Detecting Major Depression Disorder with Multiview Eye Movement Features in a Novel Oil Painting Paradigm

Tian-Fang Ma[1], Lu-Yu Liu[1], Li-Ming Zhao[2], Dan Peng[3], Yong Lu[3], Wei-Long Zheng[1], Bao-Liang Lu[1,2,3,*]
[1] Center for Brain-Like Computing and Machine Intelligence
Department of Computer Science and Engineering
Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering
Shanghai Jiao Tong University, Shanghai, 200240, China
[2] Emotion Helper, Shanghai, China
[3] RuiJin-Mihoyo Laboratory, Clinical Neuroscience Center, RuiJin Hospital
Shanghai Jiao Tong University School of Medicine, Shanghai, 200020, China

2024 International Joint Conference on Neural Networks (IJCNN) | 979-8-3503-5931-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/IJCNN60899.2024.10650558

***Abstract*—Major Depressive Disorder (MDD) is a debilitating condition marked by persistent low mood, reduced interest, cognitive impairments, and vegetative neurological symptoms such as sleep and appetite disturbances. In this paper, we collected eye movement signals from 40 patients diagnosed with MDD and 40 healthy controls to study the relation between eye movements and cognitive processes for depression detection. The eye movement data were captured during a novel emotional cognition task using oil paintings. Subsequently, the data were transformed into multiview eye movement features, including heatmaps, trajectories, and statistical vectors. Rigorous statistical analyses were then conducted on these features to identify significant patterns and correlations between eye movements and depressive symptoms. A multiview invariant & specific eye movement model (MISEYE) was proposed to fuse different eye movement features. The proposed achieved an accuracy rate of 79.88% in depression detection. This performance surpassed not only the outcomes of single-mode approaches and combinations of any two features but also outperformed other fusion methodologies. These findings not only shed light on the intricate relationship between eye movement patterns and MDD but also underscore the potential of eye-tracking technology in psychiatric research.**

***Index Terms*—major depressive disorder, eye movement, oil painting**

I. Introduction

Major depressive disorder (MDD) is a debilitating condition characterized by at least one discrete depressive episode lasting for a minimum of two weeks, frequently accompanied by dysregulation in cognitive function and emotion regulation, including impaired cognitive control, cognitive bias, and abnormal use of emotion regulation strategies [1], [2]. As a common mental illness, it has seriously affected the social function and quality of life of patients. In clinical practice, the detection, diagnosis, and treatment of MDD are challenging due to its diverse clinical manifestations, the difficulty in predicting its course and prognosis, and the variable response of patients to treatment [3].

The etiology of depression remains unclear due to the involvement of various psychological factors during its onset, as well as the existence of multiple subtypes within the disorder, each with distinct biological mechanisms. While a consensus regarding biomarkers for depression has yet to be reached, numerous scientific investigations have sought potential markers within physiological signals, including electroencephalography (EEG), functional magnetic resonance imaging (fMRI), and functional near-infrared spectroscopy (fNIRS) [4]–[6]. Additionally, some studies have explored the predictive value of external indicators, such as speech patterns or social media usage, in identifying depression [7], [8].

As an important form of human physiological signals, eye movement signals have been increasingly used in cognitive-related diseases in recent years. Eye movement signal is a kind of recording of eye movement. Compared with other biological signals, it is less costly, less noisy, easily available, and not limited by location. Eye movements not only reveal visual information selectively acquired based on moments but are also closely linked to cognitive processes [9]. Because these brain circuits that control eye movement are highly related to cognitive functions, eye movement signals could be a predictor of cognitive impairment. In the medical field, eye movement has been used to study cognitive diseases, including schizophrenia, Parkinson's disease, depression, Lewy body dementia, and so on [10]–[13].

As eye movement signals are highly correlated with what people are watching, paradigms are very important in eye movement research. This study introduced a novel visual cognitive paradigm involving the observation of oil paintings, with a focus on recording and analyzing the eye movements

This work was supported in part by grants from STI 2030-Major Projects+2022ZD0208500, National Natural Science Foundation of China (Grant No. 62376158), Shanghai Municipal Science and Technology Major Project (Grant No. 2021SHZD ZX), Shanghai Pujiang Program (Grant No. 22PJ1408600), Medical-Engineering Interdisciplinary Research Foundation of Shanghai Jiao Tong University "Jiao Tong Star" Program (YG2023ZD25, YG2024ZD25, YG2024QNA03), Shanghai Municipal Science and Technology Artificial Intelligence Support Special Project (Grant No. 22511106002), and GuangCi Professorship Program of RuiJin Hospital Shanghai Jiao Tong University School of Medicine.

* Corresponding author.

of both individuals with MDDs and healthy controls (HCs). In the past eye movement studies, most of the paradigms were resting-state [14], smooth pursuit and antisaccade tasks [15], [16], image stimulation and free viewing task [17], [18], and video stimulation [19].

The traditional paradigm has some limitations. Resting experiments and antisaccade tasks are mainly suitable for detecting eye diseases. Photographs and video stimuli are used to study emotions. However, the degree of emotion elicitation was low in photographs. Although the evoked degree of video is better than that of photos, the motion within video clips significantly influences ocular movements.

For humans, the emotional experiences induced by aesthetic encounters differ from those evoked by photographs. The perception, interpretation, and subsequent emotional responses to art are influenced by factors such as the viewer's familiarity, complexity, curiosity, and appreciation of aesthetics [20]. Given the emotional cognitive impairments commonly observed in individuals with depression, we hypothesized that the examination and rating of oil paintings with varying emotional attributes could offer valuable insights into the cognitive distinctions between depressed and non-depressed individuals.

