EMS: A Large-Scale Eye Movement Dataset, Benchmark, and New Model for Schizophrenia Recognition

Yingjie Song, Zhi Liu, *Senior Member, IEEE*, Gongyang Li, Jiawei Xie, Qiang Wu, *Senior Member, IEEE*, Dan Zeng, *Senior Member, IEEE*, Lihua Xu, Tianhong Zhang, and Jijun Wang

***Abstract*— Schizophrenia (SZ) is a common and disabling mental illness, and most patients encounter cognitive deficits. The eye-tracking technology has been increasingly used to characterize cognitive deficits for its reasonable time and economic costs. However, there is no large-scale and publicly available eye movement dataset and benchmark for SZ recognition. To address these issues, we release a large-scale Eye Movement dataset for SZ recognition (EMS), which consists of eye movement data from 104 schizophrenics and 104 healthy controls (HCs) based on the free-viewing paradigm with 100 stimuli. We also conduct the first comprehensive benchmark, which has been absent for a long time in this field, to compare the related 13 psychosis recognition methods using six metrics. Besides, we propose a novel mean-shift-based network (MSNet) for eye movement-based SZ recognition, which elaborately combines the mean shift algorithm with convolution to extract the cluster center as the subject feature. In MSNet, first, a stimulus feature branch (SFB) is adopted to enhance each stimulus feature with similar information from all stimulus features, and then, the cluster center branch (CCB) is utilized to generate the cluster center as subject feature and update it by the mean shift vector. The performance of our MSNet is superior to prior contenders, thus, it can act as a powerful baseline to advance subsequent study. To pave the road in this research field, the EMS dataset, the benchmark results, and the code of MSNet are publicly available at https://github.com/YingjieSong1/EMS.**

***Index Terms*— Benchmark, eye movement, large-scale dataset, mean shift, schizophrenia (SZ) recognition.**

I. Introduction

SCHIZOPHRENIA (SZ) is a dominating reason for disability [1] and is featured by deficits in multidomain functioning. The pervasive symptoms of SZ are cognitive deficits, flat affect, and delusions so as to impair the patients' social functions, where cognitive deficits are closer to the core cause [2], [3]. SZ not only torments the patients, but also brings burdens to their families and the society. Early detection and timely intervention can greatly enhance treatment effects and decrease associated costs [4]. The current diagnosis of SZ relies primarily on symptomatology assessments conducted by experienced clinicians [5], [6]. This process is time-consuming and places demands on medical resources, highlighting the need for the development of novel tools to assist in diagnosis.

Recently, there has been an increasing interest in applying eye-tracking technology to computer-aided diagnosis [7], [8], [9]. Abnormalities in eye movement have been identified as an "endophenotype" of SZ [10], which is also described as "a window into the psychotic mind" [11] and is an appropriate tool for evaluating cognition. Compared with healthy control (HC) group, schizophrenics exhibit scanning behavior with a restricted view pattern [12]. Contextual cues are crucial for social cognition. In association with the limited scanning, inefficient integration of contextual information may affect the ability of schizophrenics to assess facial mental states in real life [13]. In contrast to other automatic diagnostic tools, such as magnetic resonance imaging [14], [15], [16], [17], [18], electroencephalogram [19], gene [20], [21], and behavior analysis [22], [23], eye-tracking technology is better, since it takes less time, costs less money, and causes minimal discomfort to patients in clinical applications [8], [10].

Among various paradigms based on eye movement, the free-viewing paradigm is regarded as the most effective one [24], which requires participants to freely view the displayed images (also called stimuli), as shown in Fig. 1. Its simple task rule makes it easy for all age groups to finish.

Manuscript received 1 March 2024; revised 19 June 2024; accepted 7 August 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62171269 and Grant 82171544, in part by the Science and Technology Commission of Shanghai Municipality under Grant 21S31903100, and in part by the Clinical Research Plan of Shanghai Shenkang Hospital Development Center (SHDC) under Grant SHDC2022CRD026. *(Corresponding authors: Zhi Liu; Tianhong Zhang.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Shanghai Mental Health Center.

Yingjie Song, Zhi Liu, Gongyang Li, and Jiawei Xie are with the Key Laboratory of Specialty Fiber Optics and Optical Access Networks, the Joint International Research Laboratory of Specialty Fiber Optics and Advanced Communication, Shanghai Institute for Advanced Communication and Data Science, and the School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China, and also with Wenzhou Institute of Shanghai University, Wenzhou 325000, China (e-mail: songyingjie_shu@163.com; liuzhisjtu@163.com; ligongyang@shu.edu.cn; xietyler@shu.edu.cn).

Qiang Wu is with the Global Big Data Technologies Centre (GBDTC), School of Electrical and Data Engineering, University of Technology Sydney, Ultimo, NSW 2007, Australia (e-mail: Qiang.Wu@uts.edu.au).

Dan Zeng is with the School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China (e-mail: dzeng@shu.edu.cn).

Lihua Xu, Tianhong Zhang, and Jijun Wang are with Shanghai Key Laboratory of Psychotic Disorders, Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine, Shanghai 200030, China (e-mail: dr_xulihua@163.com; zhang_tianhong@126.com; dr_wangjijun@126.com).

Digital Object Identifier 10.1109/TNNLS.2024.3441928

Fig. 1. Fixation density maps of 12 subjects. The first row corresponds to the subjects in SZ group, and the second row corresponds to the subjects in HC group.

The fixation density map in Fig. 1 is computed by convolving the recorded fixations of one subject with a Gaussian kernel when freely viewing a stimulus. Besides, as shown in Fig. 1, it can be observed that the within-group variability of visual attention is significant for both groups (i.e., SZ and HC). Due to the heterogeneity within each group, subject-level diagnosis based on multiple stimuli is more reliable than only one stimulus as input. However, there has been no large-scale and publicly available eye movement dataset for subject-level SZ recognition. The existing methods mainly use their private datasets, as shown in Table I. Despite the effort of Huang et al. [25] to collect some available eye movement data, it is hard to adequately assess the generalizability of methods with such a small number of subjects. Furthermore, there has been no relevant benchmark so far, which hinders the development of subsequent studies.

Viewing multiple stimuli provides more information for accurate recognition, while how to integrate them effectively becomes a new question to be solved. The existing traditional and deep-learning-based methods mainly fused the features or scores of stimuli together through simple average or concatenation operation. There is no method to further explore how to effectively combine multiple stimuli, so as to better represent the corresponding subject. We believe that such a method is potential to push this field forward.

To address the abovementioned issues, we contribute a large-scale and publicly available eye movement dataset for SZ recognition (EMS) and establish a benchmark of the most related works. In terms of the issue of effective integration of multiple stimuli, we propose a novel mean-shift-based method for SZ recognition, namely, MSNet, which focuses on finding the cluster center from the feature subspace of stimuli to represent the corresponding subject. Our key idea is derived from the classical clustering algorithm, i.e., the mean shift algorithm [26], [27]. We aim to find the embedding lying in the densest feature subspace to represent the visual pattern of an individual, which could maximize the discriminative information from stimulus features.

Specifically, our MSNet consists of two branches, one for stimulus features and the other for cluster center. The stimulus feature branch (SFB) is responsible for enhancing the representative capacity of each stimulus feature with the help of the self-attention mechanism [28], while the cluster center branch (CCB) is responsible for generating the cluster center and guiding the cluster center to the denser feature subspace by the mean shift vector. In this way, our MSNet achieves the best performance in comparison with 12 state-of-the-art methods (SOTAs).

Our main contributions are summarized as follows.

TABLE I
FREE-VIEWING PARADIGM-RELATED DATASETS

Dataset	SZ	HC	Images	Public	Benchmark
Benson *et al.* [24]	88	88	56	✗	✗
Morita *et al.* [29]	85	252	56	✗	✗
Kacur *et al.* [30]	22	22	10	✗	✗
Zhang *et al.* [10]	108*	70	35	✗	✗
Huang *et al.* [25]	30	40	100	✓	✗
EMS	104	104	100	✓	✓

"*" means the number 108 exactly refers to clinical high risk (CHR) populations that are further diagnosed into 21 CHR-converter and 87 CHR-nonconverter subgroups.

1) We establish a large-scale and publicly available Eye Movement dataset for schizophrenia recognition (EMS), which consists of the eye movement data from 104 schizophrenics and 104 HCs based on the free-viewing paradigm with 100 stimuli.
2) We conduct the first comprehensive evaluation of the related 13 methods using six metrics in the eye movement-based SZ recognition field, which could act as a benchmark to promote the development of this field.
3) We propose a novel MSNet for eye movement-based SZ recognition to explore effectively integrating multiple stimuli. In our MSNet, we design the SFB to enhance the stimulus feature based on the self-attention mechanism and the CCB to imitate the mean shift algorithm with convolution. Extensive comparison proves that the proposed MSNet achieves superior performance compared with 12 SOTAs.

