Multimodal Depression Detection Based on Self-Attention Network With Facial Expression and Pupil

Xiang Liu, Hao Shen, Huiru Li, Yongfeng Tao, *Graduate Student Member, IEEE*, and Minqiang Yang, *Member, IEEE*

***Abstract*—Depression is a major mental health issue in contemporary society, with an estimated 350 million people affected globally. The number of individuals diagnosed with depression continues to rise each year. Currently, clinical practice relies entirely on self-reporting and clinical assessment, which carries the risk of subjective biases. In this article, we propose a multimodal method based on facial expression and pupil to detect depression more objectively and precisely. Our method first extracts the features of facial expressions and pupil diameter using residual networks and 1-D convolutional neural networks. Second, a cross-modal fusion model based on self-attention networks (CMF-SNs) is proposed, which utilizes cross-modal attention networks within modalities and parallel self-attention networks between different modalities to extract CMF features of facial expressions and pupil diameter, effectively complementing information between different modalities. Finally, the obtained features are fully connected to identify depression. Multiple controlled experiments show that compared to single modality, the multimodal fusion method based on self-attention networks has a higher ability to recognize depression, with the highest accuracy of 75.0%. In addition, we conducted comparative experiments under three different stimulation paradigms, and the results showed that the classification accuracy under negative and neutral stimuli was higher than that under positive stimuli, indicating a bias of depressed patients toward negative images. The experimental results demonstrate the superiority of our multimodal fusion method.**

***Index Terms*—Facial expression, multimodal fusion, pupil, self-attention network.**

I. Introduction

According to the World Health Organization (WHO), depression has become one of the most pressing mental health concerns in modern society [1]. Currently, over 350 million people worldwide, or roughly 5% of the global population, are affected by depression, and in developed countries, the percentage of patients with depressive disorders may be as high as 10% [2]. A person's employment, studies, and daily life can all be negatively impacted by depression. It can cause sorrow, irritation, diminished wellbeing, exhaustion, sleeplessness, difficulty concentrating, a sense of hopelessness, worry, and low self-esteem, among other symptoms. In more severe cases, depression can even lead to suicidal tendencies [3], [4], [5], [6]. Also, more frightening is the increased risk of diseases such as cardiovascular disease, cancer, and diabetes in the case of increased depression [7].

Given the serious consequences of depression, it is crucial to identify and assess the condition as early as possible to ensure patients receive timely and effective treatment. At present, the diagnosis of depression is mainly based on clinical assessments through interviews and questionnaires. However, these diagnostic techniques rely entirely on the expertise of the physicians and their clinical experience and are known to be based on their subjective evaluations [8]. As the number of individuals diagnosed with depression continues to rise, there is an increasing demand for more efficient and accurate diagnosis methods. Depending solely on physicians' subjective diagnosis and subsequent treatment can be restricted and susceptible to inaccuracies. There is an urgent need for an objective and effective method to support the diagnosis of depression, particularly given the shortage of medical professionals available to diagnose and treat this condition. With the quick advancement of artificial intelligence, machine learning-based depression assessment [9], [10] is useful for accurate and prompt diagnosis of depression, guaranteeing that patients may obtain timely and efficient treatment, and enhancing people's sense of wellbeing.

All symptoms of depression can be attributed to mood disorders characterized by depressed mood. Emotions are made up of three components: subjective feelings, physiological arousal,

Manuscript received 21 February 2024; revised 23 April 2024; accepted 21 May 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62071122, in part by the Outstanding Youth Project of Guangdong Basic and Applied Basic Research Foundation under Grant 2023B1515020064, in part by the Science and Technology Innovation 2030-Major Projects under Grant 2021ZD0202002, in part by the National Key Research and Development Program of China under Grant 2019YFA0706200, in part by the Natural Science Foundation of Gansu Province, China under Grant 22JR5RA488, in part by the Fundamental Research Funds for the Central Universities under Grant lzujbky-2023-16, and in part by Supercomputing Center of Lanzhou University. *(Xiang Liu and Hao Shen are co-first authors.) (Corresponding author: Minqiang Yang.)*

Xiang Liu is with the School of Computer Science and Technology, Dongguan University of Technology, Dongguan 523000, China (e-mail: liuxiang@dgut.edu.cn).

Hao Shen, Huiru Li, Yongfeng Tao, and Minqiang Yang are with the School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, China (e-mail: 220220942151@lzu.edu.cn; hrli20@lzu.edu.cn; taoyf21@lzu.edu.cn; yangmq@lzu.edu.cn).

Digital Object Identifier 10.1109/TCSS.2024.3405949

and external manifestations [11]. The key elements of outward manifestations are changes in voice and expression, which are two crucial signs of mood shifts. Research has indicated that individuals with depression may exhibit lower levels of social behavior compared to healthy individuals, such as fewer changes in facial expressions [12]. The residual mechanism is beneficial to extract more informative features in deeper networks [13]. To extract high-level semantic features from facial expressions, we utilize a residual network (ResNext) in our approach.

The manifestations of depression also include eye contact, which can reflect changes in a subject's motor and mental status by recording eye movements in response to specific stimuli. There is evidence that depressed individuals often look at negative emotional material for long periods of time and pay less attention to positive emotional material [14]. Several studies using eye-movement measures for diagnostic classification of depression have also been successful [15], [16], [17]. Behavioral information about eye movements is usually measured by fixation, saccade, blink, and pupil diameter. Eye-movement behaviors are actually covered in facial video streams. As a physiological feature that effectively indicates depression [18], pupil diameter deserves a fusion study with other modalities. We use 1-D convolutional neural network (CNN) to extract pupil features.

The choice of method for fusing multiple modalities of information can significantly impact the effectiveness and performance of multimodal depression recognition systems [19]. There are several main methods for multimodal data fusion, including feature-based fusion [20], [21], decision-based fusion [22], and hybrid fusion [23]. Feature-based fusion primarily takes into account the complementary nature of different modalities' features and their relationships to improve the accuracy of depression recognition. In contrast, decision-based fusion primarily relies on the integration and analysis of information from each modality to improve decision reliability by fusing features from different levels. The hybrid fusion method combines feature-based and decision-based fusion methods to further enhance the performance of depression recognition. This allows for more comprehensive and effective integration of multiple modalities for improving recognition accuracy.

In previous studies, we conducted research on depression recognition using single-modality facial expressions [24] and eye movement [25] separately. Therefore, this study focuses on the complementary nature of features between two modalities and proposes a cross-modal fusion (CMF) model based on self-attention networks for depression recognition. Specifically, we first learn the facial expression modality and pupil modality, using ResNext to obtain the spatiotemporal features of the video frame facial expression sequence and using 1-D CNN to obtain the pupil features. Furthermore, the obtained single-modality features are simultaneously input into the CMF model to obtain more effective features that capture the relationships between the two modalities. At the same time, we perform intermodal feature selection on the two modal features separately using self-attention mechanisms, obtaining the features of the two modalities within and between modalities. Finally, we connect the features and obtain the output of the depression recognition. We evaluated our model on a dataset for multimodal depression recognition, experimental results demonstrate that the proposed method achieves a classification accuracy of 75.0%, indicating its effectiveness for depression recognition research. In summary, we make the following contributions.

1) We propose a novel CMF-SN to extract fusion features from facial expressions and pupil diameter by using attention networks, which enhances information integration across modalities.
2) To enable effective feature selection for CMF and accurate depression recognition, we use parallel self-attention networks to preserve original semantic information within and between modalities.
3) To adequately fuse adjacent feature elements, we utilize the self-attention mechanism and project different modalities onto separate spaces.