In this study, we scrutinized the distinctions between MDD patients and healthy controls by processing raw eye movement data into eye heatmaps, trajectories, and statistical vectors. Initially, a thorough examination of eye movement parameters was undertaken, encompassing multiple dimensions within the paradigm. Subsequent to this, pertinent eye movement features were meticulously chosen based on statistical insights, serving as representative markers. Ultimately, we devised the (MIS-EYE) amalgamating eye movement heatmaps, trajectory, and features to detect MDD.

The main contributions of this paper can be summarized as follows:

- We introduced an innovative oil painting paradigm and verified its effectiveness in MDD detection.
- We implemented fusion model, integrating multiview eye movement data including thermograms, trajectories, and statistical vectors, culminating in superior classification accuracy between MDD patients and healthy control groups.

II. Related Work

A. Eye Movement Studies in MDD

Eye movement studies have shed light on cognitive biases observed in MDD patients. Studies have reported eye movement abnormalities in individuals with MDD compared to healthy controls. Takahashi et al. proposed that the saccade path of depressed patients was significantly shorter in the free-viewing paradigm [18]. Alghowinem et al. found differences in eye-opening time and blink time in people with depression. They extracted multi-dimensional statistical vectors from the face videos and used GMM and SVM to predict depression [21]. Wang et al. examined impairments in basic features of fixations and saccades in MDD and bipolar disorder (BPD) in smooth pursuit and free-viewing paradigms [22].

Eye movement behavior is closely related to the task. In past eye movement studies of depression, the specific tasks included multiple paradigms. The traditional paradigms used include resting state experiment, saccade-related paradigm (antisaccade task), smooth pursuit task, free-view paradigm, and visual cognitive tasks. Some researchers also combine multiple paradigms. Li et al. used three types of taks: fixation stability task, saccade task, and free-view task an investigated the differences in eye movement metrics between depressed patients and healthy controls in these paradigms. Result showed that the performance of some indicators in the two groups of people was not the same in the different paradigms. Number of saccades are significantly higher for the depression group in the fixation stability task while it is lower for the depression group in the free-view task.

Previous research has confirmed that MDD differs from healthy people in eye movement characteristics in many ways. In this paper, we have made several improvements over those previous studies. For the first time in a study of depression, we subdivided saccade categories to extract and summarize eye movement features from a more comprehensive perspective. Additionally, for the first time, we analyzed eye movement heatmaps and fused them with features for depression detection.

B. The Application of Transformer Model in Medicine

Transformer is a neural network architecture based on self-attention mechanism, which has achieved great success in natural language processing (NLP) and other fields [23]. Vision Transformer (ViT), a derivative model of Transformer, has successfully applied Transformer architecture to computer vision tasks, which has made remarkable achievements in the medical field and promoted technological progress in medical image analysis, pathological diagnosis, and medical image processing [24]–[26].

A major problem with deep networks of images is that they require a large amount of labeled data for training. The pre-training algorithm represented by Masked Autoencoder (MAE) uses unlabeled data to solve this problem [27]. For the image in painting domain, MAEs are employed to fill in missing or damaged regions in an image. By masking out portions of the image and training the MAE to reconstruct the original content, the model can effectively restore the missing information, making it valuable for image restoration and completion tasks. By using MAE, the results of image segmentation and multi-label classification in the medical field have been further improved [28].

C. Model Fusion Strategy

Fusion algorithms have become a prominent area of research in the field of data integration and decision-making. The fusion methods are commonly divided into four types: feature-level fusion, decision-level fusion, mixture-level fusion, and model-level fusion. Feature-level fusion concatenates

TABLE I. Demographics and the scores of the self-rating scales and MATRICS consensus cognitive battery of the subjects

Variable	MDDs (n = 40)	HCs (n = 40)	z	P-value
Age(year)	23.49 ± 4.34	24.98 ± 4.23	-1.57	.121
Gender	20F / 20M	20F / 20M	-	-
HAMD-17	17.83 ± 4.72	NAN	-	-
CES-D	37.558 ± 11.93	8.08 ± 4.59	**14.00**	**<.001**
PHQ-9	16.93 ± 6.71	2.10 ± 1.79	**13.50**	**<.001**
GAD-7	11.80 ± 5.81	2.02 ± 1.86	**10.14**	**<.001**
TAS-20				
F1-score	24.45 ± 5.91	10.75 ± 3.58	**12.55**	**<.001**
F2-score	17.48 ± 3.43	9.33 ± 2.71	**11.80**	**<.001**
F3-score	22.15 ± 3.44	17.25 ± 4.49	**5.48**	**<.001**
Total	64.01 ± 9.61	37.33 ± 7.72	**13.72**	**<.001**
SHAPS	32.65 ± 6.26	18.33 ± 4.60	**11.66**	**<.001**

*MDDs: major depressive disorder patients; HCs: healthy controls.

the features extracted from different modalities into a single high-dimensional feature vector immediately after extraction, using methods including Principal component analysis (PCA) and maximum relevance and minimum redundancy algorithm (mRMR) to remove redundant information. Decision level fusion, after obtaining a decision based on each modality, is achieved by applying algebraic combination rules of multiple predictive class labels. Decision level fusion combines the previous two fusion methods. The implementation of model-level fusion mainly depends on the fusion model used.