II. RELATED WORKS

A. Eye Movement Paradigms and Datasets for SZ

As early as over a 100 years ago, Diefendorf and Dodge [31] had observed the abnormal eye movement data of persons with mental disorders. Holzman [32] further confirmed the abnormalities were also present in schizophrenics. Since then, kinds of paradigms were proposed to characterize the eye movement abnormalities of schizophrenics. For example, given the oculomotor deficits of schizophrenics, the smooth pursuit paradigm was designed to capture the eye-tracking dysfunction [29], [33]. Schizophrenics could be disturbed by distractors more easily, resulting in impairment in focusing on a stationary target. As a result, the fixation stability paradigm became a candidate marker for SZ [10], [24]. Besides, many studies found the eye movement data of schizophrenics tended to be a limited scanning pattern, which could be reflected in the free-viewing paradigm [12], [34], [35].

Among the above paradigms, the free-viewing paradigm was regarded as the best one whose effect remained stable regardless of time, sex, medication, or cigarette smoking [24]. Jiang and Zhao [36] supplemented that the free-viewing paradigm was generalizable and easily applicable to most people. Therefore, we focus on the free-viewing paradigm and summarize the related datasets in Table I. For SZ recognition based on the free-viewing paradigm, the clinical eye movement data are scarce. Though these studies [10], [24], [29], [30]

focused on SZ diagnosis, they privatized their datasets, so that subsequent works could not follow them. Huang et al. [25] contributed a publicly available dataset, while the scale of dataset was small, and the benchmark had not been built. Therefore, to better foster the related research, we establish a large-scale and publicly available EMS dataset and conduct a full evaluation/benchmark of eye movement-based SZ recognition on the EMS dataset.

B. Eye Movement-Based Psychosis Recognition

Besides SZ, individuals with autism spectrum disorder (ASD) [37] and depression [38] also show atypical visual attention to various visual stimuli. To review the related works comprehensively, we take ASD and depression recognition into consideration as well. From the perspective of stimulus number, the existing methods can be divided into two categories: stimulus-level methods and subject-level methods.

Duan et al. [39] released the Saliency4ASD dataset in 2019, which was the first open dataset for eye movement-based ASD recognition based on the free-viewing paradigm. It was a pity that the annotation of subject was absent. As a result, only the stimulus-level methods, which rely on the features from a single stimulus, could be carried out. Rooted in this dataset, a set of stimulus-level methods sprang up. For example, Wu et al. [40] designed two branches of ResNet [41]. One was to capture the semantics of the stimulus, and the other was for eye movement features. Tao and Shyu [42] further associated convolutional neural networks with long short-term memory (LSTM) networks to mine the time information of scanpath. Wei et al. [43] proposed the field-of-view maps that effectively extract spatiotemporal features of scanpath. The stimulus-level methods could reflect the difference between patients and HCs, but these methods could not handle the heterogeneity within group relying on only one stimulus as input [44], resulting in limited performance.

The subject-level methods are based on a set of stimuli. Due to the lack of publicly available datasets, the relevant studies are mainly based on their private datasets. These works [10], [24], [25], [45] calculated statistics of eye movement data for each stimulus, which were then concatenated together as subject-level features. Škunda et al. [46] concatenated the fixation density maps of stimuli as input and used a convolutional neural network for classification. Xie et al. [47] came up with a two-stream deep learning network to obtain semantic features and eye movement patterns, and the features of each stimulus were concatenated subsequently. Jiang and Zhao [36] utilized VGGNet [48] to learn the different eye movement patterns between the patient group and HC group based on one stimulus. Then, they selected a subset of discriminative stimuli using the Fisher score and simply averaged the features of stimuli to represent corresponding subjects. Xia et al. [8] computed the score of each stimulus and averaged these scores to get the subject-level scores afterward.

In comparison with the stimulus-level methods, the subject-level methods need to fuse the information from multiple stimuli, which is relatively unexplored. The fusion through average or concatenation assumes that each stimulus contributes equally to psychosis recognition by default, but this is not the case. How to integrate the information of multiple stimuli effectively is a problem to be solved.

TABLE II
EXAMPLE OF THE EMS DATASET, WHICH IS PART OF THE EYE MOVEMENT DATA OF ONE STIMULUS. FIX. IS SHORT FOR FIXATIONS

Index of Fix.	Duration (ms)	X (pixel)	Y (pixel)	Pupil
1	205	518.5	371.3	1177
2	90	275.1	282.2	1262
3	305	600.3	271.1	1265
4	269	635.2	269.2	1326
5	318	597.3	266	1366
6	585	575.4	269.3	1442

C. Clustering for Deep Learning

Clustering algorithms can progressively acquire the representation of a category, i.e., cluster center. The classical clustering algorithms have achieved remarkable success, such as K-means clustering [49], mean shift algorithm [26], [27], and so on. Recently, researchers have tried to combine deep learning with classical clustering algorithms [50], [51], [52]. For example, Zheng et al. [53] adopted locality sensitive hashing to cluster the query features adaptively so as to reduce the computation cost. Inspired by the traditional clustering algorithm [49], Yu et al. [54] regarded the object queries as cluster centers and updated the centers by pooling pixel features. Benefiting from K-means clustering [49], Yu et al. [55] redesigned the cross attention by replacing the softmax operation with the argmax operation to facilitate the performance. Zeng et al. [56] used the K-nearest-neighbor-based density peaks clustering algorithm [57] to get the cluster centers and then assigned the token features to the corresponding cluster centers.

As far as eye movement-based SZ recognition is concerned, we try to represent the viewing pattern of the subjects, that is, find the cluster center of multiple stimulus features. Motivated by the mean shift algorithm [26], [27], we use the mean shift vector to guide our model to find the cluster center lying in the denser feature subspace, which is the core of the CCB.

III. DATASET

To push the progress of eye movement-based SZ recognition research, we collected a large-scale dataset named EMS. Table II provides an example for a quick overview of the EMS dataset. The details of data acquisition are as follows.

A. Subjects

A total of 104 schizophrenics from Shanghai Mental Health Center and 104 HCs from Shanghai University were recruited. The schizophrenics all met the diagnostic criteria based on the structured clinical interview for diagnostic and statistical manual of mental disorders (DSM-IV-TR) [5]. The subjects had normal vision or corrected vision, and the right hand was the dominant hand. All subjects were willing to take

TABLE III

DEMOGRAPHIC CHARACTERISTICS OF SZ AND HC GROUP. FOR AGE, EDUCATION, AND DURATION OF ILLNESS, THE REPORTED VALUES ARE MEAN (STANDARD DEVIATION). "−" REPRESENTS NO RELEVANT DATA

	SZ ($n = 104$)	HC ($n = 104$)	p value
Age (year)	50.30 (14.97)	50.32 (12.12)	0.992
Gender (male/female)	80/24	78/26	0.105
Education (year)	11.50 (2.82)	10.71 (3.76)	0.089
Duration of illness (year)	5.26 (6.90)	−	−

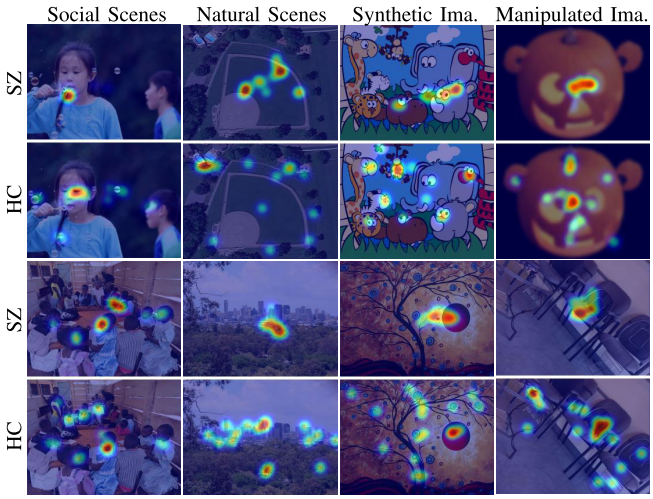
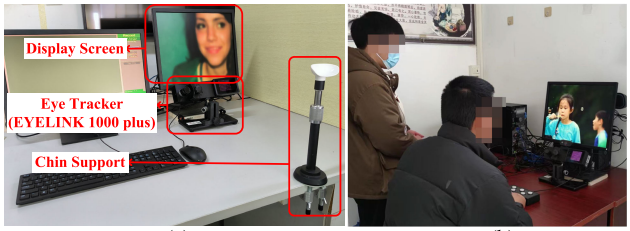


Fig. 2. Examples of stimuli. Ima. is short for images. Each row is overlaid by the fixation density maps (visualized by heatmaps) of one subject in SZ group or HC group.

part in this study and signed the informed consent. The details of exclusion criteria were the same as those in [25]. We employed independent t-tests for continuous variables and chi-square tests for categorical variables to measure the between-group differences of demographic characteristics. As shown in Table III, the two groups were statistically matched on age, gender, and educational background. The average duration of illness is 5.26 years. In experiments, we retained the eye movement data of 48 subjects as the test set, and the remaining 160 subjects were used for fourfold cross validation, where the proportion of SZ group was all 50%.