The rest of this article is structured as follows. Section II discusses related works. Section III explains the techniques used for preprocessing and extracting features from facial expressions and pupil diameter, along with the multimodal fusion strategy. Section IV outlines the experimental design and experimental protocols. Section V demonstrates the outcomes of the experiment. Section VI furnishes an elaborate examination and discourse on the conclusions inferred from the experiment.

II. Related Works

Multimodal feature fusion method primarily involves two key aspects: first, feature extraction; and second, the interaction between different modalities. Currently, there are two main approaches for extracting facial expression features in depression recognition research, namely geometry-based techniques and appearance-based techniques. Facial landmarks are the essential component of geometry-based face recognition systems, which divide the face picture into several areas and then identify between expressions by taking into account the geometric changes in these regions. For instance, differential geometric fusion network (DGFN) [26] and deep action unit graph network (DAUGN) [27]. While DAUGN employs facial signs and the Voronoi diagram (VD) technique to transform face pictures into facial graphics, DGFN specifically combines facial signs to create different characteristics corresponding to action units (AUs) [28]. Geometry-based features, on the other hand, significantly rely on landmark identification techniques and are unable to record face motions that result in landmark displacement. Appearance-based techniques utilize textual information, such as local binary patterns (LBPs) [29], [30], Gabor wavelets [31], and pyramid histogram of gradients-three orthogonal planes (PHOG-TOPs) [32] to capture the variation of facial texture in different expressions. For appearance-based features, they are affected by background noise or facial organ deformation. Since the changes of expressions are very subtle, it is possible that the noise of small images will affect the weights of similar expressions and reduce the effectiveness of the weights. The feature extraction algorithms discussed above primarily focus on low-level feature extraction, which has been found to be inadequate for accurately recognizing depression. Deep neural network models can transform lower level features into higher level feature representations through feature selection using different network architectures. This enhances the ability to capture more effective representations of information.

In the field of depression recognition, the results of multimodal fusion methods outperform those of single modalities due to multifarious significant manifestations of depressive disorder [33], [34], [35], [36]. Technically, various types of data can be used for multimodal fusion, such as speech sequences [37], facial expressions [38], text [39], and electroencephalogram (EEG) [40]. On the research of pupil related data and EEG fusion, Zhu et al. [41] used mutual information to measure the relevance between EEG and pupil area signals, and selected EEG electrodes based on mutual information. Then, they fused bimodal features using the denoising autoencoder. Zhang et al. [42] integrated eye-movement information into EEG and divide the features into groups according to their respective characteristics. To obtain a fusion representation of EEG and eye movement, they used group sparse canonical correlation analysis (GSCCA). On the research of visual feature and other features fusion, Tao et al. [43] proposed a plug-and-play multimodal spatiotemporal fusion attention module (STAT), which captured the global dependencies of temporal and spatial information of visual and acoustic sequences in a video stream. Ghosh et al. [44] proposed multimodal MT profile information encoder, utilizing image and text in Twitter tweets to infer users' depression status and emotion. Existing multimodal fusion methods tend to focus on leveraging the complementary nature of information within each modality. While research results have shown the effectiveness of this approach for depression recognition, it does not consider the potential loss of information during the fusion process. These methods only take into account information complementarity, while ignoring the importance of each modality's inherent features for decision-making.

III. Methods

The recognition of depression based on video and multimodal approaches has received wide attention due to its robustness. However, current multimodal fusion methods mainly focus on the complementarity of different modalities during fusion, while ignoring the importance of inherent information contained by each modality feature, i.e., the different information they provide when expressing the same object or concept [45], [46]. Failure to consider the complementary information during multimodal fusion may also result in information loss, which can negatively impact the performance of the model for depression recognition. To address this issue, researchers have proposed methods such as attention mechanisms and gate mechanisms, which allow the model to dynamically allocate and select weights between different modalities, to retain important information and suppress noise, better addressing the multimodal problem in facial expressions. In this article, a CMF-SN is proposed for multimodal depression recognition. First, efficient ResNext-50 (32 × 4d) and 1-D CNN are used for representation learning of facial expressions and pupil modalities, respectively, to obtain features of the two modalities. Then, the modal fusion features of facial expressions and pupils are extracted through the intramodal self-attention cross-module and parallel self-attention network between different modalities. The general structure of the model is illustrated in Fig. 1.

A. Preprocessing

1) Facial Image Preprocessing: The preprocessing of face data involved in this article mainly includes face alignment, cropping, and normalization. The frame rate of facial video is 30 fps. According to the timestamps corresponding to all trials in the paradigm, the facial expression data of each subject are divided into 40 segments. We first converted the original image to a grayscale image and set brightness between 0.3 and 1.0. Then, we used the Face_alignment library in the Python toolkit to automatically detect the faces of subjects in each video. We extract regions restricted to faces, removing background distractions such as hairstyles and accessories. Notably, for the cases that the hands of subjects cover his/her faces or something blocks the camera, it will fail to detect faces in those frames. In these cases, we just ignore the frames and interpolate. The effect of frequency on illumination is subsequently incredibly weak, as shown by [47]. This study utilized histogram equalization [48] to improve the image contrast in video frames. This technique enhances the contrast of image regions with low local contrast and can improve details in frames with underexposure or overexposure.

2) Pupil Diameter Preprocessing: Eye trackers often output sight coordinates, pupil diameter, and related parameters. The frame rate of our eye tracker is 200 fps. This article only uses pupil diameter data. Subjects may blink frequently or close their eyes for a long time during the experiment, which will cause the eye tracker to fail to capture pupils, resulting in blank or invalid values. This article directly ignores the missing and invalid values [49] to prevent their influence on data analysis. The pupil diameter data with confidence scores less than 0.5 are removed. Additionally, the pupil diameter is normalized with min-max method [50] to remove the impact of the dimension between characteristics, which might make processing subsequent data easier and faster convergence. Same as facial expression data, we divided pupil diameter into 40 segments.

B. Facial Expression Features

Considering the performance of the network and training efficiency, we used the ResNext-50 (32 × 4d) network to extract facial expression features. One typical approach to enhance a model's recognition accuracy is to enlarge the depth or width of the network. However, increasing the hyperparameters of the network is not a simple way to improve the model's performance, as it requires a balance between optimizing the model's performance and computational cost. The ResNext-50 network integrates the ResNet network's stacking approach with the package convolution technique of the initial network to create a more effective and robust neural network model. The ResNext-50 network differs from ResNet in that it replaces the three-layer convolution residual block with a residual block of identical topology structure, which is then stacked in parallel

