Three-Stream Convolutional Neural Network for Depression Detection With Ocular Imaging

Minqiang Yang, *Member, IEEE*, Ziru Weng, Yuhong Zhang, Yongfeng Tao, *Graduate Student Member, IEEE*, and Bin Hu, *Fellow, IEEE*

***Abstract*— Depression is a prevalent and severe mental disorder that significantly affects both mind and body, leading to persistent feelings of sadness, despair, and impaired functionality. Diagnosis of depression primarily relies on clinical assessment and observation of symptoms. However, due to the lack of objective indicators, the experience and skills of doctor may lead to misdiagnosis. Current researches indicate that eye movement patterns and pupil dilation can serve as potential biomarkers for emotional and cognitive dysregulation in individuals with depression. However, most studies are based on manually extracted eye movement features, overlooking a significant portion of information available in ocular imaging. This paper proposes Three-Stream Convolutional Neural Network (TSCNN) for detecting depression, leveraging both spatio-temporal information of raw ocular imaging and paradigmatic semantic features. We suggest using optical flow with different sampling intervals to capture temporal features. In the third stream, we employ an encoder to learn semantic information from paradigm images and use it as prior knowledge. Finally, we utilize a fully connected network for classification, achieving an accuracy of 79.3% on our self-collected dataset. The proposed method may demonstrate significant clinical utility in the future.**

***Index Terms*— Depression detection, eye movement, ocular imaging, three-stream convolutional neural network.**

I. Introduction

DEPRESSION is a widespread mental disorder that profoundly affects both mental and physical health. The World Health Organization (WHO) considers it a primary cause of disability [1]. Depression patients may experience a range of symptoms, including a sense of despair, loss of interest in previously enjoyed activities, sleep disturbances, changes in weight or appetite, fatigue, difficulty concentrating, and indecisiveness. These symptoms can significantly impact an individual's daily functioning, potentially leading to functional impairments in work, study, or social relationships. The most widely used diagnostic criteria for depression are found in Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [2], it provides a standardized set of symptoms and criteria, clinical professionals determine whether an individual meets the characteristics of depression through psychological assessments [3] and clinical interviews [4]. These traditional diagnostic methods rely heavily on subjective perception of patient and expertise of physician, emphasizing the urgent need for objective diagnostic tools. Disease surveillance studies based on physiological and behavioral data have been extensively conducted [5], providing many clues for objective auxiliary diagnosis of depression [6], [7], such as EEG [8], [9], [10], facial expressions [11], [12], [13], and audio [14], [15] etc.

It has been shown that gaze duration, sweep range, blinks, and pupil size in response to emotional stimuli are all key features of emotion recognition [16], [17], [18], pupil dilation is linked to activation of autonomic nervous system, particularly sympathetic branch, contains rich and genuine emotional information. Ocular movements and pupil changes objectively reflect subjects' attention to different stimuli [19], eye-tracking technology [17], [20], [21] visualizes the distribution of attention and are effective tools for analyzing attentional biases in depression patients [18], [22], [23], [24]. Abnormal ocular changes have been observed in depression patients, such as prolonged fixation duration, decreased frequency of eye slews, and reduced smoothness eye movements [25]. These changes suggest that attention processes, cognitive functions, and emotional regulation abilities are generally impaired in individuals with depression, and they may exhibit atypical pupil dilation responses to emotional stimuli [26].

Most of existing studies detect depression by defined ocular change features such as gaze points, pupil diameter, number of eye jumps, gaze duration [27], [28]. Alghowinem et al. [29] labeled each eye image with 74 points and extracted three features for each image. Al-Gawwam and Benaissa [30] extracted features of blink frequency and duration for subjects in each video frame. Shen et al. [31] proposed a

Manuscript received 17 September 2023; revised 21 November 2023; accepted 26 November 2023. Date of publication 5 December 2023; date of current version 18 December 2023. This work was supported in part by the Natural Science Foundation of Gansu Province,China (Grant No. 22JR5RA488), in part by the Fundamental Research Funds for the Central Universities (Grant No. lzujbky-2023-16), in part by the National Key Research and Development Program of China (Grant No. 2019YFA0706200), in part by the National Natural Science Foundation of China (Grant No.62227807, No.62202212), in part by the SIT2030-Major Projects (2021ZD0200800). Supported by Supercomputing Center of Lanzhou University. *(Corresponding author: Bin Hu.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee of Guangyuan Mental Health Center under Application No. GJWLSP2020012, and performed in line with the Declaration of Helsinki.

Minqiang Yang, Ziru Weng, Yuhong Zhang, and Yongfeng Tao are with the School of information Science and Engineering, Lanzhou University, Lanzhou 730000, China (e-mail: yangmq@lzu.edu.cn).

Bin Hu is with the Gansu Provincial Key Laboratory of Wearable Computing, School of Information Science and Engineering, Lanzhou University, Lanzhou, Gansu 730000, China, and also with the School of Medical Technology, Beijing Institute of Technology, Beijing 100081, China (e-mail: bh@lzu.edu.cn).

Digital Object Identifier 10.1109/TNSRE.2023.3339518

new experimental paradigm of eye movement-based cognitive psychology, extracting eye movement features during free viewing phase and attention frame tracking phase. Acquisition of eye movement events is labor-intensive, and follow-up tasks are limited to identified eye movement behaviors. Results may not be satisfactory due to small correlation between some eye movement events and detection tasks.

Depth methods can focus on undefined physiological behavioral features. With deepening of problem complexity and development of artificial intelligence techniques, neural networks have become stable and widely used. Many researchers are now using deep learning methods to process physiological signal data for emotion detection. Ma et al. [32] utilized a combination of convolutional neural networks and Long Short-Term Memory to capture depression-related characteristics in audio, resulting in a more precise audio representation. Yang et al. [33] proposed a multi-modal fusion framework consisting of deep convolutional networks and deep neural networks to process audio, video, and text streams. Tao et al. [34] utilized the transformer encoder to extract semantic information from video log data to identify anomalous emotional states. Pan et al. [35] proposed Spatio-Temporal Attention Depression Recognition Network (STA-DRN), which improves quality of extracted features by capturing global and local spatio-temporal information. de Melo et al. [36] designed DMSN architecture in order to adapt the model to different facial behaviors, capable of exploring a variety of multi-scale spatio-temporal features. To solve the problem of ignoring some undefined eye movement features, our previous work [37] proposed a novel depression detection method. This method involves directly clustering eye movement data to bypass the identification of eye movement events and acquire regions of interest (ROI).

Although recent researches focus on using deep learning methods to detect depression in various modal datasets, such as facial images, audio, text, and other modalities, there are not many research methods for ocular imaging. The main methods in the field of video recognition are Two-Stream Convolutional Neural Networks [38], Still Image Feature Aggregation, 3D Convolutional Networks [39] and Transformer [40], convolutional neural networks are very good at processing static appearance information (object shape size color, scene information, etc.) rather than motion information, so they can't handle video very well. Since this case, Two-Stream Convolutional Neural Networks use another network (optical flow network) to extract features of good motion information with good results. Although Two-Stream Convolutional Neural Networks notices spatial and temporal information, its main use is for action recognition. Unlike ocular imaging, the background of video for action recognition changes more rapidly and characters move more, ocular imaging has single background and simpler picture, requiring more attention to temporal features.

