	-	-
	SySM†(Wu et al. 2019)	0.5388	0.8071	0.2824	0.5448	-	0.5440	-	-
	SP-ASDNET* (Tao and Shyu 2019)	0.5566	**0.8771**	0.2507	0.5639	-	0.5639	-	-
	SP-ASDNET†(Tao and Shyu 2019)	0.5790	0.5921	0.5664	0.5793	-	0.5793	-	-
	SP-ASDNET‡(Tao and Shyu 2019)	0.5739	0.5936	0.5558	0.5747	-	0.5747	-	-
	CETS* (Wei et al. 2021)	0.6148	0.6857	0.5465	0.6161	-	0.6433	-	-
	CETS†(Wei et al. 2021)	0.6065	0.6964	0.5205	0.6085	-	0.6434	-	-
	SpFormer	**0.6577**	0.6046	0.6424	**0.6235**	**0.6266**	0.6240	**0.6225**	**0.6447**
Subject-Wise	HoG (Dalal and Triggs 2005)	-	-	-	-	-	0.5700	-	-
	Gist (Li and Itti 2009)	-	-	-	-	-	0.6800	-	-
	VGG16 (Simonyan et al. 2014)	-	-	-	-	-	0.6370	-	-
	DoF (Jiang and Zhao 2017)	0.9011	0.0000	**1.0000**	0.5000	**0.8500**	0.4800	0.4800	0.6033
	APM* (Chen and Zhao 2019)	0.9200	0.8600	0.9300	0.8950	-	0.8900	-	-
	APM†(Chen and Zhao 2019)	0.9286	0.8571	0.9231	0.8901	0.6267	0.8933	0.4800	0.6667
	SpFormer	**1.0000**	**1.0000**	**1.0000**	**1.0000**	0.8267	**1.0000**	**0.7433**	**0.8567**

Table 1: Performance comparison on the Saliency4ASD dataset, measured by AUC, sensitivity (*Sen.*, classification threshold is 0.5), specificity (*Sen.*, classification threshold is 0.5), balanced accuracy (BA), and accuracy under three classification thresholds, *i.e.*, 0.4, 0.5, and 0.6, respectively. "*avg.*" denotes the average accuracy performance under three classification thresholds. Results in bold denote the best performance, while the underlined ones indicate the second best. The arrow represents the direction of better performance for the metric. "*", "†", and "‡" denote the different implements, respectively.

Result	*Method*	AUC ↑	*Sen.* ↑	*Spe.* ↑	BA ↑	Accuracy ↑			
						0.4	0.5	0.6	*avg.*
Scanpath-Wise	DoF (Jiang and Zhao 2017)	0.6412	0.5323	0.6568	0.5946	0.5128	0.5984	0.5974	0.5695
	APM (Chen and Zhao 2019)	**0.7086**	0.5868	**0.7257**	0.6563	0.6465	0.6614	0.6634	0.6571
	SpFormer	0.7063	**0.6356**	0.6961	**0.6659**	**0.6571**	**0.6675**	**0.6669**	**0.6638**
Subject-Wise	DoF (Jiang and Zhao 2017)	0.8600	0.7650	0.8140	0.7895	0.5060	0.7150	0.5630	0.5647
	APM (Chen and Zhao 2019)	0.9700	0.7140	**0.9667**	0.8404	0.8110	0.8690	0.6810	0.7870
	SpFormer	**0.9800**	**0.8929**	**0.9667**	**0.9298**	**0.8426**	**0.9314**	**0.7944**	**0.8561**

Table 2: Comparison results on our collected dataset for the ASD recognition task.

the best competitors under all metrics for subject-wise results. It indicates that the SpFormer has more consistent results between images. The SpFormer improves sensitivity from 76.50% to 89.29% and accuracy (0.5 thresholds) from 86.90% to 93.14%, compared to the previous best results.

Ablation Study We carry out a series of ablation studies on the Saliency4ASD and report the subject-wise results.