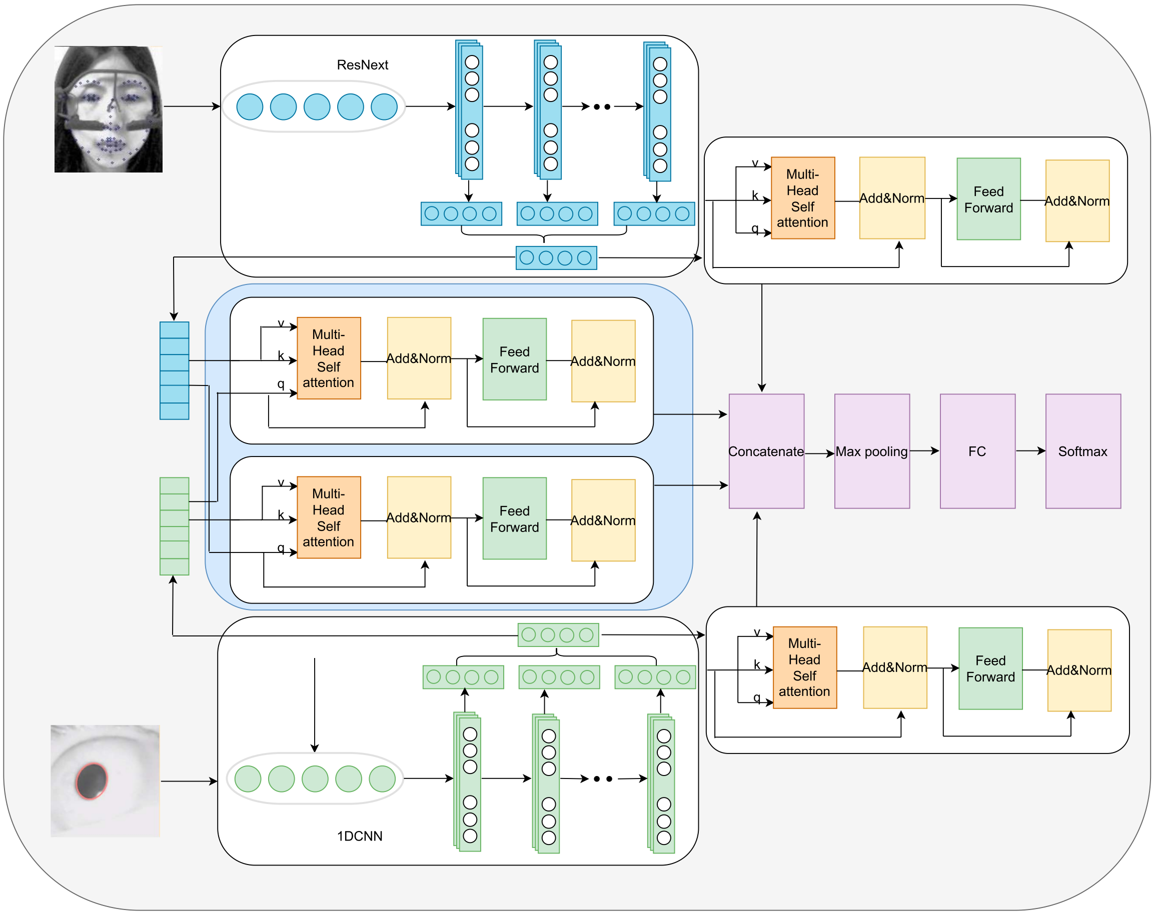

Fig. 1. Multimodal fusion flow diagram of self-attention network. The dark blue feature is facial expression feature extracted from ResNext and the green feature is pupil diameter feature extracted from 1-D CNN network. The upper and lower right sections represent the single-mode self-attention network model. The blue part in the middle represents the multimode fusion within and between modes.

to enhance its performance. The ResNext-50 network offers several advantages over the ResNet network. First, ResNext-50 networks can enhance the accuracy of modalities without raising the complexity of their parameters by leveraging their network structure, which is the first advantage. This is because the ResNext-50 network uses a technique called model compression, which increases the network's width rather than depth to improve the model's expressive ability and generalization performance. This approach not only improves the model's accuracy but also reduces the risk of overfitting. The second benefit of the ResNext-50 network is that it can decrease the number of hyperparameters required. This is because the ResNext-50 network uses a technique called "group convolution," which decomposes the convolution operation into multiple small convolution operations, reducing the parameter volume and computational cost of each convolution layer. This technique not only enhances the computational efficiency and speed of the model but also reduces the number of hyperparameters necessary. The facial expression extraction network consists of a ResNext (32 × 4d) model [51] without structure modification, and we retrained it based on pretrained model weights because of the limited amount of our collected dataset.

C. Pupil Features

CNN is a deep neural network structure that uses convolutional operations, which can learn representations of input information using hierarchical structures and achieve translation invariance. Generally speaking, 1-D CNN can capture local patterns and features of sequence data by sliding convolutional kernels on the time axis, 2-D CNN can perform convolutional operations on 2-D images to capture local features and spatial structure information, while 3-D CNN can perform convolutional operations on 3-D images and videos to capture local features and spatiotemporal structure information. Hence, choosing a suitable CNN model is crucial for various kinds of data, as it can enhance the model's efficiency and effectiveness. The pupil feature extraction technique employed in this article is based on 1-D CNN. The structure of the 1-D CNN network comprises of two convolutional layers, two pooling layers, and one fully connected layer. Just like 2-D CNN, 1-D CNN operates on the input data through convolutional and pooling layers and performs classification using the fully connected layer at the end. Meanwhile, it addresses the issue of losing a significant amount of feature information that occurs when 2-D CNN compresses the dimensions of the input data. Technically,

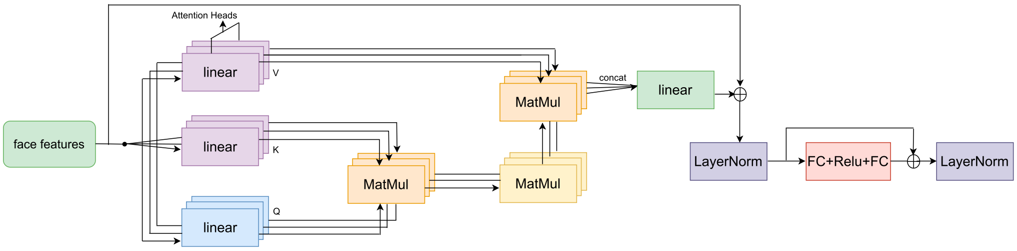

Fig. 2. Expression self-attention network structure.

we chose 3 as kernel size, 2 as stride, and did not do padding operation. ReLU is the activation function. We chose 1-D max pooling with kernel size 2 as pooling operation.

D. Multimodal Fusion

1) Self-Attention Module for Facial Expressions: The self-attention module, in parallel with the cross-attention module, is designed to capture how different modalities, such as facial expressions and pupil, interact with each other. The self-attention mechanism operates in a similar way to the cross-attention mechanism, but instead of using query, key, and value information from different modalities, it uses them from the same modality. As shown in Fig. 2. Thus, the complete process can be summarized as follows: using a multihead self-attention model, the sequence $S = [s_1, s_2, ..., s_n]$ is used as input, as shown in the following equation:

$$\text{MultiHead}(S, S, S) = \text{Concat}(\text{head}_1, ..., \text{head}_i)W^o \quad (1)$$

$$\text{head}_i = \text{Attention}(SW_i^Q, SW_i^K, SW_i^V). \quad (2)$$

The self-attention module applies linear transformations as its initial step in the model. S is multiplied by trainable parameter matrices W_i^Q, W_i^K, and W_i^V to obtain Q, K, and V, respectively. The following step involves feeding the scaled dot-product attention module, which iterates i times, as demonstrated in (3). Ultimately, the multihead attention values are obtained by concatenating the i attention results of the scaled dot-product and linearly transforming them. By using linear transformations in the self-attention module, the model can effectively identify and learn important information in distinct subspaces, which is one of the main benefits of this approach. The scaled dot-product employed in the calculation of the attention module is displayed as follows:

$$\text{Attention}\left(SW_i^Q, SW_i^K, SW_i^V\right) = \text{softmax}\left(\frac{SW_i^Q\left(SW_i^K\right)^T}{\sqrt{d_k}}\right) SW_i^V. \quad (3)$$