For properties of ocular imaging, we propose an end-to-end depression classification model i.e., Three-Stream Convolutional Neural Network (TSCNN). Given individual differences, each participant reacts to stimuli at different speeds. Sampling optical flow at different time intervals allows for a more comprehensive understanding of changes in participants' eyes. Secondly, considering the correlation between stimulus images and ocular changes, we attempt to integrate the spatio-temporal features of ocular imaging with the semantic features of paradigm images to enhance the model's classification capabilities. The main contributions of this paper are as follows:

- We present Three-Stream Convolutional Neural Network (TSCNN) method for depression detection. The method is capable of analysing raw ocular imaging and is an end-to-end model for extracting eye movement features.
- We split time block into two input streams (fast frame rate and slow frame rate) to sample subjects' ocular changes at different time intervals, allowing for a more detailed extraction of temporal features from ocular imaging.
- Latent codes of paradigm images are incorporated into the model as prior information and merged with the features from ocular imaging to aid in the learning process of the model.

The remainder of this paper is organized as follows. Section II provides an overview of different approaches used in eye-movement research and video recognition. Section III presents information on paradigm experiments and subjects involved. Section IV details structure of TSCNN. Section V shows experimental results. Section VI analyzes and discusses experimental results. Section VII summarizes work presented in this paper, identifies its shortcomings, and suggests potential avenues for future research.

II. Related Work

A. Eye Movements in Mental Disorders

The presence of mental disorders typically leads to a range of psychological issues, accompanied by negative emotions such as anxiety, fear, and depression. This condition leads to neurological dullness and eye disorientation, differences between depression patients and healthy control group can be identified through eye movement events. Eye movement events are interpretable and often have judgment criteria. Li et al. [41] confirmed abnormal eye movement metrics in depression patients compared to healthy controls via three eye movement task trials (gaze stabilization task, skip gaze task, and free viewing task). Representative features were selected for downstream tasks after acquiring eye-movement event features, such as, Li et al. [42] extracted 6 features to feed into classifier for depression detection, including Negative Preference (NP) frequency, average Fixation duration, Pupil size mean, and others. Then they conducted comparison experiments with five classifiers, including k-Nearest Neighbors (kNN), Logistic Regression, Support Vector Machine, Naive Bayes, and Random Forests (RF). On this basis, Pan et al. [43] mined eye movement data for individual attention bias features from different perspectives, including orientation, release, and transfer, confirming that individual attention bias features are indirectly revealed by reaction times. Despite eye movement events have interpretable advantages, the extraction methods are highly task-dependent. Different tasks may require different features, inevitably leading to repetitive work that is both labor-intensive and time-consuming. In addition, defined eye movement events are very limited, limiting to these features would ignore other useful undefined eye movement features.

Even many deep learning methods still rely on extracted eye movement events for depression detection. Zhao and Wang [44] used random forest to select 24 EM data features as input, EnSA model was designed to detect depression, where it consists of three modules: multi-head attention, SA, and Add. Kacur et al. [45] proposed several novel approaches (including neural network-based approaches) for automatic detection of schizophrenic patients, tested several features that enable analysis of visual tracker signals as a whole. The types of features spanned global (heat maps, gaze maps), feature sequences (mean, variance, and spectrum), static (x and y signals as 2D images), and dynamic (x and y signals as 2D images). Mao et al. [46] extracted pupil features, including position and size, obtained feature vectors of eye movements from normalized pupil information. Based on Long Short-Term Memory (LSTM) network, classifiers corresponding to each feature are built, all weak classifiers are combined to obtain a strong classifier for disease identification. None of these deep learning methods obtain features directly from raw image and still face the problem of extracting features manually.

B. Video Processing Methods

Features require enough experience to design, which is increasingly difficult with increasing amount of data. Thus end-to-end networks emerged, models learn features on their own without human intervention. There are three primary approaches in field of video processing. The first approach involves extracting features from each frame of the video. However, since the content of different frames is interrelated, directly inputting the extracted frames into the network makes it challenging to learn the associative information between them. Carreira and Zisserman [39] demonstrated experimentally that this strategy is not particularly effective. The second approach, 3D convolutional network, splits video into separate segments to train model. Xie et al. [47] improved a 3D-CNN for temporal feature extraction, using facial video recordings combined with Self-Rating Depression Scale (SDS) scores to detect depression. Although 3D convolutional neural networks have good performance to learn both spatial and temporal information, they do not converge easily and are prone to overfitting, the number of parameters is very large and computational complexity is higher than that of 2D convolutional networks, due to small size of the dataset, the training results are not satisfactory. The third network is Two-Stream Convolutional Neural Network, which is divided into two parts: spatial stream part takes a single image as input, temporal stream part takes the optical stream of multiple images as input, the two parts are subjected to late fusion after softmax.

Two-Stream Convolutional Neural Networks are mostly used for action recognition, as stated in Introduction section, ocular imaging is different from action recognition video, it does not have too many background changes, so the focus should be on processing of temporal stream. To address this issue, we split the temporal module of Two-Stream Convolutional Neural Networks into two pathways consisting of different frame frequencies (fast and slow) to capture subject's ocular changes over different time intervals. Melo et al. [48] enhanced original Two-Stream Convolutional Neural Network for depression detection. They developed a new preprocessing method for temporal part involving the extraction of complex semantic features using ResNet50 [49] as backbone. The depth of ResNet50 is capable of learning more complex features than simpler convolutional neural networks used in classic Two-Stream Convolutional Network.

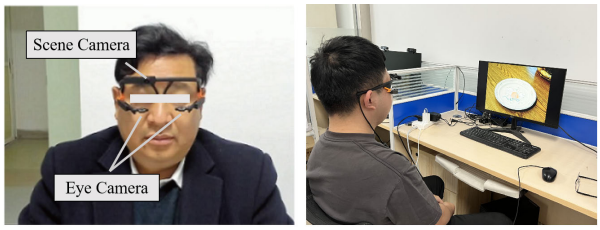

Fig. 1. The wearing standard illustration (a) and experimental real scene (b) of Pupil Core.

C. Latent Code

Latent code [50] is a low-dimensional representation of data, often used to represent important features or properties of data. In image processing, latent code is an abstract representation of image properties. Paradigm stimulus pictures triggered ocular changes, we add semantic information of paradigm pictures, combine it with spatio-temporal features of ocular imaging, model learns the correlation between ocular imaging features and paradigm pictures to provide more information to assist classification.

Based on the factors mentioned above, we chose Two-Stream Convolutional Neural Network as the foundational model. Considering the relationship TSCNN.

III. Materials and Experiment

A. Experimental Equipment

In this experiment, we use Pupil Core [51], [52], a wearable eye-tracking equipment made by Pupil Labs that is more versatile and convenient than desktop eye-tracking devices, since it does not require individuals to secure their heads in brackets. Pupil Core is equipped with three cameras that can capture video, i.e. two eye cameras record the subject's ocular imaging, and a scene camera captures environment in front of the glasses. The frame rate of ocular imaging captured by Pupil Core is approximately 200 fps, with a frame width of 320 and a frame height of 200.

Subfigure (a) of Fig. 1 illustrates Pupil Core and the proper way to wear it. The pupil camera should be located below the eyes so as not to cover them. The entire instrument must be symmetrical to subject's face.