B. Stimuli

Each subject was instructed to finish the free-viewing task. The 100 images were presented as stimuli in a random order with each one showing for 5 s. Given the atypical visual attention of schizophrenics [12], [58], we selected four categories of stimuli, i.e., social scenes, natural scenes, synthetic images, and manipulated images. The stimuli were mainly selected from the MIT dataset [59] and the OSIE dataset [60]. We also supplemented some images from the Internet. The examples of four categories are shown in Fig. 2. For the category of social scenes, as shown in the first column of Fig. 2, schizophrenics are usually unable to scan rapidly and uptake of social contextual information, especially in case of multiple faces. The poor integration of contextual cues in social scenes shows significant deficits of schizophrenics in informing social cognition [13], [61]. Besides, we purposefully select stimuli with objects distributed far from the center or with complex content in the category of natural scenes and synthetic images to characterize the restricted visual pattern of schizophrenics. The restricted visual pattern impedes the processing of environmental information and further impairs the cognitive function [12]. We make the stimuli intricate by blurring, adding noise, splicing, or rotating in the category of manipulated images, so as to better characterize cognitive impairment.



(a) (b)

Fig. 3. (a) Devices and (b) scene for data acquisition.

C. Experimental Procedure

As shown in Fig. 3, we chose the EYELINK 1000 plus desktop eye tracker to collect eye movement data in a therapy room. The sampling rate was set to 1000 Hz. The eye tracker was coupled to a 19-in display screen with a resolution of 1024×768. Subjects were positioned at a distance of 0.6 m away from the screen, and the subject's head was fixed with the special chin support. We utilized the nine-point calibration manner and recorded data of the eye with smaller calibration errors. This study was approved by the ethics committee of Shanghai Mental Health Center.

IV. PROPOSED METHOD

A. Network Overview

As shown in Fig. 4, we extract the feature map of each stimulus by the saliency prediction model, i.e., RINet [62], which is detailed in Section IV-D. For the feature map of each stimulus, the features lying in the fixation positions are concatenated as initial stimulus features $\boldsymbol{f}^i \in \mathbb{R}^{1\times N\times C}$ ($i \in \{1, 2, \ldots, 100\}$), and then, all these initial stimulus features are concatenated as input $\mathbf{F} \in \mathbb{R}^{100\times N\times C}$, where 100 is the number of stimuli, and N[1] and C[2] denote the number of fixation and the dimension of channel, respectively. The proposed MSNet consists of two branches. The initial stimulus features $\mathbf{F}$ are fed into the SFB first. This branch contains a fixation embedding layer and four stacked stages to enhance the representative ability of each stimulus feature. After the fixation embedding layer and each stage in the SFB, the stimulus features are passed to the CCB. This branch consists of cluster center generator (CCG)

[1]Since the average of fixations is 14 in the EMS dataset, the number of fixation N is set to 14. When there are fewer than 14 fixations, we append null to the end.

[2]The channel dimension C is 1056. We elaborate on the process of extracting the feature map of each stimulus in Section IV-D.

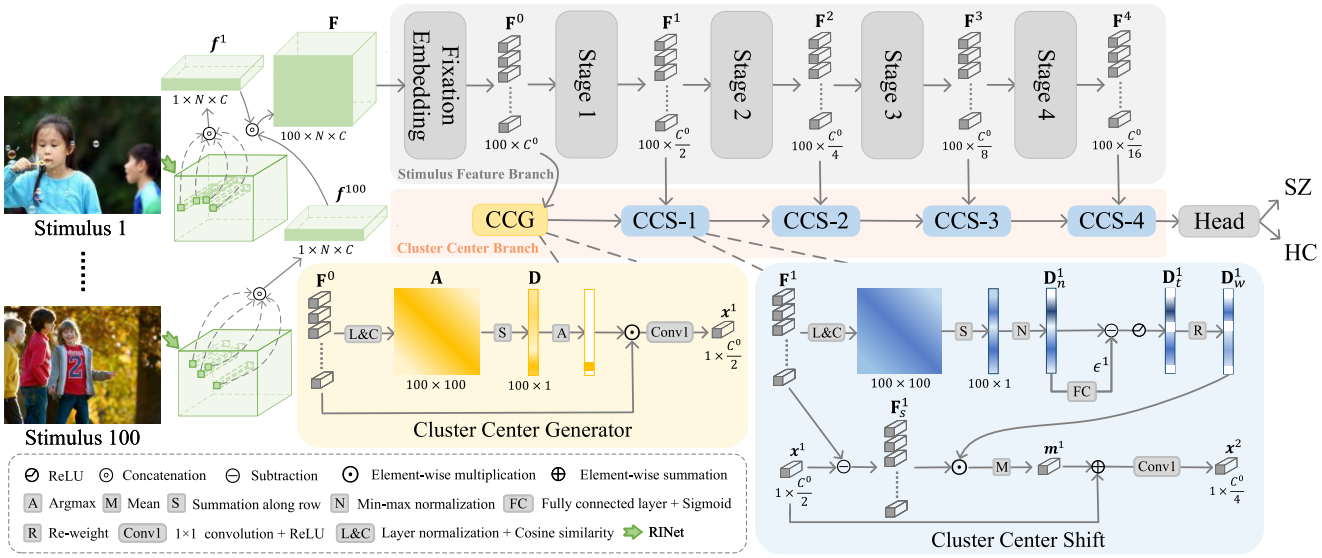


Fig. 4. Overview of the proposed MSNet. Given a set of stimuli, the features at fixation positions are concatenated as the initial stimulus features, where the fixation number N is simplified to 4 for understanding. These features are enhanced by the SFB in a multistage manner. The other branch, namely, the CCB, is designed to generate the cluster center with the CCG module and move the cluster center to the denser feature subspace with the CCS module. Finally, the updated cluster center is fed into the classification head.

module and cluster center shift (CCS) module to generate and move the cluster center. The two modules are detailed in Sections IV-C1 and IV-C2, respectively. Finally, the updated cluster center represents the visual pattern of the corresponding subject, and the classification head is exploited to distinguish whether the subject belongs to HC or SZ group based on the final cluster center. The classification head contains a layer normalization [63], a fully connected layer, and a sigmoid activation function.

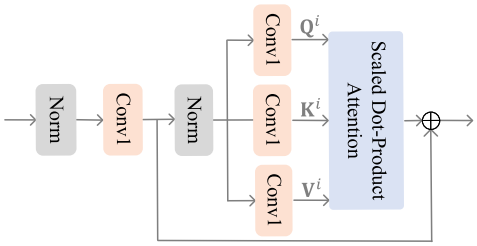


Fig. 5. Structure of each stage in SFB.

B. Stimulus Feature Branch

Because of the inherent inductive biases, convolution architecture is unable to capture long-range contextual information, while the self-attention mechanism [28] has shown great potential in acquiring long-range information [64]. Here, we design the SFB based on the self-attention mechanism. As for the initial stimulus features $\mathbf{F}$, they are only built upon each individual stimulus. There is a lack of long-range information integration among stimuli. We aim to enhance each stimulus with similar information from all stimuli based on the self-attention mechanism. Specifically, the SFB is comprised of a fixation embedding layer and four stacked stages.

First, a fixation embedding layer is adopted to reduce the channel dimension and reshape the feature $\mathbf{F}$ to $\mathbf{F}^0 \in \mathbb{R}^{100\times C^0}$, where C^0 is the dimension of channel. The process in fixation embedding layer can be described as follows:

$$\mathbf{F}^0 = \mathrm{rs}(\mathrm{conv}(\mathbf{F}; \mathbf{W}_{\mathrm{fe}})) \tag{1}$$

where $\mathrm{rs}(\cdot)$ denotes the reshape operation and $\mathrm{conv}(*; \mathbf{W}_{\mathrm{fe}})$ represents the 1×1 convolutional layer with parameters $\mathbf{W}_{\mathrm{fe}}$. The following four stages, i.e., Stage-i ($i \in \{1, 2, 3, 4\}$), are based on the same structure shown in Fig. 5. We first employ layer normalization and a convolutional layer to reduce the channel dimension by half, which makes the stimulus features compressive. After the subsequent layer normalization, the stimulus features are mapped to the queries $\mathbf{Q}^i$, keys $\mathbf{K}^i$, and values $\mathbf{V}^i$ by one convolutional layer, where i corresponds to Stage-i. Following the original Transformer [65], we compute the scaled dot-product attention $\mathrm{Attention}(\cdot)$ as follows:

$$\mathrm{Attention}\left(\mathbf{Q}^i, \mathbf{K}^i, \mathbf{V}^i\right) = \mathrm{Softmax}\left(\frac{\mathbf{Q}^i \otimes \left(\mathbf{K}^i\right)^{\mathrm{T}}}{\sqrt{d^i}}\right)\mathbf{V}^i \tag{2}$$

where $\mathrm{Softmax}(\cdot)$ and $\otimes$ denote the softmax function and matrix multiplication, respectively, and d^i is the dimension of $\mathbf{K}^i$. In terms of a certain stimulus feature, the dot-product function of its query with all keys measures the similar information between it and all stimulus features. Then, according to the dot-product function, a weighted sum of the values is computed as output, where this stimulus feature can be enhanced by aggregating similar features to its output. Besides, we also apply a residual connection after the scaled dot-product attention. The outputs of Stage-i are denoted as $\mathbf{F}^i$.