To avoid excessively large values for the last dimension of W_i^Q, W_i^K, and W_i^V, d_k is divided by the square root of d_k in the self-attention module. An excessively large value of QK^T can lead to the gradient of the softmax function becoming very small. Then, the LayerNorm layer is used to normalize the attention values of the network, improving the stability and convergence speed of the network during training. In a multilayer network architecture, the commonly used in the following equation:

$$S_l = \text{LayerNorm}(S + \text{MultiHead}(S, S, S)). \quad (4)$$

The LayerNorm used in this context refers to a normalization technique applied to each layer of the model. Normalizing the input data by adjusting its mean and variance to a consistent level can facilitate faster convergence during model training by ensuring standardized features. The output L, which has been processed through the self-attention module and LayerNorm, is then inputted into a feedforward layer that comprises two fully connected layers. According to the (5), the first layer of the feedforward layer applies the ReLU activation function, whereas the second layer does not apply any activation function

$$S_f = \max(0, S_l W_1 + b_1)W_2 + b_2. \quad (5)$$

Similarly, the output S_E is obtained by normalizing using the LayerNorm layer, as shown in the following equation:

$$S_E = \text{LayerNorm}(S_l + S_f). \quad (6)$$

2) Self-Attention Module for Pupils: Regarding the self-attention module for pupils, it is similar to the facial expression module. Using a multihead self-attention model, the sequence $T = [t_1, t_2, ..., t_n]$ is used as input, as shown in

$$\text{MultiHead}(T, T, T) = \text{Concat}(\text{head}_1, ..., \text{head}_j)W^o \quad (7)$$

$$\text{head}_j = \text{Attention}(TW_j^Q, TW_j^K, TW_j^V). \quad (8)$$

Normalization structure using LayerNorm layer

$$T_l = \text{LayerNorm}(T + \text{MultiHead}(T, T, T)). \quad (9)$$

Once the decoder stack produces the output T_l, it is fed into a feedforward layer that contains two fully connected layers. The feedforward layer that T_l passes through includes a first layer that utilizes the ReLU activation function, followed by a second layer that does not apply any activation function

$$T_f = \max(0, LW_3 + b_3)W_4 + b_4. \quad (10)$$

Normalization is used again with the LayerNorm layer to obtain the output T_E

$$T_E = \text{LayerNorm}(T_l + T_f). \quad (11)$$

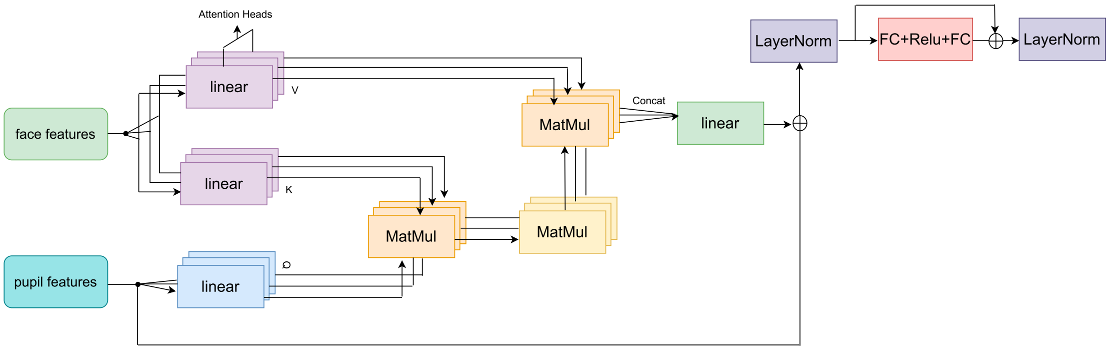

Fig. 3. Self-attentional structure used for expressions and pupils.

3) Cross-Attention Module: By encoding the two modalities, we obtain high-level semantic features of facial expressions and eye movements. To make more accurate final decisions, we use complementary intramodality information fusion between the two modalities. Specifically, we first use the self-attention mechanism to perform intramodality representation learning on the facial expression modality. The utilization of a feedforward layer featuring ReLU activation function on the self-attention module output empowers the model to acquire and adjust to the crucial high-level features of facial expressions. As a result, the model becomes more attentive to the features that carry greater weight in determining the final output result. By utilizing the self-attention mechanism, the model can effectively incorporate the impact of adjacent feature elements, while its query, key, and value components represent the modalities projected onto separate spaces. As shown in Fig. 3.

To establish a connection between facial expressions and pupils, this article optimized the multihead self-attention network by feeding data from both modalities concurrently, as opposed to the approach depicted in Fig. 2. The following equation for computation is given as:

$$\text{MultiHead}(S,T,T) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^o \quad (12)$$

$$\text{head}_h = \text{Attention}(SW_h^Q, TW_h^K, TW_h^V). \quad (13)$$

The provided equation denotes S as the input sequence of pupil data, while T is utilized to represent the input sequence of facial expression data. Within the multihead self-attention network, S is multiplied by the trainable weight matrix W_h^Q for linear transformation of attention Q, while T is multiplied by the trainable weight matrices W_h^K and W_h^V for linear transformation of attention K and V, respectively. In this context, h refers to the quantity of heads contained within the multihead training network. The incorporation of multihead self-attention allows the model to concentrate on multiple crucial regions in parallel. The calculation of the attention module requires the use of scaled dot product:

$$\text{Attention}\left(SW_h^Q, TW_h^K, TW_h^V\right) = \text{softmax}\left(\frac{SW_h^Q * TW_h^K}{\sqrt{d_k}}\right) TW_h^V. \quad (14)$$

In the equation involving $SW_h^Q * TW_h^K$, d_k represents the final dimension of S and T. To avoid potential issues with excessively large values in the multiplication of $SW_h^Q * TW_h^K$ caused by a large d_k, it is divided by the square root of d_k. This helps to prevent the gradient of the softmax function from becoming too small. After calculating the attention values in the model, a LayerNorm layer is applied to normalize these values. Introducing this normalization step enhances the stability and speed of convergence for the network while undergoing training. In the multilayer network architecture, the following equation is shown as:

$$H_l = \text{LayerNorm}(S + \text{MultiHead}(S,T,T)) \quad (15)$$

In this context, S refers to the input sequence of pupils. LayerNorm is a layer normalization technique that normalizes each feature of the input data using learnable parameters, including scaling and shifting factors, as well as statistical information such as mean and standard deviation. This resolves the issue of gradient vanishing while also enhancing generalization ability. Following its passage through the attention mechanism and LayerNorm, the output H_l proceeds to a feedforward layer containing two fully connected layers. According to the design of the network, the initial layer in the feedforward layer implements the ReLU activation function, while the subsequent layer does not use any activation function

$$H_m = \max(0, H_l W_5 + b_5)W_6 + b_6. \quad (16)$$

Similarly, normalization is used again with the LayerNorm layer, where S_T is the facial expression feature that is more closely related to the pupil feature obtained through the attention module, as shown in

$$S_T = \text{LayerNorm}(H_l + H_m). \quad (17)$$

Subsequently, we leverage multihead attention to identify the shared characteristics between facial expressions and pupils, with the calculation (18) outlined as follows:

$$\text{MultiHead}(T,S,S) = \text{Concat}(\text{head}_1, \ldots, \text{head}_l)W^o \quad (18)$$

$$\text{head}_l = \text{Attention}(TW_l^Q, SW_l^K, SW_l^V). \quad (19)$$

In (18), T represents the input sequence of facial expressions, and S represents the input sequence of pupils. In the multihead self-attention network, T is multiplied by the trainable weight matrix W_l^Q for linear transformation of attention Q, while S is multiplied by the trainable weight matrices W_l^K and W_l^V for linear transformation of attention K and V, respectively. The attention module requires the use of scaled dot-product for calculation, as shown in the following equation:

$$\begin{aligned}&\text{Attention}\left(TW_1^Q, SW_1^K, SW_1^V\right)\\&\quad= \text{softmax}\left(\frac{TW_l^Q * SW_l^K}{\sqrt{d_k}}\right) SW_l^V. \qquad (20)\end{aligned}$$

Here, d_k is the last dimension of the shapes of S and T and is divided by $\sqrt{d_k}$ to prevent $SW_h^Q * TW_h^K$ from becoming too large when d_k is too large, which can cause the softmax function's gradient to become too small. Subsequently, the attention values produced by the network undergo normalization through the LayerNorm layer, which serves to enhance the network's stability and speed of convergence throughout the training process. The formula typically utilized within multilayer network architecture is as follows:

$$H_a = \text{LayerNorm}(T + \text{MultiHead}(T, S, S)). \qquad (21)$$

Here, T is the input sequence of facial expressions, and LayerNorm is the layer normalization. Then, H_a is inputted into a feedforward layer

$$H_b = \max(0, H_a W_7 + b_7) W_8 + b_8. \qquad (22)$$

Then, normalization is used again with the LayerNorm layer, where T_S is the pupil feature that is more closely related to the facial expression feature obtained through the attention module:

The following equation represents the computation obtained by integrating the features of all four modalities:

$$K = \text{Concat}(S_E, T_E, S_T, T_S). \qquad (23)$$

The output multimodal feature K is processed further by a max pooling layer, which transforms it into a vector. The next step involves the application of a fully connected layer, along with the ReLU activation function, to subject the vector to a nonlinear transformation. Last, classification is performed utilizing a softmax layer.