B. Participants

This study was approved by Ethics Committee of Guangyuan Mental Health Center. We worked with two hospitals: the Second People's Hospital of Gansu Province and the Third People's Hospital of Guangyuan, Sichuan province. We recruited 81 subjects (Aged 18-55), including 41 patients with depression (13 males and 28 females) and 40 healthy controls (9 males and 31 females). Among the participants we recruited, there was a higher proportion of females. Some

TABLE I
DEMOGRAPHIC INFORMATION OF THE SUBJECTS

Category ofber	Number	Age(MEAN±SD)	Gender
Depression	13	37.69 ± 9.12	Male
Depression	28	35.64 ± 10.21	Female
Healthy	9	42.40 ± 8.21	Male
Healthy	31	37.91 ± 10.21	Female

TABLE II
T-TEST FOR DEMOGRAPHIC INFORMATION OF SUBJECTS. (P>0.05:NO SIGNIFICANT DIFFERENCE)

Factor	Lenvene	T-test	P-value
Age	0.996	0.648	0.519
Education level	0.099	1.279	0.205
Residence	0.044	1.820	0.079
Marital status	0.353	0.513	0.613

studies [53], [54] suggest that females often possess higher emotional sensitivity, are more prone to depression, and consequently exhibit more distinct behavioral patterns. All subjects read and signed written informed consent before experiment, all subjects had at least primary education level, all subjects had not been on psychotropic medication in the last two weeks. Patients with depression participated in experiment accepted Mini-International Neuropsychiatric Interview (M.I.N.I.), and the diagnosis was confirmed according to DSM-IV. The details of subjects are shown in Table I.

Depression may be influenced by factors such as age, education level and residence [55]. We conducted an independent sample t-test on four aspects of age, education level, residence, and marital status for subjects from Gansu and Sichuan provinces. From the results in Table II, it can be observed that the regional differences are not significant.

C. Paradigm Experiment

Paradigm experiments were conducted in a quiet and neat environment, required experimental equipment included a computer monitor and an eye-tracking device. Subjects were placed at a distance of 60-70 cm from computer screen and adjusted to the final sitting position under guidance, who guided subjects to look forward and normalized pupil camera. The experimental scenario diagram is shown in subfigure (b) of Fig. 1.

Eye-tracking experiment used classical free-viewing paradigm [56], stimulus images were derived from International Mood Picture System (IAPS) [57], which is considered to be the most reliable and valid system for experimental research on emotions. Experiment was divided into two blocks with a total of 40 stimulus pictures, including 20 neutral pictures, 10 positive pictures, and 10 negative pictures, as a way to induce eye-movement responses of subjects. The arrangement of paradigm pictures in one block is shown in Fig. 2, each picture is shown for 5 seconds.

IV. METHODS

A. Three-Stream Convolutional Neural Network

Compared with other networks such as 3D convolution neural network [58] and static image feature aggregation,

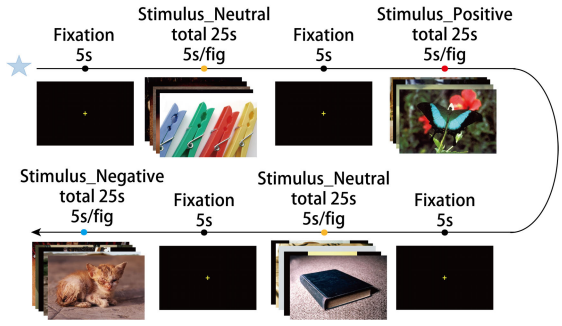

Fig. 2. One block of free exploration paradigm.

Two-Stream Convolutional Neural Networks add an optical stream path to learn temporal information of video, it integrates spatial and temporal information in visual processing tasks, which address the limitations of traditional convolutional neural networks in capturing both appearance and motion information effectively [38].

Motivated by the efficiency of Two-Stream Convolutional Neural Networks, we try to add the third stream, which contains prior information of paradigm stimulus, propose TSCNN, the model architecture is shown in Fig. 3. TSCNN consisted of three main modules: spatial module, temporal module, and paradigm module. Spatial module extracts spatial features from input frames of ocular imaging. Temporal module extracts temporal features of ocular imaging from inputs optical flow images. Ocular imaging is a time-varying sequence, each subject will have different eye movement responses to different paradigm stimulus pictures, so we focus on the processing of extracting temporal features of ocular imaging. To achieve this, we adopt the design concept of SlowFast network, where optical streams with different frame rates are input and the model learns pupil change features between different frame rates (i.e. different time periods).

B. Preprocessing

In this experiment, to improve universality and sample diversity of the model, we extract ocular imaging frame by frame. For each adjacent pair of extracted frames, estimate a dense optical flow using calcOpticalFlowFarneback module in OpenCV [59]. The first 8000 frames and optical streams of ocular imaging for each subject were chosen, separated into five equal groups of 1600 frames or optical streams each, and then disordered into smaller five samples. Paradigm video processing is congruent with ocular imaging processing. In order to increase the efficiency of training, we pre-generate frames and optical flow images before training and then import them during training.

With this processing, the inputs to the model (RGB frames, optical flow, and paradigm) are temporally aligned. In this paper, preprocessing does not require complex manual operations such as designing features and extracting eye-movement events, only requires data enhancement of raw ocular imaging to be fed into the model.

C. Spatial Module

Ocular imaging has only one eye throughout, the background information is relatively homogeneous, learning spatial

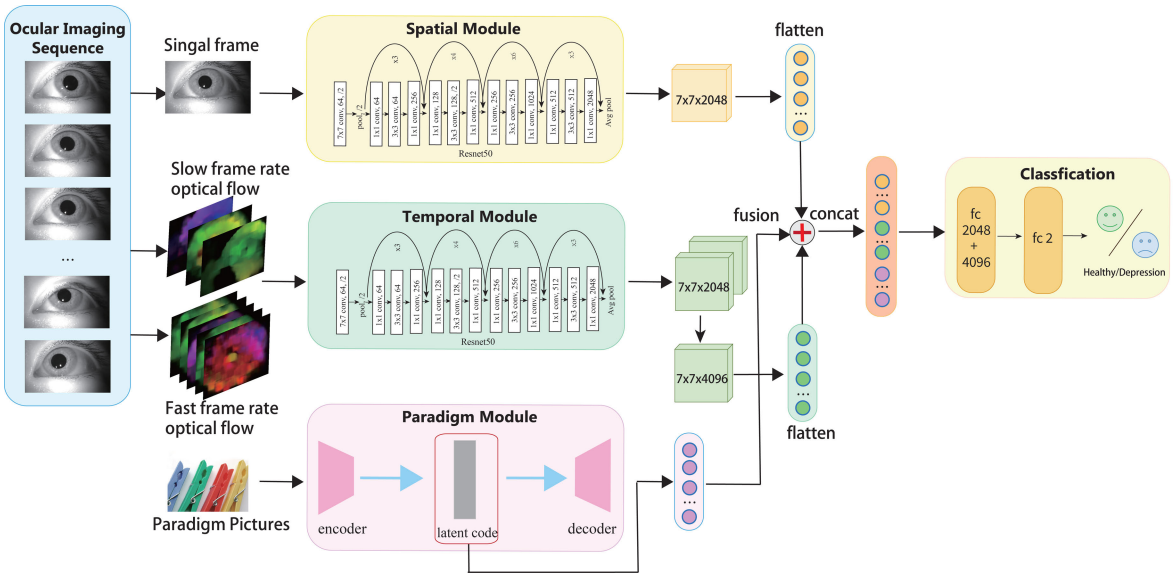

Fig. 3. Depression detection structure of Three-Stream Convolutional Neural Network. First, video frames are extracted, optical flows are computed, and they are preprocessed. second, spatio-temporal features in data stream are extracted, paradigm images are encoded to obtain latent codes. Then these feature maps are fused, finally depression detection is performed.

information using a single frame is sufficient. The depth of ResNet efficiently extracts abstract features, its unique residual module enhances responsiveness to output changes. This solves the problem of deep networks experiencing performance degradation with increasing depth. Based on these advantages, we use ResNet50 as backbone of spatial module.