C. Cluster Center Branch

After the SFB, we focus on finding the cluster center among these stimuli features to represent the visual pattern of the corresponding subject. The classical algorithm, i.e., mean shift algorithm [26], [27], has done a good job in locating the

cluster center. As a result, we imitate the mean shift algorithm with convolution to design the CCB. The original mean shift algorithm consists of the following two steps for each sample.

1) *Step 1:* Selecting a sample x as the initial cluster center randomly.
2) *Step 2:* Computing the mean shift vector $m(x)$ as Eq. (3), and shifting the cluster center to $x+m(x)$. Then, iterating this operation until the cluster center reaches a peak of sample density

$$m(x) = \frac{\sum_{x_i \in S} w(x_i)(x_i - x)}{\sum_{x_i \in S} w(x_i)} \tag{3}$$

where S denotes the set of selected samples and $w(\cdot)$ is a weight function.

Corresponding to these two steps, we design the following two modules in the CCB, namely, the CCG module and the CCS module, to generate and move the cluster center, respectively.

1) CCG Module: The CCG module is responsible for choosing an appropriate stimulus feature as the initial cluster center. Following the mean shift algorithm, the cluster centers are most likely to lie in the densest place. We redesign the affinity matrix to represent the density of features. Specifically, as shown in Fig. 4, the input $\mathbf{F}^0$ contains the features of 100 stimuli. In other words, each row of $\mathbf{F}^0$ represents the feature of one stimulus. We apply the layer normalization [63] to $\mathbf{F}^0$ and denote its i-th row and j-th row as $\boldsymbol{r}_i$ and $\boldsymbol{r}_j$, respectively. Then, the cosine similarity between every two features is calculated to obtain the affinity matrix $\mathbf{A} \in \mathbb{R}^{100\times 100}$ as follows:

$$\mathbf{A}(i, j) = \frac{\boldsymbol{r}_i \cdot \boldsymbol{r}_j}{\|\boldsymbol{r}_i\| \|\boldsymbol{r}_j\|} \quad (i, j \in \{1, 2, \ldots, 100\}) \tag{4}$$

where $\cdot$ is dot product and $\|\cdot\|$ means L2 norm. The numerator of Eq. (4) computes the dot product of features from every two stimuli, and the denominator of Eq. (4) computes the product of the L2 norms of features from every two stimuli. Based on Eq. (4), we observe that the affinity matrix $\mathbf{A}$ belongs to $[-1, 1]^{100\times 100}$. The higher value stands for the more similarity between features. Then, the affinity matrix $\mathbf{A}$ is summed along row to get the density vector $\mathbf{D} \in \mathbb{R}^{100\times 1}$. Each row of $\mathbf{A}$ indicates the similarity between a certain stimulus feature and all features, so that we regard the sum along row as the feature density. We utilize the argmax operation and elementwise multiplication to pick out the stimulus feature with the highest density as the initial cluster center. Moreover, a convolutional layer is adopted to align the dimension with the next CCS module, and the final feature is denoted as $\boldsymbol{x}^1 \in \mathbb{R}^{1\times(C^0/2)}$.

2) CCS Module: The intuition of mean shift is gradient ascent, and the original mean shift algorithm achieves this target using the mean shift vector $m(x)$. When computing $m(x)$, the contribution of each sample is different. It is more reasonable to treat samples unequally using a weight function, i.e., $w(\cdot)$. Here, we come up with the CCS module to compute $m(x)$ with a density-based weight function.

We design CCS-i ($i \in \{1, 2, 3, 4\}$) to match the corresponding four stages in SFB. Here, we take the first CCS module, i.e., CCS-1, as an example for clarification (shown in the bottom-right part of Fig. 4). Due to the prior knowledge, that is, these stimulus features all belong to one subject, we take all of them into the selected set. We get the subtraction between $\mathbf{F}^1$ and $\boldsymbol{x}^1$ to obtain the vector $\mathbf{F}^1_s \in \mathbb{R}^{100\times(C^0/2)}$ pointing from the initial cluster center to the enhanced stimulus features, which is the same as the original mean shift algorithm. As for the weight function, we calculate the density vector as mentioned in Section IV-C1 and then normalize it as $\mathbf{D}^1_n \in \mathbb{R}^{100\times 1}$ by using the min–max normalization. An adaptive threshold ϵ^1 is learned from $\mathbf{D}^1_n$ through a fully connected layer. Afterward, the portion of $\mathbf{D}^1_n$ that is smaller than the threshold ϵ^1 is set to zero using the subtraction operation and ReLU, resulting in $\mathbf{D}^1_t \in \mathbb{R}^{100\times 1}$. In the end, $\mathbf{D}^1_t$ is reweighted to sum to one so as to get the weight function $\mathbf{D}^1_w$. Similar to (3), we multiply $\mathbf{F}^1_s$ by $\mathbf{D}^1_w$ and average the result as the mean shift vector $\boldsymbol{m}^1 \in \mathbb{R}^{1\times(C^0/2)}$ as follows:

$$\boldsymbol{m}^1 = \text{mean}\left(\mathbf{D}^1_w \odot \left(\mathbf{F}^1 \ominus \boldsymbol{x}^1\right)\right) \tag{5}$$

where mean$(\cdot)$ denotes computing the mean of rows, and $\odot$ and $\ominus$ represent elementwise multiplication and subtraction, respectively. The mean shift vector $\boldsymbol{m}^1$ is utilized to shift the initial cluster center $\boldsymbol{x}^1$ by the elementwise summation operation. Subsequently, a convolutional layer is adopted to align the dimension of the shifted cluster center with the next CCS module, resulting in $\boldsymbol{x}^2 \in \mathbb{R}^{1\times(C^0/4)}$ as follows:

$$\boldsymbol{x}^2 = \text{conv}\left(\boldsymbol{x}^1 \oplus \boldsymbol{m}^1; \mathbf{W}^1_{\text{ccs}}\right) \tag{6}$$

where $\oplus$ is elementwise summation and $\mathbf{W}^1_{\text{ccs}}$ is the parameters of convolutional layer. Notably, there is no convolutional layer in CCS-4, because CCS-4 is the last module.

D. Implementation Details

The proposed MSNet was implemented by PyTorch [68] with an NVIDIA TITAN Xp GPU. The loss function adopted in the training procedure was binary cross entropy. We followed DeiT [69] and applied AdamW [70] with a momentum of 0.9 and a weight decay of 5×10^{-2} to optimize the model. The batch size was set to 8. The initial learning rate of training was set to 5×10^{-4} and decreased according to a cosine schedule [71]. To facilitate converging in the training procedure, we applied the warming-up stage for the adaptive thresholds ϵ^i ($i \in \{1, 2, 3, 4\}$) for 50 iterations. The channel dimension C^0 was 896. Our MSNet was trained for 50 epochs for each cross validation. The schizophrenics were labeled with 1, and the HC group was labeled with 0 as ground truth. We utilized the maximum between-class variance method to select a threshold for each validation set, which was the same as [8]. The subject with score above the threshold was given a positive SZ diagnosis.

As for extracting the feature map from 100 stimuli, we chose the saliency prediction model, i.e., RINet [62], and fine-tuned it on our EMS dataset for 20 epochs. The initial learning rate of fine-tuning was 1×10^{-5}, which was divided by 10 after 10 epochs. Following [36], the difference of fixation (DoF) maps were chosen as ground truth rather than the fixation density map. We aimed to distinguish two clinical populations based on what they focused on. The DoF maps depicted the subtle differences between the fixation density maps of two

TABLE IV

QUANTITATIVE PERFORMANCE COMPARISON ON THE EMS DATASET. DL-BASED IS SHORT FOR DEEP-LEARNING-BASED METHODS. ↑ REPRESENTS LARGER VALUE IS BETTER. THE BEST TWO RESULTS ARE MARKED IN RED AND BLUE, RESPECTIVELY