IV. Experiments

A. Experimental Paradigm

When exposed to the same emotional stimuli, depressive subjects and healthy controls display various behavioral patterns. Furthermore, previous research has found that individuals with depression tend to exhibit a bias toward negative information when exposed to different stimuli paradigms [52]. Therefore, in this study, different stimuli paradigms were designed to induce emotional changes in participants based on the behavioral differences observed between individuals with depression and healthy individuals. To this end, we created a stimulus task capable of evoking different expressive emotions, using a traditional psychological experimental paradigm as a basis in this process. During the assessment, participants' emotional states were evoked by freely watching videos. During the study, three different states of emotional pictures (including positive, neutral, and negative) were used to induce changes in viewer expressions and pupil behavior.

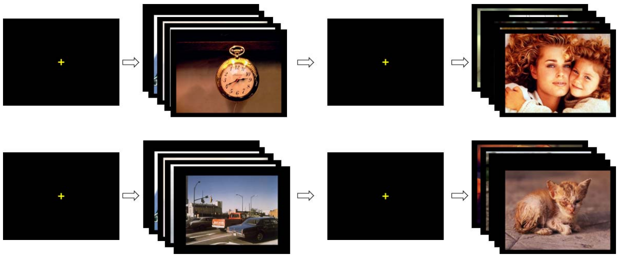

Fig. 4. Paradigm process.

The emotional images were sourced from the international affective picture system (IAPS), which was developed by the American Center for Emotion and Attention at the National Institute of Mental Health (NIMH) [53]. IAPS contains a wide variety of content and a very wide range. Different categories of pictures can induce people's emotional expressions in different states. It is used in different fields, including psychology, neurophysiology, and brain cognitive science, but most importantly, it has played a huge role in the study of emotion and attention [54], [55], [56], where multidimensional scores can help research people design targeted experimental paradigms [57]. In the IAPS emotional picture database, we selected 40 pictures of different types, including 10 positive stimulus types (valence: 5.03 ± 1.15, arousal: 2.91 ± 1.97), 20 neutral stimulus types (valence: 7.43 ± 1.48, arousal: 4.33 ± 2.27), and 10 negative stimulus types (valence 2.95 ± 1.62, arousal 5.35 ± 2.24).

The whole experimental process is mainly composed of two parts, each part is composed of four types of experiments. In each experiment, it can be further divided into two parts, one is for the refocusing of the fixation point, and the other is for the playback of the stimulus pictures. We first play the sequence of stimulus pictures in the order of neutral, positive, and negative. Then loop once in the same playback standard and order. Considering the possibility that the emotional effect of the previous group will interfere with the emotional expression of the next group, each time before the next group of pictures is played, there will be a five-second pause to buffer the possible continuous reaction. The process of experimental video playback is depicted in Fig. 4. The number of pictures displayed in each group of small experiments is 5, and the playback time of each picture is 5 s. The total video duration of the entire experiment is 245 s. Table I shows the order of picture types in the eight groups of experiments in the video paradigm. The same stimulus paradigm was also used in the work of Yang et al. [58].

B. Equipment Setup

As described in the previous section, depressed patients presented different emotional states than healthy controls under different stimulation paradigms. This requires that different cameras need to be used to record the changes in facial

TABLE I
STIMULUS PARADIGM [58]

Continuous Stimulus			Block1																							
	Trial	Focus	Neutral					Focus	Positive					Focus	Neutral					Focus	Negative					Focus
	Number	“ + ”	1	2	3	4	5	“ + ”	6	7	8	9	10	“ + ”	11	12	13	14	15	“ + ”	16	17	18	19	20	“ + ”
	Picture		Room	Basket	Outlet	Clock	Clothespins		Puppies	Family	Seal	Butterfly	Athletes		Chair	Coffee cup	Key ring	Book	Abstract art		Mutilation	Toilet	Burn	Shadow	Hand	
	Duration Time (s)	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
			Block2																							
	Trial		Neutral					Focus	Positive					Focus	Neutral					Focus	Negative					Focus
	Number		21	22	23	24	25	“ + ”	26	27	28	29	30	“ + ”	31	32	33	34	35	“ + ”	36	37	38	39	40	“ + ”
	Picture		Plate	Abstract art	Bowl	Clock	Tool		Giraffes	Bunnies	Women	Mother	Adult		Cabinet	Spoon	Light bulb	Mug	Mug		Snake	Victim	Snakes	Baby	Sick kitty	
	Duration Time (s)		5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5

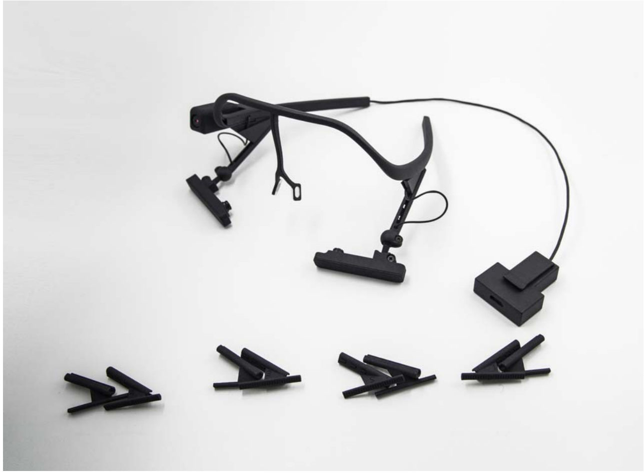

Fig. 5. Eye tracker used in this study.

expressions and eye-movement data of each subject while watching the video at the same time during the experiment.

The most popular type of camera for gathering facial expression data is an optical camera, however professional eye-tracking equipment with high accuracy is needed to record eye movements. To use the desktop eye tracker, the subjects were required to place their heads on the chin rest. Although it effectively recorded the subjects' eye movement information, it affected their facial behavior. However, the wearable eye tracking devices on the market also have some problems, such as the occlusion of key information in the face area. The eye tracker used in this article is a redesigned eye tracking device [59], which has the advantage of effectively avoiding the occlusion of the facial expression information closed area, which is shown in Fig. 5, and the eye tracking algorithm and software are ported from [60].

To collect facial expressions and pupil data synchronously, the capture processes and paradigm display process are synchronized by real-time signal, which is implemented on a hard real-time operating system. We designed the acquisition scenario as shown in Fig. 6. The parameter settings used are as follows: the eye tracking frame rate is 200 fps, and the resolution is 320 × 200P, the facial expression videos are capture by Logitech C1000 which worked at the frame rate of 30 fps and the resolution of 1920 × 1080P.