The input of ResNet50 is RGB image which has 3 channels, and the $kernel\_size$ of the first layer of the network is 7, $stride$ is 2, and $padding$ is 3.

Assume that feature map output by spatial convolutional network is $f_s^n \in \mathbb{R}^{H\times W\times C}$, the dimension of $H \times W \times C$ is $7 \times 7 \times 2048$.

D. Temporal Module With Different Time Intervals

Eye-movement reaction time to emotional stimuli will be different for each subject. SlowFast network [60] links slow and fast pathways laterally, drawing on this idea, we propose temporal module containing optical flow at different time intervals, design a slow frame rate optical flow pathway and a fast frame rate optical flow pathway, to learn features of subjects' ocular changes at different time intervals. Assuming that fast frame rate optical flow pathway is sampled after t time, slow frame rate optical flow pathway is sampled after αt time, where $\alpha > 1$, that means fast frame rate optical flow is α times denser than slow frame rate optical flow, in this experiment, the representative value of α is 8. We still chose ResNet50 as backbone.

Unlike Spatial Module, the number of channels in the first layer input is not 3, instead, stack 10 optical flow images based on frame rate as an input. Assume that feature map output by temporal convolutional network is $f_{slow_t}^n \in \mathbb{R}^{H_{slow}\times W_{slow}\times C_{slow}}$ and $f_{fast_t}^n \in \mathbb{R}^{H_{fast}\times W_{fast}\times C_{fast}}$, the dimension of both $H_{slow} \times W_{slow} \times C_{slow}$ and $H_{fast} \times W_{fast} \times C_{fast}$ are $7 \times 7 \times 2048$, then concat them, allows the model to learn optical flow information at different frame rates, which is defined as follows:

$$y_{temporal} = Concat[f_{slow_t}^n, f_{fast_t}^n] \tag{1}$$

the dimension of output feature map is $7 \times 7 \times 4096$.

E. Paradigm Module

For depression detection, paradigm stimulus are the triggers that evoke subjects' emotions. Since RGB frame, optical flow of ocular imaging and paradigm picture are temporally aligned, adding semantic features of paradigm picture as a priori knowledge has two considerations, one is to learn the relationship between paradigm picture and features of ocular imaging, the other is to learn different features of ocular changes for the same paradigm picture. Assuming $\overline{x}$ is the low-dimensional vector [61] obtained by encoding the input vector x. Loss between x and $\overline{x}$ is compared, and then convolutional network is trained to reduce the loss gradually, thus achieving unsupervised learning. In this experiment, the training set is images extracted from emotionally stimulated videos used in paradigm experiment. As shown in Fig. 4, paradigm images are fed into network with three convolutional layers and two fully connected layers for encoding, removing redundant information, extracting semantic feature, preserving the important information to produce latent code, then decoding it by network with two fully connected layers and three convolutional layers to produce reconstructed image, which is compared with paradigm image, and optimized for latent code using Mean Square Error loss (MSEloss). In this network structure, we used ReLu activation function.

The formula of Mean Square Error loss is as follows:

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - f(x_i))^2 \tag{2}$$

where y_i denotes truth value and $f(x_i)$ denotes predictive value, m is the number of samples.

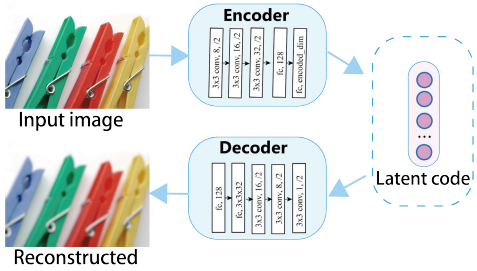

Fig. 4. Paradigm module latent code framework diagram.

To accelerate model convergence, we set up a BatchNorm layer, which normalizes the data output from each layer to the same distribution with the following equation:

$$y_{paradigm'} = \frac{x - E[x]}{\sqrt{Var[x] + \varepsilon}} * \gamma + \beta \tag{3}$$

where $E[x]$ and $Var[x]$ are mean and variance of the batch data, ε is variable added to prevent zero in the denominator, γ and β linear transformations of inputs.

In this study, the encoder compresses input paradigm images into latent space for representation. After training, paradigm images of associated video stream clip is only characterised using encoder output latent code.

F. Feature Fusion

Feature fusion methods combine multiple features and utilize their synergistic effects to produce more reliable and accurate recognition results. According to Huang et al. [62], feature fusion approaches fall into two broad categories: strong and weak fusion. There are four subcategories of strong fusion, including early (pre) fusion, deep (feature) fusion, late (post) fusion, and asymmetric fusion. The placement of fusion has a considerable impact on classification accuracy, according to the evaluation made by Guo et al. [63], spanning from traditional models to recently developed methodologies. On the basis of this, we selected the placement for feature fusion.

For each set n, the inputs of three modules are denoted as p, q, l, three feature maps X_s^n, X_t^n, X_p^n are obtained, and they respectively correspond to the outputs of spatial module, temporal module and paradigm module.

$$\begin{cases} X_s^n = ResNet(p), \\ X_t^n = ResNet(q), \\ X_p^n = CNN(l). \end{cases} \tag{4}$$

We fuse X_s^n, X_t^n, X_p^n into an output y^n. It should be noted that, X_s^n, X_t^n, $X_p^n \in \mathbb{R}^{H \times W \times C}$, $y^n \in \mathbb{R}^{H' \times W' \times C'}$. H, W, C indicate height, width, and number of channels respectively.

Simonyan and Zisserman [38] adopted late fusion method for Two-Stream Convolutional Neural Network, weighted average of the outputs of two networks, which is perhaps a bit too simple. We picked deep fusion strategy for our studies because we think that late fusion will reduce the interactivity of each branch network's characteristics. Instead of fusing the output of fully connected layer, we fuse feature maps of spatial, temporal, and paradigm modules before fully connected layer. The equation for deep fusion method is as follows.

Deep fusion refers to the conversion of different modal data into low-dimensional feature representations before fusing them in the intermediate layer of the model.

$$\begin{cases} X_s^{n'} = Flatten(X_s^n), \\ X_t^{n'} = Flatten(X_t^n), \\ X_p^{n'} = Flatten(X_p^n). \end{cases} \tag{5}$$

$$y^n = Concat[X_s^{n'}, X_t^{n'}, X_p^{n'}] \tag{6}$$

where $X_s^{n'}$, $X_t^{n'}$, $X_p^{n'}$ are X_s^n, X_t^n, X_p^n flattened as a one-dimensional vector.

V. Experimental Results

In this section, we present the partitioning of dataset and experimental details, validate our proposed method in several ways.

A. Dataset and Experiment Setup

1) Dataset: The dataset contains 81 ocular imaging of subjects with approximately 5 minutes of each video. The dataset was divided into 64 training sets (33 depressed, 31 healthy), 8 validation sets (3 depressed, 5 healthy) and 9 test sets (5 depressed, 4 healthy). The video frames, optical flow maps and paradigm stimulus images were resized to 224 × 224 then subjected to data enhancement.