III. Experiment

A. Subjects

A total of 40 individuals diagnosed with MDD and 40 HCs were recruited. Each group consisted of equal numbers of males and females. The 40 MDD patients were recruited from three hospitals and HCs were recruited from both hospitals and universities. All 80 participants had normal hearing and vision and did not present with any eye diseases or defects. The experiment was conducted by the local Ethics Committee, and all participants were fully informed about the study procedures and provided written consent.

All 80 participants in the experiment were rigorously diagnosed by doctors, and were asked if they were willing to participate in the experiment. Both patients and HCs were asked to complete several self-evaluation scales, including the Center for Epidemiologic Studies Depression Scale (CESD), the Patient Health Questionnaire-9 (PHQ-9), the Generalized Anxiety Disorder Assessment (GAD-7), the Toronto Alexithymia Scale (TAS-20), and the Pittsburgh Sleep Quality Index (PSQI). These assessments were administered under the guidance of clinical psychiatrists. MDDs were assessed by psychiatrists utilizing the International Classification of Diseases, Tenth Revision (ICD-10) criteria, and the Hamilton Rating Scale for Depression, 17-item version (HAMD-17). Patients were included in the MDD group when the psychiatrist identified a potential depressive condition during the interview and confirmed a major depressive rating based on the self-rating scale evaluation. On the other hand, subjects who did not display any signs of depression tendencies during the interview and self-rating scale assessment were included in the healthy control group. Nonparametric Wilcoxon rank-sum tests were performed for all measure scores. The z-scores and P-values were listed in Table I.

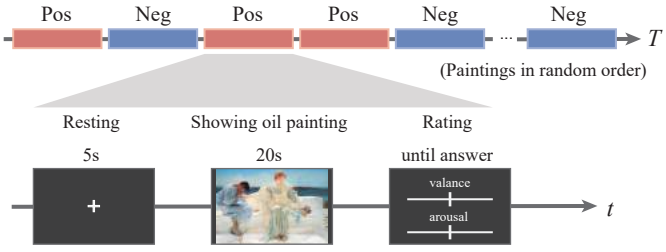

Fig. 1. The procedure of oil painting paradigm: subjects viewed the oil paintings and rated them in random order.

B. Experiment Procedure

The experimental paradigm required all participants to watch 40 oil paintings followed by the completion of an emotional cognitive task after each painting (Fig. 1). Luo et al. used 126 oil paintings to classify emotions as positive, negative, and neutral, which proved the effectiveness of oil paintings on human emotion induction [29]. For this paradigm, we recruited 20 people unrelated to the experiment to rate the emotions of these 126 oil paintings, and selected 40 oil paintings with the highest positive and negative scores. During the experimental procedure, the 40 paintings were presented in a random order, with each painting displayed for 20 seconds. Participants were given a 5-second resting period before the next painting was shown. Subsequently, participants were required to rate the paintings they had just seen based on valence and arousal dimensions (-5 to 5).

C. Eye Data Recording

The experiment was administered on Tobii-pro Fusion eye-tracker. The eye-tracker monitor was fixed on a monitor with a resolution of 2560*1440. Participants sat approximately 60-70 cm from the monitor. Tobii Studio's standard calibration was used. Different oil paintings have different proportions of width and height. Each oil painting stimulus was scaled to fit the monitor screen.

IV. Methods

A. Feature Collection

Eye movement analysis shows great potential in the early diagnosis and monitoring of neurological disorders such as Parkinson's disease, Alzheimer's disease, and autism [30], [31]. Through the examination of eye movement features, valuable insights into abnormal nervous system functioning can be gleaned. Additionally, eye movement carries cognitive and emotional implications, offering valuable clues about an individual's attention, information processing, and emotional experiences. Given that cognitive impairment and depression

are prominent symptoms of MDD [32], we posited that eye movement analysis holds potential for significant applications in depression research. It could aid in comprehending the pathological mechanisms underlying depression and facilitate diagnostic processes.

The rationale behind selecting oil paintings with varying emotional content as the paradigm lay in the notion that assessing the emotional qualities of such paintings could effectively engage individuals' comprehension and cognition of the artwork. Consequently, we could collect eye movement data from both depressed patients and healthy individuals during their emotional perception. We calculated five parameters of eye movement information, such as fixation, saccade, blink, and pupil dilation. Each parameter was carefully examined, leading to the identification of distinctive subcategories that can serve as differentiating indices.

1) Fixation & regions of interest. Fixation is widely used in cognitive psychology to study people's attention, perception, and cognitive processes. It is primarily controlled by circuits of superior colliculus, cerebellum, and reticular formation and saccades and blinks are primarily associated with visual cortex. Within the paradigm featuring oil paintings, distinct boundaries were delineated for the eye, mouth, and face of each painting. We obtained the subjects' attention to the whole picture through the fixation duration, fixation frequency, and other statistical information, and the attention to different interest regions was reflected in the eye movement heatmap.

2) Microsaccade. Microsaccades are fixational saccades that happens during fixation process. For microsaccade filtering, we use a $\lambda = 5$ velocity threshold and an 8 ms minimal duration to detect microsaccades(due to the sampling rate of Tobii Pro Fusion eye tracker) [33]. For each microsaccade detected, eye movement speed and direction were calculated.

3) Macrosaccade. Macrosaccades are characterized as sudden saccadic movements that disrupt fixation [34]. For each macrosaccade, we computed its duration, amplitude, and derived the corresponding macrosaccade velocity. Unlike microsaccades, we did not count macrosaccade direction as a statistical vectors because the direction in which the microsaccades occurred spontaneously. However, the direction in which saccades occurred was influenced by the content of the viewed painting. By contrast, although the magnitude of macrosaccades is related to the content of the painting, macrosaccade duration and velocity are much less affected by the painting content.