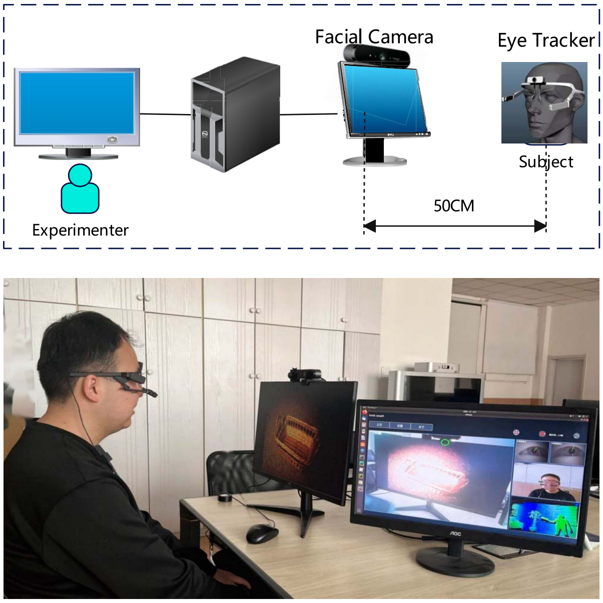

Fig. 6. Proposed experimental setup.

C. Participants Selection

Before the start of the experiment, all subjects were aware of the experimental procedure and signed a written informed consent [24]. The biomedical research consent form was approved by the local ethics committee in accordance with the World Medical Association's Code of Ethics, allowing the research to be conducted at the Third People's Hospital of Guangyuan (Declaration of Helsinki). Depressed patients in all of our subjects participated in a structured Mini International Neuropsychiatric Interview (M.I.N.I.) interview to judge whether they met the DSM-IV criteria for depression, and the results showed that the patients involved met this criterion. For this study, all participants were aged between 18 and 55 years old [61] and possessed at least an elementary school level of education. Each participant was screened for various exclusion criteria, including organic brain disease, history of epilepsy or cranial trauma, history of drug or substance abuse in the past six months, and concomitant serious physical illness or high risk of suicide. The psychiatrist scored each participant by interview and questionnaire. Among the questionnaires administered during the study were the Patient Health Questionnaire (PHQ-9) and the

TABLE II
RESULTS OF USING DIFFERENT CLASSIFIERS FOR MULTIMODAL FUSION FEATURE UNDER DIFFERENT STIMULUS MODES

		SVM	KNN	RF	DTREE
Positive	Accuracy	0.721	0.734	0.711	0.706
	Precision	0.691	0.722	0.711	0.645
	Recall	0.713	0.710	0.625	**0.783**
	F1 score	0.708	0.710	0.677	0.744
Neutral	Accuracy	0.730	0.713	0.732	0.720
	Precision	0.776	0.667	0.764	0.752
	Recall	0.723	**0.751**	0.754	0.780
	F1 score	0.772	0.712	0.702	0.741
Negative	Accuracy	0.726	0.738	0.722	0.720
	Precision	**0.781**	0.731	**0.836**	0.674
	Recall	0.702	0.629	0.695	0.750
	F1 score	0.743	0.710	0.761	0.712

Note: The bold entries represent the highest metrics of different classifiers among three stimulus modes.

TABLE III
COMPARISON BETWEEN EXISTING FEATURE EXTRACT METHODS AND OUR METHODS

Method	Accuracy	Precision	Recall	F1 Score
Facial Expression				
LBP	0.564	0.560	0.667	0.610
HOG	0.566	**0.818**	0.600	0.692
LBP-TOP	0.610	0.571	0.502	0.530
ResNext	**0.615**	0.621	**0.802**	**0.706**
Pupil				
T-frequency domain	0.673	0.670	0.405	0.509
1-D CNN	**0.723**	**0.732**	**0.801**	**0.766**

Note: The bold entries represent the best feature extract methods under accuracy, precision, recall and f1 score.

TABLE IV
COMPARISON OF DIFFERENT FUSION MODELS

Method	Accuracy	Precision	Recall	F1 Score
CCA	0.724	0.722	0.730	0.725
SCCA	0.713	0.701	0.711	0.707
TCCA	0.736	**0.752**	0.703	0.724
CMF-SN	**0.750**	0.756	**0.800**	**0.778**

Note: The bold entries in represent the best fusion models under accuracy, precision, recall and f1 score.

International Neuropsychiatric Interview (MINI). The PHQ-9 was classified as healthy controls: <5 and patients: ≥ 5 according to the PHQ-9 criteria, and this score was used as a label. In the dataset, some invalid data were removed. For example, during the acquisition process, some faces in the video disappeared for a long time, and frequent body movements covered the faces of the subjects. We used 57 valid data, including 25 depression patients (7 males and 18 females; age range 18–55 years) and 32 healthy controls (2 males and 32 females; age range 18–55 years). We divided dataset into training and test sets in a ratio of 8 to 2, without dividing the validation set. Meanwhile, we fixed the gender ratio of training dataset and test dataset when we dividing dataset multiple times. Finally, all metrics took average value of multiple experimental results.

D. Experiment Process

All subjects were individually invited to the laboratory, read written descriptions of experimental objectives and procedures and gave their informed consent. As part of the experiment, participants were situated within a soundproof and windowless room, where they were seated about 50–70 cm away from a computer screen. Ceiling lighting produced stable illumination conditions. An eye tracking device was needed for the experiment to record changes in the subjects' pupils and a desktop camera to record changes in the subjects' expressions as they viewed the film. Before the experiment began, the researchers told the subjects through a computer screen that each stimulus image started at a fixation point (white cross on a black background) until the fixation point was concentrated. If the fixation point deviation is too large, the calibration procedure is repeated until the calibration is successful to start the experiment. Subjects watched the displayed pictures naturally, and pictures of all stimulus types were equally likely to appear in each instance, only once at a time, and the experiment lasted approximately 4 min. As part of the experiment, participants were advised to limit any movements of their body or head to the greatest extent possible.

E. Experiment Details

In the course of the experiment, the expression image and pupil image used are grayscale images. In the process of the experiment, to ensure that the extracted expression and the pupil diameter change in the same frame, we choose the image of the same time frame for use, so as to ensure the consistency of expression and pupil. The relevant parameters in the learning process are obtained through random optimization. SVM, RF, KNN, and DTree algorithms are used for classification modeling, and the trials of negative, positive, and neutral emotional stimulus are investigated, respectively. In the facial expression feature extraction, the sliding window value is set to 12. The SVM algorithm selects a radial basis function (RBF) kernel for use in the analysis. The parameter K in the KNN algorithm is set to 3. The parameter criterion in the RF algorithm is set to gini. The parameter n_estimators in the DTree algorithm is set to 12.

V. RESULTS

The algorithm was applied to both facial expression and pupil data, and a series of experiments were conducted. Deep

TABLE V
COMPARISON OF CMF-SN ABLATION EXPERIMENTS

		SVM	KNN	RF	DTREE
ResNext	Accuracy	0.615	0.674	0.667	0.611
	Precision	0.621	0.724	0.751	0.673
	Recall	0.802	0.704	0.605	0.702
	F1 score	0.706	0.655	0.674	0.641
1-D CNN	Accuracy	0.723	0.721	0.720	0.716
	Precision	0.732	0.695	0.781	0.781
	Recall	0.801	0.812	0.703	0.703
	F1 score	0.766	0.780	0.745	0.745
Res+Att	Accuracy	0.674	0.705	0.679	0.691
	Precision	0.649	0.625	0.752	0.602
	Recall	0.702	0.805	0.607	0.823
	Precision	0.677	0.701	0.674	0.721
1-D CNN+Att	Accuracy	0.720	0.723	0.720	0.719
	Precision	0.718	0.703	0.679	0.741
	Recall	0.810	0.812	0.800	0.705
	Precision	0.760	0.754	0.731	0.722
Res+Att+1-D CNN+Att	Accuracy	0.723	0.716	0.720	0.721
	Precision	0.742	0.685	0.690	0.740
	Recall	0.831	0.750	0.792	0.800
	Precision	0.783	0.722	0.737	0.790
CMF-SN	Accuracy	**0.750**	**0.730**	**0.725**	**0.730**
	Precision	0.756	0.720	0.742	0.739
	Recall	0.800	0.702	0.780	0.752
	Precision	0.778	0.711	0.761	0.745

Note: The bold entries mean that our method has achieved the best performance in accuracy.