2) Experimental Details: For Temporal ConvNet, 10 optical flow maps sampled according to the frame rate are simultaneously stacked as input, while Spatial ConvNet is fed only the first frame of the current sample video stream. We extract all images of paradigm video, train them by convolutional self-encoder and compress the input images into hidden space for representation, then use latent code to characterize paradigm information, fuse them with the corresponding ocular imaging video stream segments after training is completed.

Our experiments are conducted in PyTorch framework [64]. We trained our model with initial learning rate $1e^{-3}$, using Adam optimizer [65] with momentum 0.9. To improve the generalization ability of the model and prevent overfitting, we use dropout strategy, in temporal module, we set dropout to 0.1. The Forward Pass calculation for TSCNN is 417.442G and the number of parameters is 89.877M.

In the training process, we choose cross-entropy loss function as the loss function, which is defined as follows:

$$L = \frac{1}{N} L_i = \frac{1}{N} \sum_{i=1} -[y_i \cdot log(p_i) + (1 - y_i) \cdot log(1 - p_i)] \tag{7}$$

where y_i denotes labels of ocular imaging, positive class is 1 (health), negative class is 0 (depression), p_i denotes the probability that ocular $imaging_i$ is predicted to be a positive class.

We trained and tested the final network on an A100 GPU, TSCNN took about 3 hours to train on a data-augmented dataset, and had a faster convergence rate compared to the same type of input optical streaming model. Algorithm 1 shows our train process.

TABLE III

EXPERIMENTAL RESULTS FOR DIFFERENT NETWORKS. THE PERFORMANCE OF DIFFERENT SUBSTRUCTURES IS DESCRIBED BY MEAN ± STD. T-TEST WAS USED FOR SIGNIFICANCE ANALYSIS, IN WHICH TSCNN IN THE FINAL GROUP WHICH CONTAINING ALL KIND OF INPUT WAS THE CONTROL GROUP, ** : EXTREMELY SIGNIFICANT DIFFERENCE ($P < 0.01$), * : SIGNIFICANT DIFFERENCE ($P < 0.05$)

Data Modalities	Method	Accuracy	Precision	F1 score	Recall	Specificity
One-eye single frame data	C3D	0.626 ± 0.07**	0.588 ± 0.07**	0.662 ± 0.07**	0.716 ± 0.20*	0.752 ± 0.16**
	R3D	0.582 ± 0.08**	0.555 ± 0.09**	0.552 ± 0.09**	0.551 ± 0.10**	0.608 ± 0.08**
	R2Plus1D	0.547 ± 0.05**	0.520 ± 0.05**	0.550 ± 0.11**	0.657 ± 0.27**	0.650 ± 0.18**
	SlowFast	0.762 ± 0.03*	0.777 ± 0.09*	0.751 ± 0.05	**0.726 ± 0.02**	0.809 ± 0.02*
Single frame+Optical flow	Two-Stream	0.722 ± 0.02**	0.753 ± 0.06**	0.690 ± 0.07*	0.636 ± 0.05**	0.810 ± 0.05*
Single frame+Optical flow+Latent code	Three-Stream	**0.793 ± 0.02**	**0.840 ± 0.06**	**0.752 ± 0.07**	0.680 ± 0.05	**0.880 ± 0.1**

Algorithm 1 Training Scheme of Our Proposed Method

Input: Ocular imaging dataset $\mathcal{D} = \{(\mathbf{D_1}\}$, where $\mathbf{D_1} = (X_s, X_t)$;
Paradigm dataset $\mathcal{D} = \{(\mathbf{D_2}\}$, where $\mathbf{D_2} = (X_p)$.
Output: Prediction $\hat{y}$.

for $i = 1$ to Epoches **do**
  for $k = 1$ to K **do**
    Extraction of temporal and spatial features of ocular imaging:
    $X_s^{out} = \text{ResNet}(X_s')$;
    $X_t^{out} = \text{ResNet}(X_t')$;
    Generate latent code:
    $X_p^{out} = \text{CNN}(X_p')$.
  end for
  Compute the final label $\hat{y}$:
  $\hat{y} = \text{FC}\left(\text{Concat}(X_s^{out}, X_t^{out}, X_p^{out})\right)$.
  Compute loss L.
  Backward L and update parameters.
end for

B. Comparison of Baseline Methods

In order to demonstrate the effectiveness of our proposed method, we choose some classical methods in the field of video recognition as baseline, and classify them according to modality of the input data required by these methods. As shown in Table III, C3D, R3D, R2Plus1D and SlowFast only require input video frames, Two-Stream Convolutional Neural Network requires input video frame and optical flow, while TSCNN requires input video frame, optical flow and latent code of paradigm. Each method was run 5 times, with the results of TSCNN serving as the control group. The t-tests were conducted using ttest_ind function to compare the results with those of other networks. From the results in Table III, TSCNN has the highest classification accuracy of 79.3%, precision of 84.0%, F1 Score of 75.2% and Specificity of 88.0%, which is higher than the baseline and has good performance. To visualize the experimental results, we use a bar chart to represent the distribution of rating metrics as shown in Fig. 5.

C. Ablation

Effectiveness of TSCNN is demonstrated by ablation experiments. We discuss the optimal combination of fast and slow frame rates and the advantages of adding paradigm information.

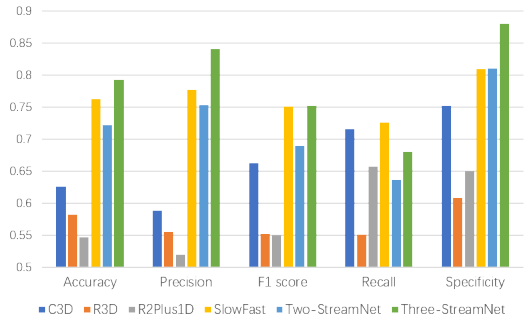

Fig. 5. Comparison of Three-Stream Convolutional Neural Network with various baseline methods on multiple evaluation metrics.

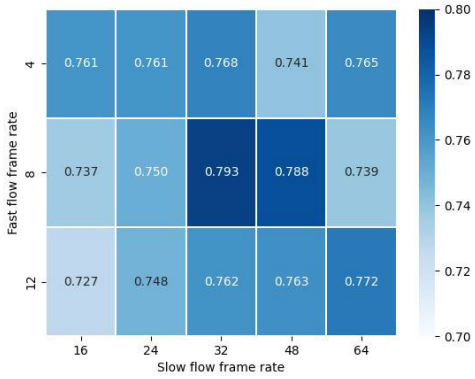

Fig. 6. The frame rate of optical flow is adjusted in TSCNN (Batch Size = 4), using grid search to find the most suitable frame rate.

To find the fast and slow frame rates with the best classification result, we used grid search and adjusted frame rates of optical flow (Batch Size = 4), Fig. 6 shows the result. Eventually, we found that higher accuracy can be obtained when the frame rates of two optical flow were multiples of 8. Specifically, we obtained the highest accuracy when the frame rates were 8 and 32, respectively.