4) Blink. Blinking represents a spontaneous eye movement. To capture and record all instances of blinking during the viewing of oil paintings, we implemented a blink filter with an interval ranging from 0.1s to 0.4s [35].

5) Pupil dilation. Because tricyclic antidepressants taken by MDD patients could dilate the pupils, we did not analyze the mean of the pupil diameter in the task. We examined the relative changes in pupil diameter as our primary measure. Iris tremor refers to a subtle oscillation observed in the pupil diameter within the frequency range of 0.05 to 0.3Hz, with an amplitude of 1mm. Research indicates that during a state of relaxation or negative emotions, the pupil diameter exhibits the iris tremor effect within the low-frequency band. However, this effect diminishes when individuals engage in mental activities [36]. To characterize the iris tremor effect, we derived the power spectral density of the pupil diameter within the frequency bands of [0,1] Hz and calculated its corresponding differential entropy features.

From the previous analysis, we defined five types of gaze features: fixation features, microsaccade features, macrosaccade features, blink features, and pupil diameter features. The description for each feature type is as follows:

- Fixation - fixation frequency (n/second), average fixation duration (second), fixation duration per second (second);
- Microsaccade - microsaccade frequency (n/second), average microsaccade velocity (°/second);
- Macrosaccade - macrosaccade frequency (n/second), average macrosaccade duration (second), average macrosaccade velocity (°/second), average macrosaccade amplitude (°);
- Blink - blink frequency (n/second), average blink duration (second), blink duration standard deviation (second);
- Pupil diameter - power spectral density and differential entropy of 0-1 Hz (bit).

All the features formed a vector of 16 dimensions, which were used as the input to the model.

B. Heatmap

On the basis of the hypothesis that there are differences in attention to different regions of interest between MDDs and HCs, we transformed each subject's attention to each oil painting into a heatmap to distinguish the difference,(Fig. 2). For each subject viewing data in an oil painting, the raw data of all fixation points were converted into a hist of dimensions of w and h (w and h denote the width and height resolution of the screen). The value corresponding to each point in hist represents the length of time fixated on that point.

The heatmaps were generated from the hist with 2D Gaussian filter. The 2D Gaussian kernel is:

$$G(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}}, \tag{1}$$

where x and y are the independent variables representing the coordinate values along the x and y axes, respectively, in this case, the width and height resolution. σ_x and σ_y are the standard deviation along the x and y axis. To keep the scale uniform, we define sigma as the width and height of the screen resolution multiplied by a coefficient α, thus $\sigma_x = \alpha w$ and $\sigma_y = \alpha h$. We use $\alpha = 0.05$ in this paper.

C. Trajectory

Eye-tracking data are time series data. We generated eye trajectory data using the X-and Y-axis coordinates of the eye gaze points on the two-dimensional coordinate axis of the screen (Fig. 2). Each oil painting was viewed for 20 seconds, so the duration of the eye-tracking data was 20 seconds, and the axis dimension was measured in pixels.

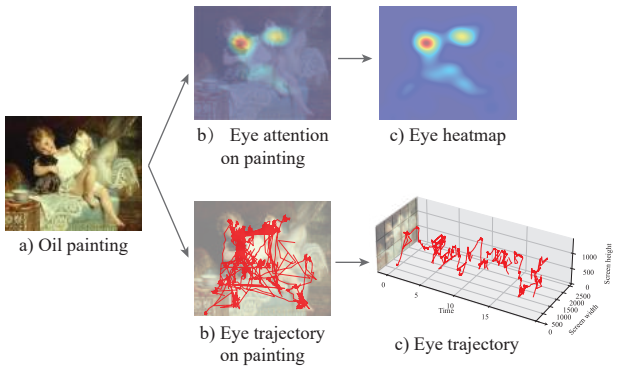

Fig. 2. The eye movement heatmap and trajectory of the oil painting.

D. Model

To detect MDDs based on eye movement patterns, A multi-view invariant & specific model for eye movement data (MISEYE) implemented. Eye heatmaps, trajectories, and statistical vectors are core inputs of the model (Fig. 4). Extracted several eye movement features are higher-level information of the original eye movement data. They are all data from the same original eye movement signal, so they will contain common information but also have their own unique information. The model comprises three primary components. The initial segment employs a masked autoencoder, with the encoder condensing input data into a more concise representation. Pre-training of the model was executed on the Emod Dataset, encompassing eye movement data from diverse individuals observing images eliciting varied emotional responses [3]. Within this dataset, eye heatmaps underwent a 75% masking before processing by the ViT encoder, compelling it to extract salient features from the unmasked regions. Utilizing the default ViT base model, the decoder then reconstructed the original input from this encoded data. Following pre-training, the model underwent refinement using the training data and subsequent evaluation on a test set. The model's second segment harnesses a pre-trained Transformer model, processing eye-tracking time series data. During its pre-training phase, this component utilized proprietary multi-task eye-tracking data, employing a 20-second eye-tracking window to anticipate the ensuing 4-second sequence. For the depression detection task, the pre-trained Transformer encoder exclusively facilitated eye movement data extraction, enabling the derivation of high-dimensional spatial attributes. The final component amalgamates eye movement features vector delineated in Section III, which was not processed before fusion.

The fusion part employs the modality-invariant and specific representation structure [38]. Processed heatmap and trajectory data, along with the statistical vectors, traverse both a morphology-invariant public encoder and a morphology-specific private encoder, and respectively get their public and private representations:

$$h_m^c = E_c(u_m; \theta^c), h_m^p = E_c(u_m; \theta^p), \tag{2}$$

E_c shares the parameters θ^c across all features and E_p assigns separate parameters θ^p for each eye feature.