	Methods	Validation Set						Test Set					
		Acc.	Sen.	Spe.	AUC	Pre.	F1-score	Acc.	Sen.	Spe.	AUC	Pre.	F1-score
Traditional	EDB_SVM [45]	0.7313	0.7156	0.7396	0.8086	0.7240	0.7195	0.6875	0.6667	0.7083	0.7813	0.6957	0.6809
	EDB_QDA [45]	0.7063	0.6026	0.7993	0.7205	0.7403	0.6638	**0.7500**	0.6667	**0.8333**	0.8073	**0.8000**	0.7273
	EDB_BYS [45]	0.7000	0.6065	0.7785	0.7987	0.7061	0.6510	0.7292	0.5833	**0.8750**	0.7413	**0.8235**	0.6829
	ESR_SVM [25]	0.7938	**0.7946**	0.7847	0.8498	0.7859	0.7889	**0.7500**	0.7083	0.7917	0.7760	0.7727	**0.7391**
	ESR_RF [25]	0.7625	0.7540	0.7639	0.8521	0.7549	0.7538	0.7292	0.7083	0.7500	0.7760	0.7391	0.7234
DL-based	IAS [43]	0.7563	0.7634	0.7500	0.8180	0.7770	0.7572	**0.7500**	0.6667	**0.8333**	**0.8229**	**0.8000**	0.7273
	LVA [36]	**0.8188**	0.7766	**0.8585**	**0.8987**	**0.8552**	**0.8096**	0.7083	0.6667	0.7500	0.8038	0.7273	0.6957
	DDB_DoF [66]	0.7250	0.7872	0.6727	0.7645	0.7254	0.7446	0.6667	**0.8333**	0.5000	0.7483	0.6250	0.7143
	DDB [66]	0.7313	0.7866	0.6873	0.7772	0.7394	0.7512	0.7083	0.6667	0.7500	0.7882	0.7273	0.6957
	DVP [8]	0.7875	0.7548	0.8056	0.8516	0.8206	0.7769	**0.7500**	0.7083	0.7917	0.8194	0.7727	**0.7391**
	GPI_LSTM [67]	0.7125	0.6881	0.7424	0.7953	0.7291	0.6947	0.6875	**0.7917**	0.5833	0.7726	0.6552	0.7170
	GPI_GRU [67]	0.7063	0.7045	0.7188	0.7798	0.7219	0.6983	**0.7500**	0.7083	0.7917	0.8125	0.7727	**0.7391**
	MSNet	**0.8313**	**0.8051**	**0.8708**	**0.8972**	**0.8575**	**0.8244**	**0.8125**	**0.8333**	0.7917	**0.8854**	**0.8000**	**0.8163**

groups, so that the DoF maps were more suitable for eye movement-based SZ recognition. We employed the feature before the last fusion module of RINet as the feature map of each stimulus. As a result, the channel dimension C was 1056.

V. BENCHMARK

We conduct the first comprehensive benchmark where the related 13 methods are taken into comparison, and six metrics are adopted to measure them. The experimental results are analyzed, and the potential research topics are discussed in this section.

A. Evaluation Metrics

To assess the performance of the proposed MSNet and other SOTAs comprehensively, we adopt six widely used evaluation metrics as follows.

1) Accuracy (Acc.) $=$ (TP $+$ TN)/(TP $+$ TN $+$ FP $+$ FN).
2) Sensitivity (Sen.) $=$ TP/(TP $+$ FN).
3) Specificity (Spe.) $=$ TN/(TN $+$ FP).
4) Precision (Pre.) $=$ TP/(TP $+$ FP).
5) F1-score $=$ (2 $\times$ Pre. $\times$ Sen.)/(Sen. $+$ Pre.).
6) AUC is the area under the curve of receiver operating characteristic (ROC) analysis, which varies the threshold for classification and plots the true positive rate versus false positive rate.

TP denotes true positive, FP represents false positive, TN indicates true negative, and FN stands for false negative. For all metrics, larger values are regarded to be better. For robust assessment, the average metrics of fourfold cross validation are regarded as the performance of validation, and the model with the highest AUC will be tested on the test set.

B. Comparison Methods

For baseline experiments, we compare 13 psychosis recognition methods, including our MSNet and 12 SOTAs, on the EMS dataset to conduct a comprehensive evaluation. These methods are all subject-level methods or adjusted to subject-level methods. As shown in Table IV, the compared methods include five traditional methods, which are based on the statistics of eye movement data, and eight deep-learning-based methods, which extract features from the eye movement data by deep-learning-based models. Apart from MSNet, the other methods are elaborated as follows.

EDB_SVM, **EDB_QDA**, and **EDB_BYS** are the traditional recognition methods using support vector machine (SVM), quadratic discriminant analysis (QDA), and Bayesian (BYS) algorithm as classifier in [45], respectively. Though there are three paradigms utilized to test in [45], for a fair comparison, we only choose the free-viewing paradigm and select the corresponding five variables for classification.

ESR_SVM and **ESR_RF** represent the traditional recognition method using SVM and random forest (RF) algorithm as classifier in [25], respectively. Because the complete feature set proposed in [25] is proved to be the most effective in identifying schizophrenics, we deploy the two methods with the complete feature set proposed in [25].

IAS [43] is a stimulus-level method originally. We compute the scores of 100 stimuli for each subject and average the scores to represent the corresponding subject. In terms of the threshold, we apply the same way as MSNet and [8], that is, exploit the maximum between-class variance method to select the threshold for each validation set.

LVA denotes the convolution-based recognition method proposed by [36]. The number of stimuli in the EMS dataset is 100, which is equal to the number after the stimuli selection in [36], so we deploy this method with all 100 stimuli.

DDB_DoF stands for the method extracting the stimulus features based on the DoF prediction task in [66].

DDB is the recognition method based on both the DoF prediction task and semantic segmentation task in [66].

DVP presents the subject-level method proposed in [8]. Xia et al. [8] utilized an LSTM network to encode the scanpath of one stimulus and then classified based on the average score of all stimuli.

GPI_LSTM and **GPI_GRU** denote the recognition methods using LSTM network and gated recurrent unit (GRU) network, respectively, to learn the mapping between inputs and the labels in [67]. Because the combination of position

and duration is proven to be a better choice as input, we feed position and duration into LSTM network and GRU network.

With regard to the name of the above methods, for simplicity, we adopt the initial letters of the first three words in the publication title to stand for the corresponding publication.

C. Performance Comparison and Analysis

Table IV reports the quantitative performance comparison on the EMS dataset.

Among the traditional methods, EDB_QDA achieves the overall best performance on the test set. Compared with EDB_SVM and EDB_BYS, it can be observed that the choice of classifier can affect classification performance. As for the handcrafted features, compared with [45], Huang et al. [25] designed a set of additional features. According to the comparison between ESR_SVM and EDB_SVM, these additional features can effectively improve performance on the validation and test sets.

In terms of the deep-learning-based methods, the performance of MSNet is better than other methods on both validation and test sets, showing effective integration of multiple stimuli can indeed improve the performance significantly. Among the deep-learning-based methods, IAS gets the second-best performance on the test set overall. This is because the restricted visual pattern is the main characteristic of schizophrenics; that is, besides semantic information, temporal and spatial information of scanpath also plays an important role. IAS combines the statistics of temporal and spatial information of scanpath with the semantic information lying in the fixations, whose competitive performance proves adding the spatial and temporal information is a potential direction for performance improvement.

D. Potential Research Topics

In this section, we analyze some challenges in the EMS dataset, including performance improvement, accurate recognition with fewer stimuli, and clinical interpretability. Furthermore, we discuss the possible solutions to these challenges. The EMS dataset can also be used in general algorithm research, such as data augmentation, transfer learning, contrastive learning, explainable learning, and so on. The potential research topics are briefly summarized as follows.

1) Performance Improvement: The EMS dataset is designed for the clinical applications of eye movement-based SZ recognition to assist clinicians in rapid and accurate diagnosis. As shown in Table IV, we can observe that the highest accuracy on the test set is 0.8125. The further improvement is essential for clinical applications. According to the analysis in Section V-C, how to effectively utilize the spatial and temporal information of scanpath is a possible direction for improvement. The alterations in pupil size can reflect the subjects' concentration and processing load [72], which is a potential biomarker to differentiate schizophrenics from HC group [25]. Integrating the information of pupil size into deep-learning-based method may further improve performance. In the field of computer vision, data augmentation schemes, such as flipping, cropping, and so on, are common and effective ways to fully leverage datasets. Whether these common operations can be directly applied to eye movement data remains to be verified. The specific data augmentation methods for eye movement data need to be explored. In this work, we attempt to transfer the knowledge of HC group's visual attention in the SALICON dataset [73] and the MIT1003 dataset [59] to the eye movement-based SZ recognition task. How to transfer knowledge from other related datasets to the eye movement-based SZ recognition task is one possible direction to realize the next improvement.

2) Accurate Recognition With Fewer Stimuli: Finishing viewing 100 stimuli takes up about 10 min. Although 10 min is already acceptable for clinical application, accurate recognition with fewer stimuli would be better to save time cost and medical resources. For the eye movement-based SZ recognition task, the number of stimuli, i.e., the duration of the subject's participation, can be considered as the efficiency of methods, which is different from the common efficiency of methods, i.e., inference time. So far, there has been no metric designed to quantify the efficiency of eye movement-based SZ recognition, and the relevant studies on improving efficiency are also absent. Empirically speaking, more stimuli will lead to an increase in recognition performance [8], [47]. Contrastive learning is a possible solution to make the features with fewer stimuli closer to the features with more stimuli, so as to improve the efficiency of eye movement-based SZ recognition methods.

3) Clinical Interpretability: For eye movement-based SZ recognition, in addition to binary groups, i.e., SZ and HC groups, it is meaningful to refine the recognition results, such as the severity of illness. Moreover, it is essential to propose an explainable model associated with relevant symptoms, which will be helpful for disease analysis.

VI. Experimental Results of MSNet

In the benchmark, we briefly analyze the performance of all methods. Here, we analyze in detail the experimental results and advantages of our MSNet.