By extracting the x,y coordinates of gaze region and visualizing them in paradigm picture. We found that attention bias to negative stimuli was more pronounced in depression patients, i.e., in subfigure a1 of Fig. 7, the gaze points of patients with depression were mostly concentrated around the injured eye of the cat, while the healthy controls had a relatively larger gaze range and more dispersed gaze points. For positive stimuli,

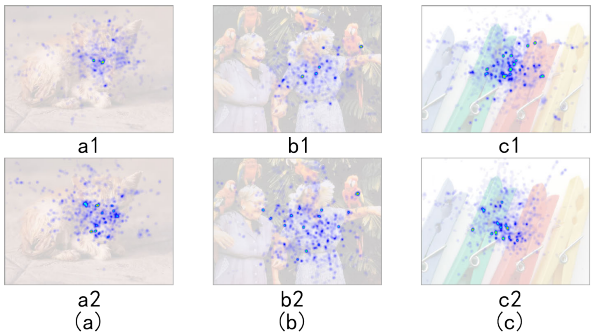

Fig. 7. The heatmaps of gaze points for subjects under three types of stimulus images (negative, positive, and neutral). (a) depicts the gaze points of participants under negative stimuli. (b) shows the gaze points under positive stimuli. (c) illustrates the gaze points under neutral stimuli. Figures a1, b1, c1 represent the gaze point distribution of depression patients, while figures a2, b2, c2 show the gaze point distribution of the healthy control group.

TABLE IV
THE EFFECTIVENESS OF THE PROPOSED AUTO-ENCODER IS TESTED ON THE CLASSICAL TWO-STREAM CONVOLUTIONAL NEURAL NETWORK AND TSCNN APPROACHES. THE TWO METHODS ARE EXPERIMENTED ON SCENARIOS WITH AND WITHOUT INPUT PARADIGM INFORMATION, RESPECTIVELY

Networks	With PD code	Without PD code
Two-Stream Network	**0.726**	0.722
Three-Stream Network	**0.793**	0.779

the range of gaze was similarly larger in healthy controls. For neutral stimuli, the difference between depression patients and healthy controls was not significant. The semantic information of paradigm is necessary for depression recognition.

To test and verify the effectiveness of latent code, we performed ablation experiments on classical Two-Stream Convolutional Neural Network and our method. We found that the accuracy of both models was significantly improved by adding latent code corresponding to paradigm pictures. Table IV records the test results on classical Two-Stream Convolutional Neural Network and our method, the model with addition of latent semantic code is more effective. Accuracy of Two-Stream Convolutional Neural Network improved from 72.2% to 72.6%, and from 77.9% to 79.3% for TSCNN.

VI. DISCUSSION

A. Comparative Analysis

Data-driven modeling approach can automatically learn features and laws of complex systems from a large amount of data without setting complex assumptions and prior knowledge artificially. Two-stream Convolutional Neural Network uses RGB images and optical flow as input to network with good results, optical flow is often considered as "black boxes". So, what makes optical flow effective? Sevilla-Lara et al. [66] experimentally established representational appearance invariance in optical flow. Setting fast and slow frame rates help model learn the process of pupil change and learn eye-movement properties from multiple time dimensions. TSCNN completely considers the characteristics of ocular imaging. Paradigm information is used as prior information input for model to learn some key features. By this way, the performance of the model is maximized. The metrics of the model are listed in Table III, and we can see that the metrics of TSCNN outperform Two-Stream Convolutional Neural Network, which can illustrate the effectiveness of our method.

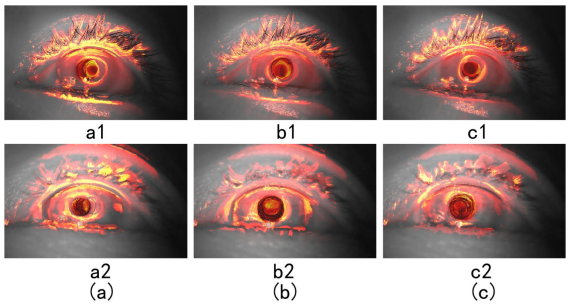

Fig. 8. Dense optical flow maps of different attribute paradigm pictures viewed by two subjects. Subfigures (a), (b), (c) are subjected to positive, neutral, and negative stimuli, respectively. Subfigures a1, b1, c1 are ocular movement optical flow maps of depressed patients. Subfigures a2, b2, c2 are ocular movement optical flow maps of healthy controls.

B. Analysis of the Choice of Fast and Slow Optical Flow

In experiment to find the optimal combination of fast and slow frame rates for optical flow, the experimental result was fast and slow frame rates of 8 and 32, respectively, we speculate the reasons for this as follows. The idea of fast and slow frame rate is using fast frame rate to collect ocular change information when staying within an image, and slow frame rate to collect ocular change information when switching between images. If the difference between fast and slow frame rate is not too big, it will degenerate towards Two-Stream Convolutional Neural Network. If the difference is too big, slow frame rate with a wide range of extraction times, will lose the correlation between extracted frames, and may also have a semantic gap between two frames. When fast and slow frame rates are 8 and 32 respectively, above conditions are just met, so the best results can be achieved.

C. Analysis of Differences in Ocular Changes Between Patients With Depression and Healthy Controls

We extracted dense optical flow of subjects' ocular images when they viewed paradigm pictures with different attributes. As shown in Fig. 8, circles in pupils of patients with depression are clearer, whereas healthy controls have traces of circles around pupils left by movement, suggesting that more flexible pupil movements in healthy controls, individuals with depression have dull eyes. To obtain numerical features relevant to this conclusion, we analyzed pupil trajectories using sparse optical flow.

It was mentioned in section Experimental Results, ocular movement under negative stimuli differed between patients with depression and healthy controls. We set the size of ocular images to 320 × 200, select ocular images of all subjects with a duration of 5 seconds while viewing paradigm picture (injured cat) shown in subfigure (a) of Fig. 7. Using Lucas-Kanade (LK) optical flow method, corner point was set on pupil, its horizontal and vertical coordinates were recorded, the distances of its movement were calculated to derive the mean, standard deviation, and median of pupil displacements

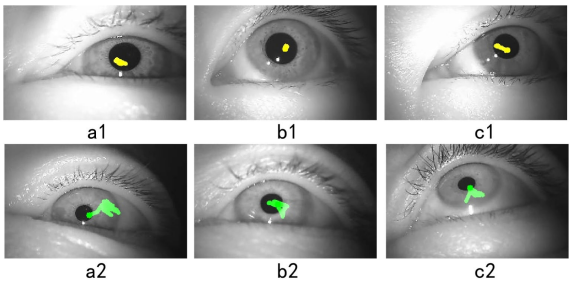

Fig. 9. Lucas-Kanade (LK) optical flow maps of ocular movement in subjects exposed to negative stimuli. Subfigures a1, b1, c1 are optical flow maps of ocular movement trajectories in patients with depression (yellow). Subfigures a2, b2, c2 are optical flow maps of ocular movement trajectories in healthy controls (green).

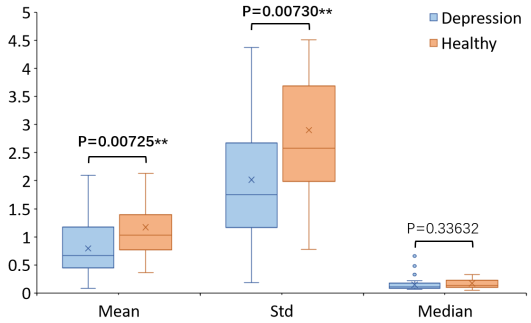

Fig. 10. Box-plot of mean, standard deviation (std), and median pupil movement distance for all patients with depression (blue) as well as healthy controls (orange). T-test was performed for these three characteristics labeled in box-plot, ** : extremely significant difference ($P<0.01$), * : significant difference ($P< 0.05$).

for each subject, obtaining an optical flow map of subject's eye movement trajectory, sparse optical flow maps are shown in Fig. 9. A smaller range of pupil movement trajectories can be seen in patients with depression and more flexible pupil movements in healthy controls.