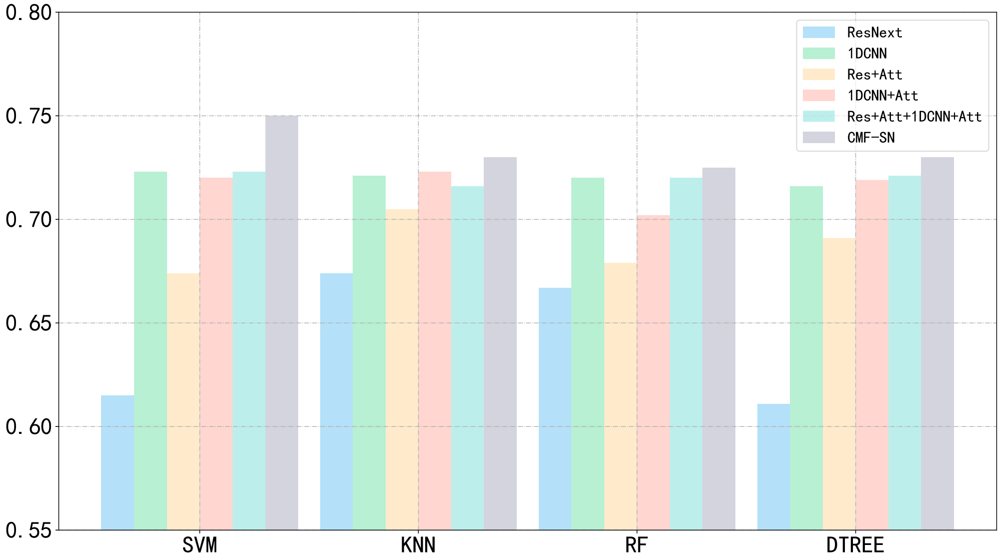

Fig. 7. Comparison of ablation results.

information was extracted from both modalities and CMF was performed. To ensure that the original semantic information was not lost during the interaction within and between the modalities, this study also used parallel self-attention networks for feature selection of both facial expression and pupil features during CMF, thus obtaining features within and between the modalities. Finally, the obtained features were fully connected for effective depression recognition. Several commonly used classification models in Python, including SVM, KNN, RF, and DTREE, were selected for depression classification. The experimental results, as shown in Table II, demonstrated that the classification accuracy reached over 70% for all three stimulus states using SVM as an example. In addition, the effective combination of ResNexts and self-attention networks helps to ensure the integrity of information interaction between different modalities. Therefore, this study further evaluated the performance of this method on a private dataset, and the experimental results proved its effectiveness, achieving a classification accuracy rate of 75.0%. Furthermore, previous research [52] has shown that depressed patients may have different reactions to different stimuli and exhibit more attentional biases toward negative stimuli. Therefore, we further analyzed and validated the study on this basis, and the results showed that the classification accuracy rates for negative stimuli and neutral stimuli were 72.6% and 73.0%, respectively, while the classification accuracy rate for positive stimuli was 72.1%.

In the single-modal experiments, as shown in the table, we compared our deep models with previous studies using traditional methods. In the facial expression modality, we used traditional methods such as LBP, HOG, and LBP-TOP for experimentation in Table III, while in the pupil modality, we used time-frequency domain features and then used an SVM classification model for classification. The tabular data indicate that traditional learning models had a relatively low classification accuracy for depression recognition utilizing exclusively facial expressions or pupil features, when compared to the higher accuracy achieved through the use of deep learning-based facial expression and pupil feature extraction techniques. We took SVM as an example and examined the variance and bias in the classification model by obtaining the learning curve. To validate the classification performance of our proposed approach, we compared it with existing feature fusion techniques using our collected dataset. The comparison results are presented in Table IV. Canonical correlation analysis (CCA) [62] is a statistical technique for analyzing multiple variables or factors, which can be utilized to investigate the correlation between two datasets. It projects the two datasets into new low-dimensional feature spaces and maximizes their correlation. Kernel canonical correlation analysis (KCCA) [63] is a commonly used multivariate statistical analysis method that maps data to high-dimensional feature spaces through kernel functions and calculates the correlation between the two datasets in the feature space. Supervised canonical correlation analysis (SCCA) [64] can find enhanced correlation subspaces for two observation spaces by adding class label information, in which the mapping of the same pattern to the observation has the highest correlation, and the enhanced correlation subspaces have stronger pattern recognition and semantic discrimination ability. The results showed that the model achieved a relatively good accuracy in depression identification, further demonstrating the effectiveness of the fusion results obtained by the proposed method.

To gain a deeper understanding of the impact of self-attention on the ultimate recognition performance of the two components within CMF-SN, we divided them and carried out comparative experiments independently. The experimental outcomes are presented in Table V. Res+Att denotes the use of ResNext-50 (32 × 4d) for feature extraction of facial expressions, and then selecting features using self-attention network. Similarly, 1-D CNN+Att refers to the use of 1-D CNN for feature extraction of pupils, and then selecting features using self-attention network. Res+Att+1-D CNN+Att refers to the parallel fusion of features from both modalities using self-attention network, without including fusion features within modalities. CMF-SN is the result of considering both intramodality and intermodality fusion. It can be seen that using only features extracted by a single-modality deep network for depression recognition results in a relatively lower recognition accuracy compared to using self-attention network feature selection. Both facial expressions and pupils experience a certain degree of reduction in accuracy across different classifiers. However, despite being compared to traditional single-modality methods for depression recognition, the proposed approach still demonstrated a considerable improvement in recognition accuracy.

The experimental findings depicted in Fig. 7 provide evidence that the proposed CMF-SN feature fusion approach utilizing self-attention networks can effectively amalgamate information from various modalities. The final recognition performance is improved compared to the results of individual modalities. Based on the conclusive experimental results, it is apparent that the recognition performance obtained solely through self-attention network fusion between modalities is inferior to the performance achieved through utilizing both intramodality and intermodality fusion. This is because the optimization of the multihead self-attention network in this article learns the similarity between facial expression and pupil features, obtaining pupil features that are more closely related to facial expression features, and facial expression features that are more closely related to pupil features, achieving complementarity of information. This can better capture some key information that was missed in the fusion between modalities, indicating that the proposed method can more comprehensively model the fusion between multiple modalities and accurately fuse the information expressed by different inputs, leading to better prediction of depression. Furthermore, this further validates the superiority of the feature fusion algorithm proposed in this article.

VI. Conclusion

The main focus of this study was to propose a CMF model for depression identification based on self-attention networks. This method considered the complementarity within and between different modalities, used cross-modal blocks to fuse the features within each modality, and then utilized the complementarity of information between modalities for multimodal fusion, achieving feature fusion. Based on the experimental results, the proposed method exhibited effectiveness and superiority in depression recognition research. However, a certain number of samples still have not been correctly classified, and most of them were males. We speculated that two factors cause such a phenomenon. First, there was imbalanced gender distribution problem in our collected dataset. Besides, the amount of our dataset was limited, which may lead to insufficient model training. Therefore, we were continuously collecting new data. In the future work, we may alleviate these problems. The results also suggested that there may be a bias toward negative stimuli. In subsequent studies, we will further compare different datasets to explore this. Another limitation of this article is that only the fusion of pupil diameter and facial expression is considered, which may weaken the recognition accuracy. In the following studies, the fusion of multiple features can be considered to improve the accuracy of depression recognition.