A. Comparison With SOTAs

As shown in Table IV, it can be observed that our MSNet reaches a remarkable classification performance. To be specific, on the validation set, our method is better than other methods on the five metrics of accuracy, sensitivity, specificity, precision, and F1-score and ranks second on AUC. As for the test set, the results are consistent with the ones on the validation set, where our MSNet shows competitive performance on accuracy, sensitivity, precision, AUC, and F1-score. In particular, our method outperforms the second-best method by 8.33% on accuracy (0.7500 $\rightarrow$ 0.8125). In terms of specificity, our method only ranks third on the test set. The reason is that sensitivity shows the true positive rate, and specificity represents the true negative rate. These methods with better specificity, i.e., EDB_BYS, EDB_QDA, and IAS, significantly sacrifice sensitivity. In contrast, our MSNet achieves a better balance between sensitivity and specificity.

TABLE V
ABLATION STUDY ON MAIN COMPONENTS OF MSNET. THE BEST RESULT OF EACH METRIC IS SHOWN IN BOLD

Models	Acc.	Sen.	Spe.	AUC	Pre.	F1-score
Baseline	0.6458	0.5417	0.7500	0.7240	0.6842	0.6047
+SFB	0.7292	0.7083	0.7500	0.7743	0.7391	0.7234
+SFB+CCB *w/o* $\mathbf{D}_w^i$	0.7708	0.7500	**0.7917**	0.8108	0.7826	0.7660
+SFB+CCB (Ours)	**0.8125**	**0.8333**	**0.7917**	**0.8854**	**0.8000**	**0.8163**

TABLE VI
ABLATION STUDY ON FEATURE EXTRACTION. THE BEST RESULT OF EACH METRIC IS SHOWN IN BOLD

Models	Acc.	Sen.	Spe.	AUC	Pre.	F1-score
DeepLabv3 [74]	0.7500	0.7083	**0.7917**	0.8160	0.7727	0.7391
RegSeg [75]	0.7500	0.8333	0.6667	0.8403	0.7143	0.7692
RINet [62]	0.7917	**0.8750**	0.7083	0.8333	0.7500	0.8077
Ours	**0.8125**	0.8333	**0.7917**	**0.8854**	**0.8000**	**0.8163**

TABLE VII
ABLATION STUDY ON THE STEP-BY-STEP MANNER. THE BEST RESULT OF EACH METRIC IS SHOWN IN BOLD

Models	Acc.	Sen.	Spe.	AUC	Pre.	F1-score
Stage 1	0.7083	0.7500	0.6667	0.7535	0.6923	0.7200
Stage 1,2	0.7500	0.7083	0.7917	0.8038	0.7727	0.7391
Stage 1,2,3	0.7708	0.7083	**0.8333**	0.8420	**0.8095**	0.7556
Stage 1,2,3,4 (Ours)	**0.8125**	**0.8333**	0.7917	**0.8854**	0.8000	**0.8163**

B. Ablation Studies

In this section, we conduct comprehensive ablation studies on the EMS test set. Specifically, we analyze the following: 1) the contributions of each key component in our MSNet; 2) the effectiveness of feature extraction based on DoF maps; and 3) the necessity of the step-by-step manner.

1) Contributions of Main Components: As shown in Table V, to investigate the contribution of the two branches, i.e., SFB and CCB, we establish four variants: 1) baseline: average the feature $\mathbf{F}$ along the number of stimuli and classify it using the classification head; 2) +SFB: average the output feature of Stage 4 in SFB and classify it using the classification head; 3) +SFB+CCB w/o $\mathbf{D}_w^i$: the complete MSNet without the weight function $\mathbf{D}_w^i$; and 4) +SFB+CCB: the complete MSNet. The quantitative results are reported in Table V.

We can observe that baseline only achieves 0.6458 on accuracy, 0.7240 on AUC, and 0.6047 on F1-score. SFB promotes baseline by 0.0834 (0.6458 $\rightarrow$ 0.7292) on accuracy, 0.0503 (0.7240 $\rightarrow$ 0.7743) on AUC, and 0.1187 (0.6047 $\rightarrow$ 0.7234) on F1-score. The +SFB+CCB w/o $\mathbf{D}_w^i$ variant evaluates the effect of the mean shift vector with uniform weight for all samples. The results show that the plain mean shift vector can improve the performance to some extent. In contrast, the mean shift vector with weight functions $\mathbf{D}_w^i$ (i.e., +SFB+CCB) can optimize the performance more. For example, the score of accuracy is promoted by 0.0833 (0.7292 $\rightarrow$ 0.8125) using the mean shift vector weighted by $\mathbf{D}_w^i$, while only 0.0416 (0.7292 $\rightarrow$ 0.7708) adopting the plain mean shift vector. Although the variant of +SFB+CCB does not further improve specificity, it promotes sensitivity effectively, also resulting in higher values on the other four metrics.

2) Ablation for Feature Extraction: Due to the restricted visual patterns, schizophrenics tend to miss some informative regions, resulting in semantic differences from the HC group. We try to transfer the semantic knowledge from other datasets to benefit the eye movement-based SZ recognition task. We deploy two popular semantic segmentation methods (i.e., DeepLabv3 [74] and RegSeg [75]) and the original RINet based on fixation prediction task for comparison in Table VI. For DeepLabv3, we utilize the publicly available models trained on the MS-COCO dataset [76] whose backbone is ResNet-101 [41]. For RegSeg, we adopt the released models trained on the Cityscapes dataset [77]. For the original RINet based on fixation prediction task, we employ the model trained on the SALICON dataset [73] and fine-tuned on the MIT1003 dataset [59]. As we observed, the initial stimulus features $\mathbf{F}$ extracted by semantic segmentation models are overall weaker than fixation prediction model (i.e., RINet). This is because the fixation prediction model, in addition to semantic information, provides extra knowledge about whether the fixations belong to HC group. In our MSNet, we adopt the RINet fine-tuned with the DoF maps of EMS dataset, which belong to the 160 subjects for cross validation, and achieve the best performance. In comparison with the fixation density maps, the DoF maps can describe the subtle differences between the two groups, so as to benefit the task of eye movement-based SZ recognition.

3) Ablation for Step-by-Step Manner: To study the necessity of the step-by-step manner, we modify the number of Stage-i and the corresponding CCS-i module and offer three variants, as shown in Table VII. By comparing the results of "Stage 1," "Stage 1,2," "Stage 1,2,3," and "Stage 1,2,3,4," it can be observed that the addition of the number of Stage-i and the corresponding CCS-i module realizes continuous performance improvements. For instance, the accuracy and AUC increase persistently (e.g., Acc.: 0.7083 $\rightarrow$ 0.7500 $\rightarrow$ 0.7708 $\rightarrow$ 0.8125 and AUC: 0.7535 $\rightarrow$ 0.8038 $\rightarrow$ 0.8420 $\rightarrow$ 0.8854). As the number of layers increases, the stimulus features are enhanced, and the more discriminative features are emphasized, so as to improve the performance. In other words, the step-by-step manner plays a vital role in our MSNet.

VII. CONCLUSION

In this article, we have presented a novel EMS, which is a large-scale and publicly available dataset and compensates for the lack of relevant datasets in the research community. It is based on the free-viewing paradigm with 100 carefully selected stimuli. The 104 schizophrenics and 104 HCs participated in this test. To ensure the privacy of the subjects, their identities and personal information have been removed. Furthermore, we conduct the first comprehensive evaluation of the relevant 13 methods and measure them by six metrics to act as a benchmark for this field. To integrate multiple stimuli effectively, we further proposed a novel MSNet for eye movement-based SZ recognition, which imitates the mean shift algorithm with convolution. Our MSNet consists of an SFB and a CCB. The SFB aggregates similar information for each stimulus feature to enhance the representative ability. The CCB chooses an appropriate stimulus feature as the initial cluster center in the CCG module and shifts it to the denser feature subspace in the CCS module. Comprehensive experimental

results have certified the better performance of our MSNet compared with 12 SOTAs.

ACKNOWLEDGMENT

The authors would like to thank the participation of subjects and doctors at Shanghai Mental Health Center and the subjects and students at Shanghai University.

REFERENCES

[1] S. Brissos, A. Molodynski, V. Dias, and M. Figueira, "The importance of measuring psychosocial functioning in schizophrenia," *Ann. Gen. Psychiatry*, vol. 10, no. 1, p. 18, 2011.

[2] S. M. Couture, "The functional significance of social cognition in schizophrenia: A review," *Schizophrenia Bull.*, vol. 32, no. 1, pp. S44–S63, Aug. 2006.

[3] S. J. Abplanalp et al., "Understanding connections and boundaries between positive symptoms, negative symptoms, and role functioning among individuals with schizophrenia: A network psychometric approach," *JAMA Psychiatry*, vol. 79, no. 10, pp. 1014–1022, Aug. 2022.

[4] T. H. McGlashan, "Early detection and intervention of schizophrenia: Rationale and research," *Brit. J. Psychiatry*, vol. 172, no. S33, pp. 3–6, Jun. 1998.

[5] J. Cooper, "Diagnostic and statistical manual of mental disorders (4th edn, text revision) (DSM-IV-TR)," *Br. J. Psychiatry*, vol. 179, p. 85, Jun. 2001.