The three privately encoded hidden vectors and three publicly encoded hidden vectors are stacked into a matrix $M = [h_h^c, h_t^c, h_f^c, h_h^p, h_t^p, h_f^p] \in \mathbb{R}^{6*d_h}$. Then a Transformer encoder is used as an fusion network, which generates a new matrix $\bar{M} = [\bar{h_h^c}, \bar{h_t^c}, \bar{h_f^c}, \bar{h_h^p}, \bar{h_t^p}, \bar{h_f^p}]$. Finally, the Transformer outputs are concatenated to a single vector and sent to a linear classifier to get the classification result.

The model comprises four distinct loss components. The similarity loss quantifies the difference in shared representations across features. For this similarity loss, we used the cosine distance metric:

$$-dist(h_1, h_2) = -\frac{||h_1||_2||h_2||_2 - h_1 \cdot h_2}{||h_1||_2||h_2||_2} \tag{3}$$

$$\mathcal{L}_{sim} = \sum_{\substack{(m_1,m_2)\in \\ \{(h,t),(t,f),\\ (f,h)\}}} dist(h_{m_1}^c, h_{m_2}^c) \tag{4}$$

The difference loss comprises two components. The first part quantifies the distance between the public hidden representations of individual features and the collective private hidden vector. The second part measures the discrepancy in private hidden representations across modal pairs. For this difference loss, we employed the negative cosine distance metric:

$$\mathcal{L}_{diff} = \sum_{\substack{m\in\{h,\\ t,f\}}} -dist(h_m^c, h_m^p) + \sum_{\substack{(m_1,m_2)\in \\ \{(h,t),(t,f),\\ (f,h)\}}} -dist(h_{m_1}^p, h_{m_2}^p) \tag{5}$$

The reconstruction loss defines the loss between the reconstructed features of the decoder $\hat{u}_m = D([h_m^c, h_m^p], \theta^D)$ and the input feature of the encoder u_m. The reconstruction loss is the mean squared error between u_m and $\hat{u}_m$.

$$\mathcal{L}_{recon} = \sum_{m\in\{h,t,f\}} \frac{||u_m - \hat{u}_m||_2^2}{d_h} \tag{6}$$

where d_h denoted the dimension of the feature.

The task loss is the cross-entropy loss for classification result.

$$\mathcal{L}_{task} = -\frac{1}{N}\sum_{i=0}^{N} y_i \cdot log(\hat{y}_i) \tag{7}$$

The overall learning of the model is performed by minimizing the following loss:

$$\mathcal{L} = \mathcal{L}_{task} + \mathcal{L}_{sim} + \mathcal{L}_{diff} + \mathcal{L}_{recon} \tag{8}$$

E. Evaluation Details

The data was divided into five folds based on subject categories, ensuring an equitable representation in each population category. Each fold was used as a test set, while the remaining four folds constituted the training set. Due to the unique crowd feature patterns observed in each painting, a separate model was deployed for every painting. Consequently, each test set