Effectiveness of ASF. We first consider the aligned fixation feature. We replace the aligned fixation feature with the previous typical misaligned feature (Chen and Zhao 2019). As seen in the results of (a) and (b) in Table 3, our method exhibits a sizeable improvement when using the aligned spatial fixation feature, particularly with a 15.67% enhancement in accuracy at the 0.5 threshold.

Effectiveness of temporal correlation. Then, we replace the proposed temporal modeling with the vanilla Transformer to carry out the ablation study. Comparing (c) and (d) of Table 3, we can conclude that the proposed temporal correlation achieves a significant performance gain on AUC from 0.8901 improve to 0.9835, further demonstrating the significance of the temporal mask and meta scanpath in causal modeling, as well as in reducing the semantic redundancy and visual noise.

Effectiveness of PDA. Furthermore, we plug the PDA into after each Transformer block to progressively fuse the duration and enhance the scanpath features, achieving a 3.333% accuracy improvement over the model without fusing the duration information, which is based on the comparison between (c) and (d) in Table 3.

Effectiveness of consistency loss $\mathcal{L}_{\text{meta}}$. We investigate the effect of the consistency loss $\mathcal{L}_{\text{meta}}$. There is also a performance gain shown in the results of (d) and (e) of Table 3, which indicates that the consistency loss $\mathcal{L}_{\text{meta}}$ plays an essential role in the local meta attention learning.

Effect of Hyper-parameters. We set $\alpha = \sigma = 0.5$, $\beta = 1 - \sigma = 0.5$ in Eq. (20), and $\lambda = 0.1$ in Eq.(24). Finally, we conduct experiments to explore the impact of the hyper-parameters on the results (see Tab. 4). For instance, a high value of σ can lead to a *rapid decay* of duration, which means that duration information cannot be effectively integrated into the model, resulting in performance degradation.

id	ASF	FTC	PDA	$\mathcal{L}_{\text{meta}}$	AUC	Accuracy 0.5
(a)	✗	✗	✗	✗	0.8681	0.7433
(b)	✓	✗	✗	✗	0.8901	0.9000
(c)	✓	✓	✗	✗	0.9835	0.9333
(d)	✓	✓	✓	✗	0.9835	0.9667
(e)	✓	✓	✓	✓	**1.0000**	**1.0000**

Table 3: Ablation studies on main modules of different design choices. "ASF" denotes the aligned spatial fixation feature. "FTC" is the fixation temporal correlation. "PDA" represents the progressive duration aggregation.

σ	0.1	0.3	0.5	0.7	0.9
AUC	0.9615	0.9890	**1.0000**	0.9945	0.9835
Accuracy	0.9333	0.9333	**1.0000**	0.9667	0.9267
λ	0.01	0.05	0.1	0.2	0.5
AUC	0.9945	0.9890	**1.0000**	0.9560	0.9505
Accuracy	0.9333	0.9667	**1.0000**	0.9333	0.8933

Table 4: Influence of hyper-parameters on Saliency4ASD.

Method	AUC	Accuracy 0.5
HoG (2005)	-	0.4700
Gist (2009)	-	0.5700
VGG16 (2014)	-	0.8290
DoF (2017)	0.6692	0.5500
APM (2019)	0.7525	0.7000
CAET (2019)	0.8400	0.8300
CETSMap* (2021)	0.7580	0.7560
CETSMap†(2021)	0.8300	0.8300
SpFormer	**0.8586**	**0.8750**

Table 5: Comparison results on the TAP benchmark.

Toddler Age Prediction

Datasets Identifying different age groups is another application of scanpaths, because eye movement patterns change with age (Munoz et al. 1998; Davidson et al. 2006; Dalrymple et al. 2019). To evaluate the model performance in age prediction, we utilize a toddler age prediction (TAP) dataset[1] obtained from (Dalrymple et al. 2019), which consists of thirty-seven 18-month-old toddlers and thirty-six 30-month-old toddlers. The stimuli are comprised of one hundred images from the object and semantic images eye-tracking (OSIE) database (Xu et al. 2014), which contains 700 image stimuli with abundant attributes.

Experimental Settings We also comply with previous experimental protocols (Rahman et al. 2021). The experimental setting and training details were the same as those in ASD recognition unless specified otherwise.