[6] M. Lavelle, P. G. T. Healey, and R. McCabe, "Nonverbal behavior during face-to-face social interaction in schizophrenia: A review," *J. Nervous Mental Disease*, vol. 202, no. 1, pp. 47–54, Jan. 2014.

[7] S. Wang, X. Ouyang, T. Liu, Q. Wang, and D. Shen, "Follow my eye: Using gaze to supervise computer-aided diagnosis," *IEEE Trans. Med. Imag.*, vol. 41, no. 7, pp. 1688–1698, Jul. 2022.

[8] C. Xia et al., "Dynamic viewing pattern analysis: Towards large-scale screening of children with ASD in remote areas," *IEEE Trans. Biomed. Eng.*, vol. 70, no. 5, pp. 1622–1633, May 2023.

[9] H. Yang et al., "An automatic detection method for schizophrenia based on abnormal eye movements in reading tasks," *Expert Syst. Appl.*, vol. 238, Mar. 2024, Art. no. 121850.

[10] D. Zhang et al., "Eye movement indices as predictors of conversion to psychosis in individuals at clinical high risk," *Eur. Arch. Psychiatry Clin. Neurosci.*, vol. 273, no. 3, pp. 553–563, Jul. 2022.

[11] K. N. Thakkar, V. A. Diwadkar, and M. Rolfs, "Oculomotor prediction: A window into the psychotic mind," *Trends Cognit. Sci.*, vol. 21, no. 5, pp. 344–356, May 2017.

[12] P. E. G. Bestelmeyer, B. W. Tatler, L. H. Phillips, G. Fraser, P. J. Benson, and D. St. Clair, "Global visual scanning abnormalities in schizophrenia and bipolar disorder," *Schizophrenia Res.*, vol. 87, nos. 1–3, pp. 212–222, Oct. 2006.

[13] M. J. Green, J. H. Waldron, I. Simpson, and M. Coltheart, "Visual processing of social context during mental state perception in schizophrenia," *J. Psychiatry Neurosci.*, vol. 33, no. 1, pp. 34–42, Jan. 2008.

[14] N. Sengupta, C. B. McNabb, N. Kasabov, and B. R. Russell, "Integrating space, time, and orientation in spiking neural networks: A case study on multimodal brain data modeling," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5249–5263, Nov. 2018.

[15] L.-L. Zeng et al., "Gradient matching federated domain adaptation for brain image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 1, no. 1, pp. 1–15, Nov. 2022.

[16] Z.-A. Huang, R. Liu, Z. Zhu, and K. C. Tan, "Multitask learning for joint diagnosis of multiple mental disorders in resting-state fMRI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 1, no. 1, pp. 1–15, Feb. 2022.

[17] R. Liu, Z.-A. Huang, Y. Hu, Z. Zhu, K.-C. Wong, and K. C. Tan, "Attention-like multimodality fusion with data augmentation for diagnosis of mental disorders using MRI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 1, no. 1, pp. 1–15, Nov. 2022.

[18] R. Liu, Z.-A. Huang, Y. Hu, Z. Zhu, K.-C. Wong, and K. C. Tan, "Spatial–temporal co-attention learning for diagnosis of mental disorders from resting-state fMRI data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 1, no. 1, pp. 1–15, Feb. 2023.

[19] S. Siuly, S. K. Khare, V. Bajaj, H. Wang, and Y. Zhang, "A computerized method for automatic detection of schizophrenia using EEG signals," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 11, pp. 2390–2400, Nov. 2020.

[20] Z.-A. Huang, J. Zhang, Z. Zhu, E. Q. Wu, and K. C. Tan, "Identification of autistic risk candidate genes and toxic chemicals via multilabel learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 9, pp. 3971–3984, Sep. 2021.

[21] Y. Zhang, H. Zhang, L. Xiao, Y. Bai, V. D. Calhoun, and Y.-P. Wang, "Multi-modal imaging genetics data fusion via a hypergraph-based manifold regularization: Application to schizophrenia study," *IEEE Trans. Med. Imag.*, vol. 41, no. 9, pp. 2263–2272, Sep. 2022.

[22] M. Bishay, P. Palasek, S. Priebe, and I. Patras, "SchiNet: Automatic estimation of symptoms of schizophrenia from facial behaviour analysis," *IEEE Trans. Affect. Comput.*, vol. 12, no. 4, pp. 949–961, Oct. 2021.

[23] A. Davies, E. Fried, H. Aung, and O. Costilla-Reyes, "Individual behavioral insights in schizophrenia: A network analysis and mobile sensing approach," 2023, *arXiv:2312.01216*.

[24] P. J. Benson, S. A. Beedie, E. Shephard, I. Giegling, D. Rujescu, and D. S. Clair, "Simple viewing tests can detect eye movement abnormalities that distinguish schizophrenia cases from controls with exceptional accuracy," *Biol. Psychiatry*, vol. 72, no. 9, pp. 716–724, Nov. 2012.

[25] L. Huang et al., "Effective schizophrenia recognition using discriminative eye movement features and model-metric based features," *Pattern Recognit. Lett.*, vol. 138, pp. 608–616, Oct. 2020.

[26] K. Fukunaga and L. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," *IEEE Trans. Inf. Theory*, vols. IT-21, no. 1, pp. 32–40, Jan. 1975.

[27] Y. Cheng, "Mean shift, mode seeking, and clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 8, pp. 790–799, Aug. 1995.

[28] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[29] K. Morita et al., "Eye movement as a biomarker of schizophrenia: Using an integrated eye movement score," *Psychiatry Clin. Neurosciences*, vol. 71, no. 2, pp. 104–114, Feb. 2017.

[30] J. Kacur, J. Polec, E. Smolejova, and A. Heretik, "An analysis of eye-tracking features and modelling methods for free-viewed standard stimulus: Application for schizophrenia detection," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 11, pp. 3055–3065, Nov. 2020.

[31] A. R. Diefendorf and R. Dodge, "An experimental study of the ocular reactions of the insane from photographic records," *Brain*, vol. 31, no. 3, pp. 451–489, 1908.

[32] P. S. Holzman, "Eye-tracking dysfunctions in schizophrenic patients and their relatives," *Arch. Gen. Psychiatry*, vol. 31, no. 2, p. 143, Aug. 1974.

[33] P. S. Holzman, C. M. Solomon, S. Levin, and C. S. Waternaux, "Pursuit eye movement dysfunctions in schizophrenia: Family evidence for specificity," *Arch. Gen. Psychiatry*, vol. 41, no. 2, pp. 136–139, Feb. 1984.

[34] S. A. Beedie, D. M. St. Clair, and P. J. Benson, "Atypical scanpaths in schizophrenia: Evidence of a trait- or state-dependent phenomenon?" *J. Psychiatry Neurosci.*, vol. 36, no. 3, pp. 150–164, May 2011.

[35] T. Zhang et al., "Inefficient integration during multiple facial processing in pre-morbid and early phases of psychosis," *World J. Biol. Psychiatry*, vol. 23, no. 5, pp. 361–373, Dec. 2021.

[36] M. Jiang and Q. Zhao, "Learning visual attention to identify people with autism spectrum disorder," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3287–3296.

[37] S. Wang et al., "Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking," *Neuron*, vol. 88, no. 3, pp. 604–616, Nov. 2015.

[38] J. L. Taylor and C. H. John, "Attentional and memory bias in persecutory delusions and depression," *Psychopathology*, vol. 37, no. 5, pp. 233–241, Oct. 2004.

[39] H. Duan et al., "A dataset of eye movements for the children with autism spectrum disorder," in *Proc. 10th ACM Multimedia Syst. Conf.*, Jun. 2019, pp. 255–260.

[40] C. Wu, S. Liaqat, S.-C. Cheung, C.-N. Chuah, and S. Ozonoff, "Predicting autism diagnosis using image with fixations and synthetic saccade patterns," in *Proc. IEEE Int. Conf. Multimedia Expo. Workshops (ICMEW)*, May 2019, pp. 647–650.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[42] Y. Tao and M.-L. Shyu, "SP-ASDNet: CNN-LSTM based ASD classification model using observer scanpaths," in *Proc. IEEE Int. Conf. Multimedia Expo. Workshops (ICMEW)*, Jul. 2019, pp. 641–646.