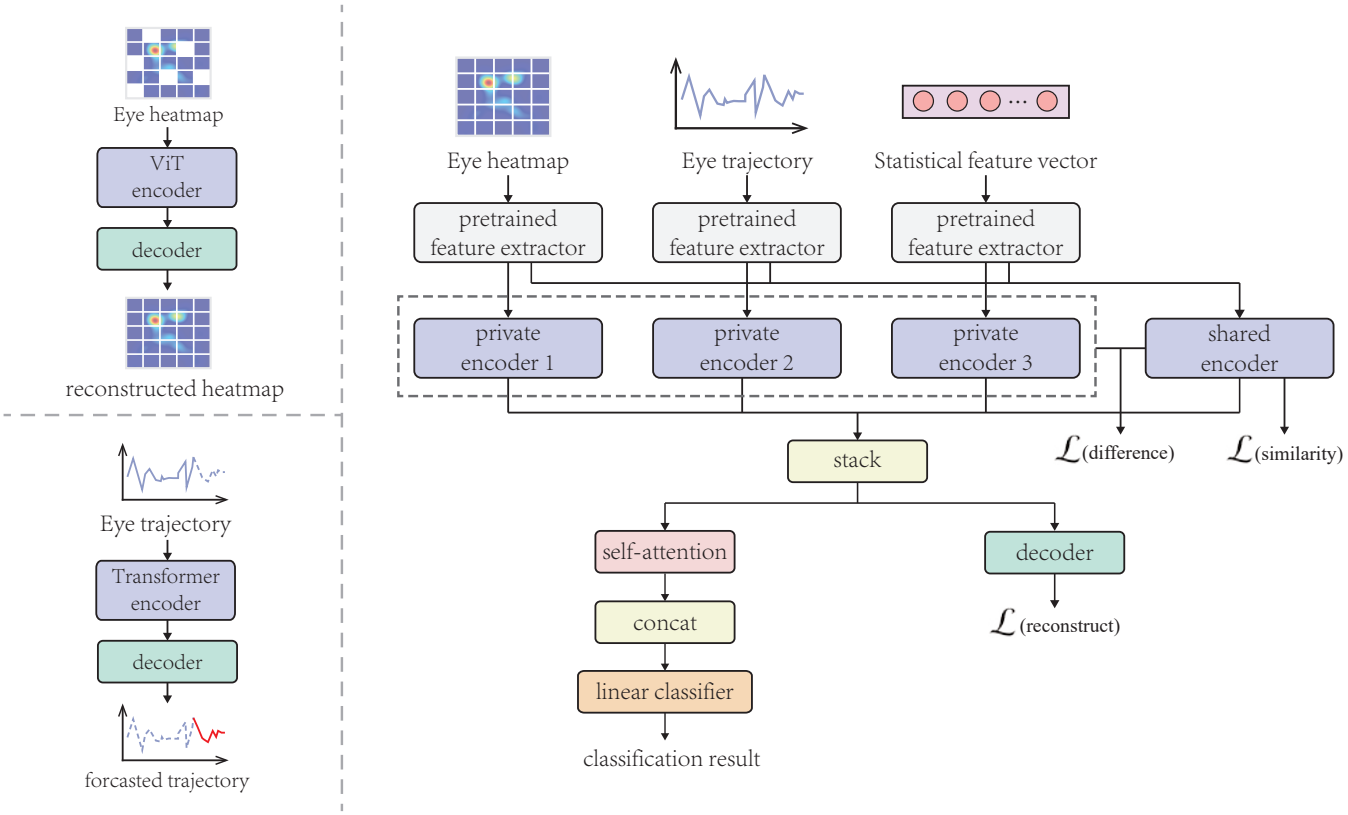

Fig. 3. The proposed fusion model structure. The left showed the pre-training process of eye heatmaps and trajectories. The right showed the MIS-EYE model of eye movement data.

encompassed 40 models, yielding 40 sets of results. For the prediction results of the test set, we use the mean accuracy rate and the voting accuracy rate as two evaluation criteria:

- 1. The mean accuracy of all 40 oil paintings across all subjects obtained by the subjects in the test set.
- 2. The voting accuracy entailed each test set subject receiving 40 predicted classification labels from all oil painting models, with the subject being categorized based on the label that occurs most frequently.

The final result was computed with the average on the five folds. Since the use of voting accuracy will sharply reduce the number of test samples, the fluctuation will also be high. Although the accuracy rate will be improved after voting, we take mean accuracy as the standard to measure the model, and the voting accuracy is only used as a reference.

V. Results

Primarily, we conducted a comparative analysis of three distinct eye movement features. We fine-tuned the heatmaps by utilizing a pre-trained ViT encoder to derive classification results. Similarly, a pre-trained Transformer encoder was employed to refine the trajectory for classification purposes. On the other aspect, the classification of statistical vectors was achieved using a single-layer linear classifier. Figure 4 presents the confusion matrix for the independent predictions of the three features. Our findings indicate that eye movement heatmaps, trajectories, and statistical vectors data effectively distinguish MDDs and healthy controls.

Table II presents the pairwise combinations of the three features alongside the classification results for depressed and healthy controls. Notably, the fusion model, which integrates any two features, and demonstrates higher accuracy compared to models trained on individual features. When all three featuress are combined, the resultant fusion model achieves the highest performance. Comparing the independent result of three features, the heatmap got a low mean accuracy but a relatively high voting accuracy. Since the voting accuracy is ascertained by averaging the outcomes from 40 predictions, this indicates that using heatmap only has a low decision confidence. The mean accuracy results of trajectory and statistical vector is higher, but meanwhile, they suffer from a lower voting accuracy. Ultimately, the combined results from the three features excel in both basic and voting accuracy, underscoring the model's robust predictive confidence.

We evaluated the MIS-EYE model against conventional fusion methods as detailed in Table III. Among those methods, the mean, max, and fuzzy fusion are the fusion methods of decision level. Interestingly, both mean and max fusion methods yielded lower base accuracy than the trajectory and feature single modes. In contrast, the fuzzy fusion approach outperformed any standalone feature in terms of accuracy. The bimodal deep auto-encoder (BDAE), multimodal transformer

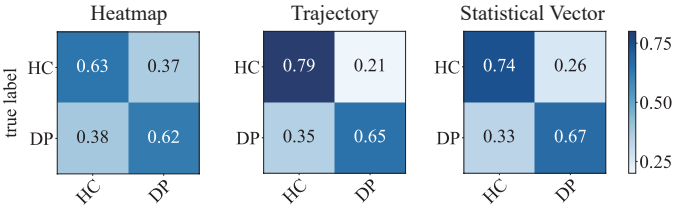

Fig. 4. Confusion matrices of different features.

TABLE II. Accuracy and standard deviations of standalone feature only.

Feature	Mean Acc.	Std.	Voting Acc.	Std.
Heatmap	60.81	1.76	83.75	3.43
Trajectory	71.81	3.43	75.00	**3.06**
Feature	70.41	4.61	81.25	5.59
Heatmap & Trajectory	76.81	**0.85**	85.00	7.50
Heatmap & Feature	76.19	1.86	80.00	9.19
Trajectory & Feature	77.56	3.38	83.75	10.16
All	**79.88**	2.62	**88.75**	6.37

and MIS-EYE used deep model fusion strategies, but they differ in their architectures. Specifically, the bimodal deep auto-encoder utilizes only the private encoder component depicted in Fig. 3, omitting the Transformer fusion network presented in the shared encoder and final fusion layer, as shown in Fig. 3. The multimodal transformer directly fuses the multiview eye features, without encoding them into public and private representations. In the MIS-EYE model, we integrated the pre-trained encoders for both the eye heatmaps and trajectories, along with the statistical vectors, into the transformer encoder. This amalgamation enabled concurrent training of the two networks, culminating in the final classification results. Ultimately, among the diverse fusion methodologies assessed, the MIS-EYE model consistently demonstrated superior performance in both basic and voting accuracy.