Main Results Table 5 presents the comparison results on the TAP benchmark. It can be found that our SpFormer receives the best results under all metrics and outperforms the previous methods by a considerable margin. The SpFormer achieves an average accuracy improvement of 11.67%, which achieves a 4.5% accuracy improvement at the threshold of 0.5, and a 1.586% improvement on AUC compared with the previous best results.

[1]https://osf.io/ugvj4

Method	AUC	Accuracy 0.5
HoG (2005)	-	0.7000
Gist (2009)	-	0.8100
VGG16 (2014)	-	0.7490
DoF (2017)	0.6640	0.4639
APM (2019)	0.9894	0.8214
PTEM (2016)	-	0.8438
CETSMap* (2021)	-	0.8635
CETSMap†(2021)	-	0.8420
SpFormer	**0.9974**	**0.9750**

Table 6: Comparison results on the VPT dataset.

Visual Perceptual Task Prediction

Datasets Different visual tasks can elicit varying visual behaviors, even when presented with the same visual scene. Therefore, scanpaths can also be applied to identify the visual tasks of the subjects. Previous methods mainly focus on visual behavior without any specific guidance, known as free-viewing. Koehler et al. (2014) proposed a visual perceptual task (VPT) dataset[2], which contains 800 natural images and four visual tasks: free-viewing, explicit perceptual judgments, saliency search, and cued object search tasks.

Experimental Settings Following the experimental form (Rahman et al. 2021) and (Boisvert and Bruce 2016), we divide the dataset with a series binary classification to classify each two visual tasks. Without loss of generality, we selected the free-viewing and cued object search takes to report results. The experimental setting and training details were the same as those in ASD recognition.

Main Results Table 6 presents the results on the VPT dataset. It can be observed that the SpFormer outperforms other advanced methods. Specifically, our model achieves 0.9974 of AUC, achieving a remarkable improvement of 11.15% in accuracy under the 0.5 threshold. These findings demonstrate the superiority of our proposed SpFormer.

Conclusion

This paper proposes a new model SpFormer to model the spatio-temporal characteristics of scanpath. For the modeling of spatial information, we extract the spatially aligned fixation to represent scanpaths. For temporal cues, we introduce the local meta attention to model the VWM mechanism and progressively aggregate the fixation duration to enhance the fixation feature. Experimental results show that the SpFormer is effective and achieves state-of-the-art performance in multiple scanpath-based tasks.

Acknowledgements

The authors gratefully acknowledge the support of the National Natural Science Foundation of China under Grants

[2]https://data.mendeley.com/datasets/8rj98pp6km/1

62172334, 62027813, 62036005, 62293543, 62202015, and 62322605.

References

Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; and Schmid, C. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6836–6846.

Arru, G.; Mazumdar, P.; and Battisti, F. 2019. Exploiting visual behaviour for autism spectrum disorder identification. In *IEEE International Conference on Multimedia & Expo Workshops*, 637–640. IEEE.

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is space-time attention all you need for video understanding? In *International Conference on Machine Learning*, 4.

Boisvert, J. F.; and Bruce, N. D. 2016. Predicting task from eye movements: On the importance of spatial distribution, dynamics, and image features. *Neurocomputing*, 207: 653–668.

Brunyé, T. T.; Drew, T.; Weaver, D. L.; and Elmore, J. G. 2019. A review of eye tracking for understanding and improving diagnostic interpretation. *Cognitive Research: Principles and Implications*, 4: 1–16.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, 213–229. Springer.

Chen, S.; and Zhao, Q. 2019. Attention-based autism spectrum disorder screening with privileged modality. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1181–1190.

Clark, K.; Luong, M.-T.; Le, Q. V.; and Manning, C. D. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Comon, P. 2014. Tensors: a brief introduction. *IEEE Signal Processing Magazine*, 31(3): 44–53.

Dalal, N.; and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, 886–893. Ieee.

Dalrymple, K. A.; Jiang, M.; Zhao, Q.; and Elison, J. T. 2019. Machine learning accurately classifies age of toddlers based on eye tracking. *Scientific Reports*, 9(1): 1–10.