[43] W. Wei et al., "Identify autism spectrum disorder via dynamic filter and deep spatiotemporal feature extraction," *Signal Process., Image Commun.*, vol. 94, May 2021, Art. no. 116195.
[44] P. Lanillos, D. Oliva, A. Philippsen, Y. Yamashita, Y. Nagai, and G. Cheng, "A review on neural network models of schizophrenia and autism spectrum disorder," *Neural Netw.*, vol. 122, pp. 338–363, Feb. 2020.
[45] D. Zhang et al., "Effective differentiation between depressed patients and controls using discriminative eye movement features," *J. Affect. Disorders*, vol. 307, pp. 237–243, Jun. 2022.
[46] J. Škunda, J. Polec, B. Nerusil, and E. Málisová, "Schizophrenia detection using convolutional neural network," in *Proc. Int. Symp. ELMAR*, Sep. 2021, pp. 151–154.
[47] J. Xie et al., "A two-stream end-to-end deep learning network for recognizing atypical visual attention in autism spectrum disorder," 2019, *arXiv:1911.11393*.
[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[49] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vols. IT-28, no. 2, pp. 129–137, Mar. 1982.
[50] H. Wang, M. Yao, G. Jiang, Z. Mi, and X. Fu, "Graph-collaborated auto-encoder hashing for multiview binary clustering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 1, no. 1, pp. 1–13, Jan. 2023.
[51] P. Zeng, H. Zhang, L. Gao, X. Li, J. Qian, and H. T. Shen, "Visual commonsense-aware representation network for video captioning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 1, no. 1, pp. 1–12, Dec. 2023.
[52] J. Kauffmann, M. Esders, L. Ruff, G. Montavon, W. Samek, and K.-R. Muller, "From clustering to cluster explanations via neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 2, pp. 1–15, Jul. 2022.
[53] M. Zheng, P. Gao, R. Zhang, K. Li, H. Li, and H. Dong, "End-to-end object detection with adaptive clustering transformer," in *Proc. BMVC*, Nov. 2020, p. 226.
[54] Q. Yu et al., "CMT-DeepLab: Clustering mask transformers for panoptic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2550–2560.
[55] Q. Yu et al., "K-means mask transformer," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 288–307.
[56] W. Zeng et al., "Not all tokens are equal: Human-centric visual analysis via token clustering transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11091–11101.
[57] M. Du, S. Ding, and H. Jia, "Study on density peaks clustering based on k-nearest neighbors and principal component analysis," *Knowledge-Based Syst.*, vol. 99, pp. 135–145, May 2016.
[58] B. R. Manor et al., "Eye movements reflect impaired face processing in patients with schizophrenia," *Biol. Psychiatry*, vol. 46, no. 7, pp. 963–969, Oct. 1999.
[59] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 2106–2113.
[60] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao, "Predicting human gaze beyond pixels," *J. Vis.*, vol. 14, no. 1, pp. 1–28, Jan. 2014.
[61] G. H. Patel et al., "What you see is what you get: Visual scanning failures of naturalistic social scenes in schizophrenia," *Psychol. Med.*, vol. 51, no. 16, pp. 2923–2932, Jun. 2020.
[62] Y. Song et al., "RINet: Relative importance-aware network for fixation prediction," *IEEE Trans. Multimedia*, vol. 25, pp. 1–15, 2023.
[63] J. Lei Ba, J. Ryan Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.
[64] Y. Liu et al., "A survey of visual transformers," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 1, no. 1, pp. 1–21, Mar. 2023.
[65] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inform. Process. Syst. (NIPS)*, 2017, pp. 5998–6008.
[66] Y. Lin, H. Ma, Z. Pan, and R. Wang, "Depression detection by combining eye movement with image semantics," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 269–273.
[67] W. Zhou, M. Yang, J. Tang, J. Wang, and B. Hu, "Gaze patterns in children with autism spectrum disorder to emotional faces: Scanpath and similarity," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 32, pp. 865–874, 2024.
[68] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. NIPS*, Dec. 2019, pp. 8024–8035.
[69] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, "Training data-efficient image transformers distillation through attention," in *Proc. ICML*, vol. 139, 2021, pp. 10347–10357.
[70] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, Dec. 2018, pp. 1–10.
[71] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. ICLR*, Feb. 2017, pp. 1–16.
[72] J. Hyönä, J. Tommola, and A.-M. Alaja, "Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks," *Quart. J. Experim. Psychol. Sect. A*, vol. 48, no. 3, pp. 598–612, Aug. 1995.
[73] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "SALICON: Saliency in context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1072–1080.
[74] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
[75] R. Gao, "Rethinking dilated convolution for real-time semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 4675–4684.
[76] T. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
[77] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.

Yingjie Song received the B.E. degree from Shanghai University, Shanghai, China, in 2019. She is currently pursuing the Ph.D. degree with the School of Communication and Information Engineering, Shanghai University, Shanghai.

Her research interests include fixation prediction and autism spectrum disorder/schizophrenia recognition.

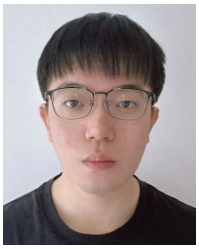
Zhi Liu (Senior Member, IEEE) received the B.E. and M.E. degrees from Tianjin University, Tianjin, China, and the Ph.D. degree from the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China, in 1999, 2002, and 2005, respectively.

From August 2012 to August 2014, he was a Visiting Researcher with the SIROCCO Team, IRISA/INRIA-Rennes, Rennes, France, with the support by EU FP7 Marie Curie Actions. He is currently a Professor at the School of Communication and Information Engineering, Shanghai University, Shanghai. He has published more than 200 refereed technical papers in international journals and conferences. His research interests include image/video processing, machine learning, computer vision, and multimedia communication.

Dr. Liu is an Area Editor of *Signal Processing: Image Communication* and served as a Guest Editor for the Special Issue on Recent Advances in Saliency Models, Applications and Evaluations in *Signal Processing: Image Communication*.

Gongyang Li received the Ph.D. degree from Shanghai University, Shanghai, China, in 2022.

From 2021 to 2022, he was a Visiting Ph.D. Student with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. From 2022 to 2024, he was a Post-Doctoral Researcher at Shanghai University, where he is currently an Associate Professor at the School of Communication and Information Engineering. His research interests include saliency detection, multimodel image processing, and image/video segmentation.



Jiawei Xie received the B.E. and M.E. degrees from Shanghai University, Shanghai, China, in 2021 and 2024, respectively.

His research interests include computer vision and image/video processing.

Qiang Wu (Senior Member, IEEE) received the B.Eng. and M.Eng. degrees from Harbin Institute of Technology, Harbin, China, in 1996 and 1998, respectively, and the Ph.D. degree from the University of Technology Sydney, Ultimo, NSW, Australia, in 2004.

He is currently an Associate Professor and a Core Member with the Global Big Data Technologies Centre, University of Technology Sydney. His research outcomes have been published in many premier international conferences, including ECCV, CVPR, ICCV, ICIP, and ICPR, and the major international journals, such as IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, IEEE TRANSACTIONS ON MULTIMEDIA, *International Journal of Computer Vision*, *Public Relations*, and *Physical Review Letters*. His research interests include computer vision, image processing, pattern recognition, machine learning, and multimedia processing. The application fields where the research outcomes are applied span over video security surveillance, biometrics, video data analysis, and human–computer interaction.

Dan Zeng (Senior Member, IEEE) received the B.S. degree in electronic science and technology and the Ph.D. degree in circuits and systems from the University of Science and Technology of China, Hefei, China, in 2003 and 2008, respectively.

She is a Full Professor and the Dean of the Department of Communication Engineering, Computer Vision and Pattern Recognition Laboratory, Shanghai University, Shanghai, China. Her main research interests include computer vision, multimedia analysis, and machine learning.

Dr. Zeng is serving as a TC Member for IEEE MSA and IEEE MMSP. She is serving as an Associate Editor for IEEE TRANSACTIONS ON MULTIMEDIA and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY.

Lihua Xu received the M.D. degree from Soochow University, Suzhou, China, in 2011, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2017.

She is currently working as a Psychiatrist and an Attending Doctor. Her research direction is the biomarker research of high risk syndrome of psychosis. She is responsible for the clinical evaluation, follow-up, and data management of high risk syndrome of psychosis.

Tianhong Zhang is a Senior Psychiatrist and a Professor with Shanghai Jiao Tong University School of Medicine, Shanghai, China; the Director of Shanghai Engineering Research Center of Intelligent Psychological Evaluation and Intervention, Shanghai; a Clinical PI for ShangHai At Risk for Psychosis (SHARP) Program, Shanghai Mental Health Centre, Shanghai; and the M.D. and Ph.D.'s Tutor at Shanghai Jiao Tong University School of Medicine.

Dr. Zhang is a Secretary and a Member of CSNP Schizophrenia Research Alliance, a Member of Schizophrenia Collaborative Group of Psychiatric Society of Chinese Medical Association, and various roles in Editorial Board in Psychiatry Research, BMC Psychiatry, Frontier in Psychiatry, and Schizophrenia Bulletin Open.

Jijun Wang received the M.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 1994, and the Ph.D. degree from Ryukyu University, Japan, in 2004.

He is a Chief Physician and the Doctoral Supervisor of Shanghai Mental Health Center (mental health center affiliated with Shanghai Jiao Tong University School of Medicine, Shanghai), the Director of Brain Film Image Eye Movement Research Office, and a PI of Shanghai Heavy Mental Disease Laboratory, Shanghai.

Dr. Wang is a Member of Psychiatry Basic and Clinical Branch of Chinese Neuroscience Society, a Member of EEG and EMG Branch of Shanghai Medical Association, and the Director of Youth Committee of Shanghai Overseas Chinese joint committee. He is a Reviewer of various international journals, including the *Cochrane Database Systematic Reviews*, *Biological Psychology*, and *International Journal of Psychiatry*. He is the Editor of the *Journal of Psychiatry* and *BMC Psychiatry*.