T-test concluded that mean and standard deviation are significantly different, their P-values are 0.00725 and 0.00730, respectively, specific results are shown in Fig. 10.

VII. Conclusion

In this paper, we propose TSCNN that combines spatial, temporal and paradigmatic information to extract features directly from raw ocular imaging to detect depression. In temporal module, we design two data streams with fast and slow frame rate to extract the ocular change features of subjects under different viewing states. The experimental results show that our method outperforms other methods in each evaluation index, this may provide ideas for future research on detecting depression based on ocular imaging. By analyzing the results of ablation experiments, introducing prior information (latent code of paradigm images) helps to improve the classification accuracy, which may inspire other experiments, when experimental results do not meet expectations, a prior information can be introduced to assist the model in learning.

From the analysis in this paper, patients with depression are more sensitive to negative stimuli and have more pronounced behavioral performance or eye movement responses, compared with healthy controls, patients with depression have less pronounced pupil changes and relatively dull eyes when facing stimuli. In future studies, the specific content, emotional validity, and arousal of the paradigm pictures can be considered in conjunction with ocular imaging features in order to fully utilize paradigm stimulus pictures. In addition, multimodal fusion can provide complementary information to extracted features as a way to facilitate depression detection. In the future, we will try to multimodal fusion of eye movements with facial expressions, audio, etc.

References

[1] *The Global Burden of Disease: 2004 Update*, World Health Organization, Geneva, Switzerland, 2008.
[2] *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*, American Psychiatric Association, Washington, DC, USA, 2013, vol. 5, no. 5.
[3] D. A. Dunstan, N. Scott, and A. K. Todd, "Screening for anxiety and depression: Reassessing the utility of the Zung scales," *BMC Psychiatry*, vol. 17, no. 1, pp. 1–8, Dec. 2017.
[4] M. N. Norhayati, N. H. N. Hazlina, A. R. Asrenee, and W. M. A. W. Emilin, "Magnitude and risk factors for postpartum symptoms: A literature review," *J. Affect. Disorders*, vol. 175, pp. 34–52, Apr. 2015.
[5] J. Chen, S. Sun, L.-B. Zhang, B. Yang, and W. Wang, "Compressed sensing framework for heart sound acquisition in Internet of medical things," *IEEE Trans. Ind. Informat.*, vol. 18, no. 3, pp. 2000–2009, Mar. 2022.
[6] A. T. Drysdale et al., "Resting-state connectivity biomarkers define neurophysiological subtypes of depression," *Nature Med.*, vol. 23, no. 1, pp. 28–38, Jan. 2017.
[7] Z. Dai, H. Zhou, Q. Ba, Y. Zhou, L. Wang, and G. Li, "Improving depression prediction using a novel feature selection algorithm coupled with context-aware analysis," *J. Affect. Disorders*, vol. 295, pp. 1040–1048, Dec. 2021.
[8] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, H. Adeli, and D. P. Subha, "Automated EEG-based screening of depression using deep convolutional neural network," *Comput. Methods Programs Biomed.*, vol. 161, pp. 103–113, Jul. 2018.
[9] K. M. Puk et al., "Emotion recognition and EEG analysis using ADMM-based sparse group lasso," *IEEE Trans. Affect. Comput.*, vol. 13, no. 1, pp. 199–210, Jan. 2022.
[10] H. Cai et al., "A pervasive approach to EEG-based depression detection," *Complexity*, vol. 2018, Jan. 2018, Art. no. 5238028.
[11] W. C. de Melo, E. Granger, and M. B. López, "MDN: A deep maximization-differentiation network for spatio-temporal depression detection," *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 578–590, Jan. 2023.
[12] M. Yang, Y. Ma, Z. Liu, H. Cai, X. Hu, and B. Hu, "Undisturbed mental state assessment in the 5G era: A case study of depression detection based on facial expressions," *IEEE Wireless Commun.*, vol. 28, no. 3, pp. 46–53, Jun. 2021.
[13] M. Yang, Y. Wu, Y. Tao, X. Hu, and B. Hu, "Trial selection tensor canonical correlation analysis (TSTCCA) for depression recognition with facial expression and pupil diameter," *IEEE J. Biomed. Health Informat.*, to be published, doi: 10.1109/JBHI.2023.3322271.
[14] T. Al Hanai, M. M. Ghassemi, and J. R. Glass, "Detecting depression with audio/text sequence modeling of interviews," in *Proc. Interspeech*, Sep. 2018, pp. 1716–1720.
[15] G. Sharma, A. Dhall, and J. Cai, "Audio-visual automatic group affect analysis," *IEEE Trans. Affect. Comput.*, to be published.
[16] Y. Wang et al., "The similar eye movement dysfunction between major depressive disorder, bipolar depression and bipolar mania," *World J. Biol. Psychiatry*, vol. 23, no. 9, pp. 689–702, Oct. 2022.
[17] T. Suslow, A. Hußlack, A. Kersting, and C. M. Bodenschatz, "Attentional biases to emotional information in clinical depression: A systematic and meta-analytic review of eye tracking findings," *J. Affect. Disorders*, vol. 274, pp. 632–642, Sep. 2020.
[18] T. Armstrong and B. O. Olatunji, "Eye tracking of attention in the affective disorders: A meta-analytic review and synthesis," *Clin. Psychol. Rev.*, vol. 32, no. 8, pp. 704–723, Dec. 2012.
[19] M. Yang, X. Feng, R. Ma, X. Li, and C. Mao, "Orthogonal-moment-based attraction measurement with ocular hints in video-watching task," *IEEE Trans. Computat. Social Syst.*, to be published.

[20] X. Ding, X. Yue, R. Zheng, C. Bi, D. Li, and G. Yao, "Classifying major depression patients and healthy controls using EEG, eye tracking and galvanic skin response data," *J. Affect. Disorders*, vol. 251, pp. 156–161, May 2019.

[21] Y. Yuan and Q. Wang, "Detection model of depression based on eye movement trajectory," in *Proc. IEEE Int. Conf. Data Sci. Adv. Anal. (DSAA)*, Oct. 2019, pp. 612–613.

[22] A. C. Mennen, K. A. Norman, and N. B. Turk-Browne, "Attentional bias in depression: Understanding mechanisms to improve training and treatment," *Current Opinion Psychol.*, vol. 29, pp. 266–273, Oct. 2019.

[23] R. H. Kaiser, H. R. Snyder, F. Goer, R. Clegg, M. Ironside, and D. A. Pizzagalli, "Attention bias in rumination and depression: Cognitive mechanisms and brain networks," *Clin. Psychol. Sci.*, vol. 6, no. 6, pp. 765–782, Nov. 2018.

[24] A. M. Klein, L. de Voogd, R. W. Wiers, and E. Salemink, "Biases in attention and interpretation in adolescents with varying levels of anxiety and depression," *Cogn. Emotion*, vol. 32, no. 7, pp. 1478–1486, Oct. 2018.