Davidson, M. C.; Amso, D.; Anderson, L. C.; and Diamond, A. 2006. Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia*, 44(11): 2037–2078.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Duan, H.; Zhai, G.; Min, X.; Che, Z.; Fang, Y.; Yang, X.; Gutiérrez, J.; and Callet, P. L. 2019. A dataset of eye movements for the children with autism spectrum disorder. In *Proceedings of the 10th ACM Multimedia Systems Conference*, 255–260.

Epelboim, J.; and Suppes, P. 2001. A model of eye movements and visual working memory during problem solving in geometry. *Vision Research*, 41(12): 1561–1574.

Fan, D.-P.; Li, T.; Lin, Z.; Ji, G.-P.; Zhang, D.; Cheng, M.-M.; Fu, H.; and Shen, J. 2021. Re-thinking co-salient object detection. *IEEE transactions on pattern analysis and machine intelligence*, 44(8): 4339–4354.

Han, J.; Zhang, D.; Cheng, G.; Liu, N.; and Xu, D. 2018. Advanced deep-learning techniques for salient and category-specific object detection: a survey. *IEEE Signal Processing Magazine*, 35(1): 84–100.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

Huang, X.; Shen, C.; Boix, X.; and Zhao, Q. 2015. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 262–270.

Jiang, M.; and Zhao, Q. 2017. Learning Visual Attention to Identify People With Autism Spectrum Disorder. In *Proc. IEEE Int. Conf. Comput. Vis.*, 3267–3276.

Jones, W.; and Klin, A. 2013. Attention to eyes is present but in decline in 2–6-month-old infants later diagnosed with autism. *Nature*, 504(7480): 427–431.

Judd, T.; Ehinger, K.; Durand, F.; and Torralba, A. 2009. Learning to predict where humans look. In *2009 IEEE 12th International Conference on Computer Vision*, 2106–2113.

Kim, B.; Chang, H. J.; Kim, J.; and Choi, J. Y. 2022. Global-local motion transformer for unsupervised skeleton-based action learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, 209–225. Springer.

Klaib, A. F.; Alsrehin, N. O.; Melhem, W. Y.; Bashtawi, H. O.; and Magableh, A. A. 2021. Eye tracking algorithms, techniques, tools, and applications with an emphasis on machine learning and Internet of Things technologies. *Expert Systems with Applications*, 166: 114037.

Koehler, K.; Guo, F.; Zhang, S.; and Eckstein, M. P. 2014. What do saliency models predict? *Journal of Vision*, 14(3): 14–14.

Kok, E. M.; and Jarodzka, H. 2017. Before your very eyes: The value and limitations of eye tracking in medical education. *Med. education*, 51(1): 114–122.

Lang, C.; Cheng, G.; Tu, B.; and Han, J. 2022. Learning what not to segment: A new perspective on few-shot segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8057–8067.

Li, Z.; and Itti, L. 2009. Gist based top-down templates for gaze prediction. *Journal of Vision*, 9(8): 202–202.

Lin, H.; Ma, Z.; Ji, R.; Wang, Y.; and Hong, X. 2022. Boosting crowd counting via multifaceted attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19628–19637.

Liu, N.; Han, J.; Zhang, D.; Wen, S.; and Liu, T. 2015. Predicting eye fixations using convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 362–370.

Liu, W.; Li, M.; and Yi, L. 2016. Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework. *Autism Research*, 9(8): 888–898.

Luck, S. J.; and Vogel, E. K. 1997. The capacity of visual working memory for features and conjunctions. *Nature*, 390(6657): 279–281.

Marsh, P. J.; and Williams, L. M. 2006. ADHD and schizophrenia phenomenology: visual scanpaths to emotional faces as a potential psychophysiological marker? *Neuroscience & Biobehavioral Reviews*, 30(5): 651–665.

Mastergeorge, A. M.; Kahathuduwa, C.; and Blume, J. 2021. Eye-tracking in infants and young children at risk for autism spectrum disorder: A systematic review of visual stimuli in experimental paradigms. *Journal of Autism and Developmental Disorders*, 51: 2578–2599.