In our MIS-eye model, an ablation study is presented as supplementary material in Table IV. This study systematically evaluated the parameters of the private encoders and the public encoder, resulting in configurations with exclusively public or private encoders. The accuracies of these isolated models were diminished compared to the integrated model. The superior ac-

TABLE III. Accuracy and standard deviations of different models.

Method	Mean Acc.	Std.	Voting Acc.	Std.
Mean	66.81	4.28	78.75	8.83
Max	66.13	3.59	77.50	7.29
Fuzzy fusion	75.97	3.38	85.00	9.35
BDAE [39]	77.81	**1.75**	87.50	**3.95**
Multimodal transformer [40]	78.94	3.01	85.00	8.48
MIS-EYE	**79.88**	2.62	**88.75**	6.37

TABLE IV. Ablation study of the MIS-EYE model.

Method	Mean Acc.	Std.	Voting Acc.	Std.
Public only	76.72	3.50	82.50	4.68
Private only	77.90	1.90	82.50	10.46

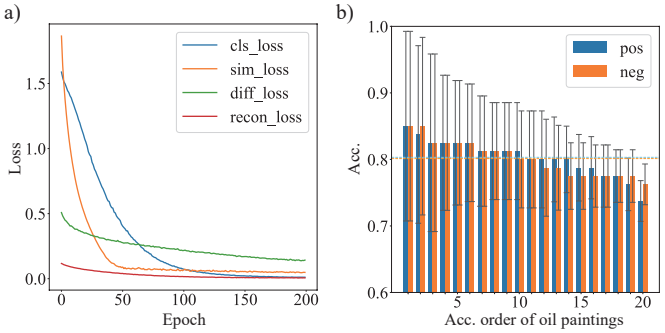

Fig. 5. a) Decrease of different losses in the training process. b) Acc./Std. % of each individual oil paintings in descending order.

curacy of the private-only configuration, relative to the public-only one, is attributable to the presence of a shared encoder complemented by three private encoders, which collectively retain more comprehensive information. Additionally, we list the decline of the four losses in Eq. (8) during training (Fig. 5 a). This illustrated the effectiveness of several losses in model training. We delineated the accuracy for each of the 40 oil paintings individually, as illustrated in Fig. 5 b. The 20 positive oil paintings and the 20 negative ones were ranked based on discrimination accuracy. In the figure, distinct dashed lines of two colors each represent sets of 20 paintings. Notably, the accuracy did not exhibit a significant disparity between the positive and negative oil paintings. Moreover, irrespective of emotion type, paintings with elevated accuracy also manifested higher standard deviations across the five folds.

VI. Conclusion

Differences in eye movement behaviors exist between depressed patients and healthy people in emotional cognition tasks. In this study, we have introduced a novel paradigm that utilizes oil paintings to detect differences in eye movement patterns between MDD patients and healthy controls. We extracted heatmaps, trajectories, and statistical vectors from the original eye movement data, and established an MIS-EYE model for depression detection. Within the model's architecture, the shared and distinct components of eye movement data from various features are effectively integrated. The proposed model attained an accuracy of 79.88%, demonstrating notable efficacy in differentiating MDD patients from healthy controls.

References

[1] N. Sadek and J. Bona, "Subsyndromal symptomatic depression: a new concept," Depression and Anxiety, vol. 12, no. 1, pp. 30–39, 2000.
[2] W. Gao, X. Yan, and J. Yuan, "Neural correlations between cognitive deficits and emotion regulation strategies: understanding emotion dysregulation in depression from the perspective of cognitive control and cognitive biases," Psychoradiology, vol. 2, no. 3, pp. 86–99, 2022.
[3] World Health Organization, "World mental health report: transforming mental health for all," 2022.
[4] S. Olbrich and M. Arns, "EEG biomarkers in major depressive disorder: discriminative power and prediction of treatment response," International Review of Psychiatry, vol. 25, no. 5, pp. 604–618, 2013.