[25] N. Carvalho et al., "Eye movement in unipolar and bipolar depression: A systematic review of the literature," *Frontiers Psychol.*, vol. 6, p. 1809, Dec. 2015.

[26] K. L. Burkhouse, G. J. Siegle, M. L. Woody, A. Y. Kudinova, and B. E. Gibb, "Pupillary reactivity to sad stimuli as a biomarker of depression risk: Evidence from a prospective study of children," *J. Abnormal Psychol.*, vol. 124, no. 3, pp. 498–506, Aug. 2015.

[27] J. Bittencourt et al., "Saccadic eye movement applications for psychiatric disorders," *Neuropsychiatric Disease Treatment*, vol. 9, pp. 1393–1409, Sep. 2013.

[28] M. Li et al., "Method of depression classification based on behavioral and physiological signals of eye movement," *Complexity*, vol. 2020, pp. 1–9, Jan. 2020.

[29] S. Alghowinem, R. Goecke, M. Wagner, G. Parker, and M. Breakspear, "Eye movement analysis for depression detection," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2013, pp. 4220–4224.

[30] S. Al-gawwam and M. Benaissa, "Depression detection from eye blink features," in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol. (ISSPIT)*, Dec. 2018, pp. 388–392.

[31] R. Shen, Q. Zhan, Y. Wang, and H. Ma, "Depression detection by analysing eye movements on emotional images," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 7973–7977.

[32] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, "DepAudioNet: An efficient deep model for audio based depression classification," in *Proc. 6th Int. Workshop Audio/Visual Emotion Challenge*, Oct. 2016, pp. 35–42.

[33] L. Yang, D. Jiang, X. Xia, E. Pei, M. C. Oveneke, and H. Sahli, "Multimodal measurement of depression using deep learning models," in *Proc. 7th Annu. Workshop Audio/Visual Emotion Challenge*, Oct. 2017, pp. 53–59.

[34] Y. Tao, M. Yang, Y. Wu, K. Lee, A. Kline, and B. Hu, "Depressive semantic awareness from vlog facial and vocal streams via spatio-temporal transformer," *Digit. Commun. Netw.*, to be published, doi: 10.1016/j.dcan.2023.03.007.

[35] Y. Pan et al., "Spatial–temporal attention network for depression recognition from facial videos," *Expert Syst. Appl.*, vol. 237, Mar. 2024, Art. no. 121410.

[36] W. C. de Melo, E. Granger, and M. B. Lopez, "Facial expression analysis using decomposed multiscale spatiotemporal networks," *Expert Syst. Appl.*, vol. 236, Feb. 2024, Art. no. 121276.

[37] M. Yang, C. Cai, and B. Hu, "Clustering based on eye tracking data for depression recognition," *IEEE Trans. Cognit. Develop. Syst.*, doi: 10.1109/TCDS.2022.3223128.

[38] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 568–576.

[39] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4724–4733.

[40] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[41] Y. Li et al., "Eye movement indices in the study of depressive disorder," *Shanghai Arch. Psychiatry*, vol. 28, no. 6, p. 326, 2016.

[42] X. Li, T. Cao, S. Sun, B. Hu, and M. Ratcliffe, "Classification study on eye movement data: Towards a new approach in depression detection," in *Proc. IEEE Congr. Evol. Comput. (CEC)*, Jul. 2016, pp. 1227–1232.

[43] Z. Pan, H. Ma, L. Zhang, and Y. Wang, "Depression detection based on reaction time and eye movement," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 2184–2188.

[44] J. Zhao and Q. Wang, "Eye movement attention based depression detection model," in *Proc. IEEE 9th Int. Conf. Data Sci. Adv. Anal. (DSAA)*, Oct. 2022, pp. 1–2.

[45] 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.

[46] Y. Mao, Y. He, L. Liu, and X. Chen, "Disease classification based on synthesis of multiple long short-term memory classifiers corresponding to eye movement features," *IEEE Access*, vol. 8, pp. 151624–151633, 2020.

[47] W. Xie, L. Liang, Y. Lu, H. Luo, and X. Liu, "Deep 3D-CNN for depression diagnosis with facial video recording of self-rating depression scale questionnaire," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 2007–2010.

[48] W. Carneiro de Melo, E. Granger, and M. B. Lopez, "Encoding temporal information for automatic depression recognition from facial analysis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 1080–1084.

[49] 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.

[50] T. Karras et al., "Alias-free generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 852–863.

[51] A. T. Duchowski, "A breadth-first survey of eye-tracking applications," *Behav. Res. Methods, Instrum., Comput.*, vol. 34, no. 4, pp. 455–470, Nov. 2002.

[52] M. Yang, Y. Gao, L. Tang, J. Hou, and B. Hu, "Wearable eye-tracking system for synchronized multimodal data acquisition," *IEEE Trans. Circuits Syst. Video Technol.*, to be published, doi: 10.1109/TCSVT.2023.3332814.

[53] J. Girgus, K. Yang, and C. Ferri, "The gender difference in depression: Are elderly women at greater risk for depression than elderly men?" *Geriatrics*, vol. 2, no. 4, p. 35, Nov. 2017.

[54] M. Piccinelli and G. Wilkinson, "Gender differences in depression: Critical review," *Brit. J. Psychiatry*, vol. 177, no. 6, pp. 486–492, Dec. 2000.

[55] Y. Gou et al., "Province- and individual-level influential factors of depression: Multilevel cross-provinces comparison in China," *Frontiers Public Health*, vol. 10, May 2022, Art. no. 893280.

[56] D. R. van Renswoude, M. E. J. Raijmakers, A. Koornneef, S. P. Johnson, S. Hunnius, and I. Visser, "Gazepath: An eye-tracking analysis tool that accounts for individual differences and data quality," *Behav. Res. Methods*, vol. 50, no. 2, pp. 834–852, Apr. 2018.

[57] C. Jayaro, I. De La Vega, M. Díaz-Marsá, A. Montes, and J. Carrasco, "The use of the international affective picture system for the study of affective dysregulation in mental disorders," *Actas Españolas de Psiquiatría*, vol. 36, no. 3, pp. 177–182, 2008.

[58] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[59] G. Farneback, "Two-frame motion estimation based on polynomial expansion," in *Proc. 13th Scand. Conf. Image Anal. (SCIA)*, Halmstad, Sweden, Berlin, Germany: Springer, Jan. 2003, pp. 363–370.

[60] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Mali, Oct. 2019, pp. 6201–6210.

[61] E. Tiu. (Apr. 2020). *Understanding Latent Space in Machine Learning*. [Online]. Available: https://towardsdatascience.com/understanding-latent-space-in-machine-learning-de5a7c687d8d

[62] K. Huang, B. Shi, X. Li, X. Li, S. Huang, and Y. Li, "Multi-modal sensor fusion for auto driving perception: A survey," 2022, *arXiv:2202.02703*.

[63] W. Guo, J. Wang, and S. Wang, "Deep multimodal representation learning: A survey," *IEEE Access*, vol. 7, pp. 63373–63394, 2019.

[64] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 8024–8035.

[65] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[66] L. Sevilla-Lara, Y. Liao, F. Güney, V. Jampani, A. Geiger, and M. J. Black, "On the integration of optical flow and action recognition," in *Proc. 40th German Conf. Pattern Recognit. (GCPR)*, Stuttgart, Germany, Cham, Switzerland: Springer, Oct. 2019, pp. 281–297.