Mohammadhasani, N.; Caprì, T.; Nucita, A.; Iannizzotto, G.; and Fabio, R. A. 2020. Atypical visual scan path affects remembering in ADHD. *Journal of the International Neuropsychological Society*, 26(6): 557–566.

Munoz, D.; Broughton, J.; Goldring, J.; and Armstrong, I. 1998. Age-related performance of human subjects on saccadic eye movement tasks. *Experimental Brain Research*, 121: 391–400.

Piumsomboon, T.; Lee, G.; Lindeman, R. W.; and Billinghurst, M. 2017. Exploring natural eye-gaze-based interaction for immersive virtual reality. In *IEEE Symposium on 3D User Interfaces*, 36–39. IEEE.

Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.; et al. 2018. Improving language understanding by generative pre-training.

Rahman, S.; Rahman, S.; Shahid, O.; Abdullah, M. T.; and Sourov, J. A. 2021. Classifying eye-tracking data using saliency maps. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 9288–9295. IEEE.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Startsev, M.; and Dorr, M. 2019. Classifying autism spectrum disorder based on scanpaths and saliency. In *IEEE International Conference on Multimedia & Expo Workshops*, 633–636. IEEE.

Strudel, R.; Garcia, R.; Laptev, I.; and Schmid, C. 2021. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7262–7272.

Sun, W.; Chen, Z.; and Wu, F. 2019. Visual scanpath prediction using IOR-ROI recurrent mixture density network. *IEEE transactions on pattern analysis and machine intelligence*, 43(6): 2101–2118.

Tao, Y.; and Shyu, M.-L. 2019. SP-ASDNet: CNN-LSTM based ASD classification model using observer scanpaths. In *IEEE International Conference on Multimedia & Expo Workshops*, 641–646. IEEE.

Ungerleider, L. G.; Courtney, S. M.; and Haxby, J. V. 1998. A neural system for human visual working memory. *Proceedings of the National Academy of Sciences*, 95(3): 883–890.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, J.; Bertasius, G.; Tran, D.; and Torresani, L. 2022. Long-short temporal contrastive learning of video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14010–14020.

Wang, W.; Shen, J.; Xie, J.; Cheng, M.-M.; Ling, H.; and Borji, A. 2019. Revisiting video saliency prediction in the deep learning era. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(1): 220–237.

Wei, W.; Liu, Z.; Huang, L.; Wang, Z.; Chen, W.; Zhang, T.; Wang, J.; and Xu, L. 2021. Identify autism spectrum disorder via dynamic filter and deep spatiotemporal feature extraction. *Signal Processing: Image Communication*, 94: 116195.

Wu, C.; Liaqat, S.; Cheung, S.-c.; Chuah, C.-N.; and Ozonoff, S. 2019. Predicting autism diagnosis using image with fixations and synthetic saccade patterns. In *IEEE International Conference on Multimedia & Expo Workshops*, 647–650. IEEE.

Xia, C.; Han, J.; Qi, F.; and Shi, G. 2019. Predicting human saccadic scanpaths based on iterative representation learning. *IEEE Trans. Image Process.*, 28(7): 3502–3515.

Xia, C.; Zhang, D.; Li, K.; Li, H.; Chen, J.; Min, W.; and Han, J. 2022. Dynamic Viewing Pattern Analysis: Towards Large-Scale Screening of Children With ASD in Remote Areas. *IEEE Transactions on Biomedical Engineering*.

Xu, J.; Jiang, M.; Wang, S.; Kankanhalli, M. S.; and Zhao, Q. 2014. Predicting human gaze beyond pixels. *Journal of Vision*, 14(1): 1–20.

Yang, F.; Yang, H.; Fu, J.; Lu, H.; and Guo, B. 2020. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5791–5800.

Zhang, D.; Wang, B.; Wang, G.; Zhang, Q.; Zhang, J.; Han, J.; and You, Z. 2022. Onfocus detection: Identifying individual-camera eye contact from unconstrained images. *Science China Information Sciences*, 65(6): 160101.

Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P. H.; et al. 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6881–6890.