[5] S. Grimm et al., "Imbalance between left and right dorsolateral prefrontal cortex in major depression is linked to negative emotional judgment: an fMRI study in severe major depressive disorder," Biological Psychiatry, vol. 63, no. 4, pp. 369–376, 2008.
[6] S. F. Husain et al., "Validating a functional near-infrared spectroscopy diagnostic paradigm for Major Depressive Disorder," Scientific Reports, vol. 10, no. 1, p. 9740, 2020.
[7] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in Proceedings of the International AAAI Conference on Web and Social Media, 2013, vol. 7, no. 1, pp. 128–137.
[8] M. Cannizzaro, B. Harel, N. Reilly, P. Chappell, and P. J. Snyder, "Voice acoustical measurement of the severity of major depression," Brain and Cognition, vol. 56, no. 1, pp. 30–35, 2004.
[9] Y. Lu, W.-L. Zheng, B. Li, and B.-L. Lu, "Combining eye movements and EEG to enhance emotion recognition.," in IJCAI, 2015, vol. 15, pp. 1170–1176.
[10] K. Morita, K. Miura, K. Kasai, and R. Hashimoto, "Eye movement characteristics in schizophrenia: A recent update with clinical implications," Neuropsychopharmacology Reports, vol. 40, no. 1, pp. 2–9, 2020.
[11] F. Chan, I. T. Armstrong, G. Pari, R. J. Riopelle, and D. P. Munoz, "Deficits in saccadic eye-movement control in Parkinson's disease," Neuropsychologia, vol. 43, no. 5, pp. 784–796, 2005.
[12] M. Gorges et al., "The association between alterations of eye movement control and cerebral intrinsic functional connectivity in Parkinson's disease," Brain Imaging and Behavior, vol. 10, pp. 79–91, 2016.
[13] U. P. Mosimann, R. M. Müri, D. J. Burn, J. Felblinger, J. T. O'Brien, and I. G. McKeith, "Saccadic eye movement changes in Parkinson's disease dementia and dementia with Lewy bodies," Brain, vol. 128, no. 6, pp. 1267–1276, 2005.
[14] X. Liu et al., "Spatial and temporal abnormalities of spontaneous fixational saccades and their correlates with positive and cognitive symptoms in schizophrenia," Schizophrenia Bulletin, p. sbad039, 2023.
[15] D. P. Munoz and S. Everling, "Look away: the anti-saccade task and the voluntary control of eye movement," Nature Reviews Neuroscience, vol. 5, no. 3, pp. 218–228, 2004.
[16] P. Van Donkelaar and A. S. Drew, "The allocation of attention during smooth pursuit eye movements," Progress in Brain Research, vol. 140, pp. 267–277, 2002.
[17] B. Pfleging, D. K. Fekety, A. Schmidt, and A. L. Kun, "A model relating pupil diameter to mental workload and lighting conditions," in Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 2016, pp. 5776–5788.
[18] J. Takahashi et al., "Eye movement abnormalities in major depressive disorder," Frontiers in Psychiatry, vol. 12, p. 673443, 2021.
[19] T. Kosch, M. Hassib, P. W. Woźniak, D. Buschek, and F. Alt, "Your eyes tell: Leveraging smooth pursuit for assessing cognitive workload," in Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 2018, pp. 1–13.
[20] A. Chatterjee, P. Widick, R. Sternschein, W. B. Smith, and B. Bromberger, "The assessment of art attributes," Empirical Studies of the Arts, vol. 28, no. 2, pp. 207–222, 2010.
[21] S. Alghowinem, R. Goecke, M. Wagner, G. Parker, and M. Breakspear, "Eye movement analysis for depression detection," in 2013 IEEE International Conference on Image Processing, 2013, pp. 4220–4224.
[22] Y. Wang et al., "The similar eye movement dysfunction between major depressive disorder, bipolar depression and bipolar mania," The World Journal of Biological Psychiatry, vol. 23, no. 9, pp. 689–702, 2022.
[23] A. Vaswani et al., "Attention is all you need," Advances in Neural Information Processing Systems, vol. 30, 2017.
[24] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.
[25] J. Chen et al., "Transunet: Transformers make strong encoders for medical image segmentation," arXiv preprint arXiv:2102.04306, 2021.
[26] F. Shamshad et al., "Transformers in medical imaging: A survey," Medical Image Analysis, p. 102802, 2023.
[27] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16000–16009.
[28] J. Xiao, Y. Bai, A. Yuille, and Z. Zhou, "Delving into masked autoencoders for multi-label thorax disease classification," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 3588–3600.
[29] S. Luo, Y.-T. Lan, D. Peng, Z. Li, W.-L. Zheng, and B.-L. Lu, "Multimodal emotion recognition in response to oil paintings," in 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2022, pp. 4167–4170.
[30] H. Ashraf, M. H. Sodergren, N. Merali, G. P. Mylonas, H. Singh, and A. Darzi, "Eye-tracking technology in medical education: A systematic review," Medical Teacher, vol. 40, no. 1, pp. 62–69, Nov. 2017.
[31] C. Bodkin and C. H. Schenck, "Rapid Eye Movement Sleep Behavior Disorder in Women: Relevance to general and Specialty Medical practice," Journal of Womens Health, vol. 18, no. 12, pp. 1955–1963, Dec. 2009.
[32] M. J. Weightman, T. M. Air, and B. T. Baune, "A review of the role of social cognition in major depressive disorder," Frontiers in Psychiatry, vol. 5, p. 179, 2014.
[33] R. Engbert and K. Mergenthaler, "Microsaccades are triggered by low retinal image slip," in Proceedings of the National Academy of Sciences, vol. 103, no. 18, pp. 7192–7197, 2006.
[34] J. Otero-Millan, A. Serra, R. J. Leigh, X. G. Troncoso, S. L. Macknik, and S. Martinez-Conde, "Distinctive features of saccadic intrusions and microsaccades in progressive supranuclear palsy," Journal of Neuroscience, vol. 31, no. 12, pp. 4379–4387, 2011.
[35] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing AI created fake videos by detecting eye blinking," in 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 2018, pp. 1–7.
[36] H. Bouma and L. Baghuis, "Hippus of the pupil: periods of slow oscillations of unknown origin," Vision Research, vol. 11, no. 11, pp. 1345–1351, 1971.
[37] S. Fan et al., "Emotional attention: A study of image sentiment and visual attention," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7521–7531.
[38] D. Hazarika, R. Zimmermann, S. Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in Proceedings of the 28th ACM international conference on multimedia, 2020, pp. 1122–1131.
[39] W. Liu, W. Zheng, B. Lu, "Emotion recognition using multimodal deep learning," in Neural Information Processing: 23rd International Conference, ICONIP 2016, Kyoto, Japan, October 16–21, 2016, Proceedings, Part II 23, 2016, pp. 521–529.
[40] Tsai, Y., et al, "Multimodal transformer for unaligned multimodal language sequences," in Proceedings of the conference. Association for Computational Linguistics. Meeting, 2019, pp. 6558.