References

[1] D. Santomauro, J. Amm Herrera, P. Zheng, and A. Ferrari, "Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic," *Lancet*, vol. 398, no. 10312, pp. 1700–1712, 2021.

[2] K. Skonieczna-Żydecka et al., "Faecal short chain fatty acids profile is changed in polish depressive women," *Nutrients*, vol. 10, no. 12, 2018, Art. no. 1939.

[3] J. Park and I. Lee, "Factors influencing suicidal tendencies during COVID-19 pandemic in Korean multicultural adolescents: A cross-sectional study," *BMC Psychol.*, vol. 10, no. 1, 2022, Art no. 158.

[4] A. Chadha and B. Kaushik, "Performance evaluation of learning models for identification of suicidal thoughts," *Comput. J.*, vol. 65, no. 1, pp. 139–154, 2022.

[5] Z. Sarhan, H. Shinnawy, M. Eltawil, Y. Elnawawy, W. Rashad, and M. Mohammed, "Global functioning and suicide risk in patients with depression and comorbid borderline personality disorder," *Neurology, Psychiatry Brain Res.*, vol. 31, pp. 37–42, 2019, doi: 10.1016/j.npbr.2019.01.001.

[6] S. Ghosh, A. Ekbal, and P. Bhattacharyya, "A multitask framework to detect depression, sentiment and multi-label emotion from suicide notes," *Cogn. Comput.*, vol. 14, no. 1, pp. 110–129, Jan. 2022.

[7] J. Sotelo and C. Nemeroff, "Depression as a systemic disease," *Pers. Med. Psychiatry*, vol. 1, pp. 11–25, 2017, doi: 10.1016/j.pmip.2016.11.002.

[8] A. Nassibi, C. Papavassiliou, and S. Atashzar, "Depression diagnosis using machine intelligence based on spatiospectrotemporal analysis of multi-channel EEG," *Med. Biol. Eng. Comput.*, vol. 60, no. 11, pp. 3187–3202, 2022.

[9] T. Richter, B. Fishbain, E. Fruchter, G. Richter-Levin, and H. Okon-Singer, "Machine learning-based diagnosis support system for differentiating between clinical anxiety and depression disorders," *J. Psychiatric Res.*, vol. 141, pp. 199–205, 2021, doi: 10.1016/j.jpsychires.2021.06.044.

[10] B. Hu, Y. Tao, and M. Yang, "Detecting depression based on facial cues elicited by emotional stimuli in video," *Comput. Biol. Med.*, vol. 165, 2023, Art. no. 107457.

[11] X. Li, W. Guo, and H. Yang, "Depression severity prediction from facial expression based on the DRR_DepressionNet network," in *Proc. IEEE Int. Conf. Bioinf. Biomed., (BIBM)*. Piscataway, NJ, USA: IEEE Press, 2020, pp. 2757–2764.

[12] A. Pampouchidou et al., "Automatic assessment of depression based on visual cues: A systematic review," *IEEE Trans. Affect. Comput.*, vol. 10, no. 4, pp. 445–470, Oct.–Dec. 2019.

[13] Q. Li, T. Zhang, C. Chen, K. Yi, and L. Chen, "Residual GCB-Net: Residual graph convolutional broad network on emotion recognition," *IEEE Trans. Cogn. Devel. Syst.*, vol. 15, no. 4, pp. 1673–1685, Dec. 2023.

[14] C. Nicolas et al., "Eye movement in unipolar and bipolar depression: A systematic review of the literature," *Frontiers Psychol.*, vol. 6, 2015, Art. no. 1809.

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

[16] 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),* Piscataway, NJ, USA: IEEE Press, 2021, pp. 7973–7977.

[17] M. Yang, Z. Weng, Y. Zhang, Y. Tao, and B. Hu, "Three-stream convolutional neural network for depression detection with ocular imaging," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 4921–4930, 2023.

[18] N. Sekaninova et al., "Oculometric behavior assessed by pupil response is altered in adolescent depression," *Physiol. Res.*, vol. 68, no. 3, pp. 325–338, 2019.

[19] P. Atrey, M. Hossain, A. Saddik, and M. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey," *Multimedia Syst.*, vol. 16, no. 6, pp. 345–379, 2010, doi: 10.1007/s00530-010-0182-0.

[20] S. Tripathi, S. Tripathi, and H. Beigi, "Multi-modal emotion recognition on iemocap dataset using deep learning," 2018, *arXiv:1804.05788*.

[21] H. Zhang, H. Wang, S. Han, W. Li, and L. Zhuang, "Detecting depression tendency with multimodal features," *Comput. Methods Programs Biomed.*, vol. 240, 2023, Art. no. 107702.

[22] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning affective features with a hybrid deep model for audio-visual emotion recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 3030–3043, Oct. 2018.

[23] J. Liu et al., "Multimodal emotion recognition with capsule graph convolutional based representation fusion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP),* Piscataway, NJ, USA: IEEE Press, 2021, pp. 6339–6343.

[24] 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.

[25] M. Yang, C. Cai, and B. Hu, "Clustering based on eye tracking data for depression recognition," *IEEE Trans. Cogn. Devel. Syst.*, vol. 15, no. 4, pp. 1754–1764, Dec. 2023.

[26] T. Yan, X. Zhang, and H. Wang, "Geometric-convolutional feature fusion based on learning propagation for facial expression recognition," *IEEE Access*, vol. 6, pp. 42532–42540, 2018.

[27] Y. Liu, X. Zhang, Y. Lin, and H. Wang, "Facial expression recognition via deep action units graph network based on psychological mechanism," *IEEE Trans. Cogn. Devel. Syst.*, vol. 12, no. 2, pp. 311–322, Jun. 2020.

[28] P. Shivanasab and R. Abbaspour, "An incremental algorithm for simultaneous construction of 2D Voronoi diagram and Delaunay triangulation based on a face-based data structure," *Adv. Eng. Softw.*, vol. 169, 2022, Art. no. 103129.

[29] X. Huang, S. Wang, X. Liu, G. Zhao, X. Feng, and M. Pietikäinen, "Discriminative spatiotemporal local binary pattern with revisited integral projection for spontaneous facial micro-expression recognition," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 32–47, Jan.–Mar. 2017.

[30] L. Liu, S. Lao, P. Fieguth, Y. Guo, X. Wang, and M. Pietikäinen, "Median robust extended local binary pattern for texture classification," *IEEE Trans. Image Process.*, vol. 25, no. 3, pp. 1368–1381, Mar. 2016.

[31] N. Maddage, R. Senaratne, L. Low, M. Lech, and N. Allen, "Video-based detection of the clinical depression in adolescents," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* Piscataway, NJ, USA: IEEE Press, 2009, pp. 3723–3726.

[32] X. Fan and T. Tjahjadi, "A spatial-temporal framework based on histogram of gradients and optical flow for facial expression recognition in video sequences," *Pattern Recognit.*, vol. 48, no. 11, pp. 3407–3416, 2015.

[33] J. Ye et al., "Multi-modal depression detection based on emotional audio and evaluation text," *J. Affect. Disorders*, vol. 295, pp. 904–913, Dec. 2021.

[34] J. Zhu et al., "Content-based multiple evidence fusion on EEG and eye movements for mild depression recognition," *Comput. Methods Programs Biomed.*, vol. 226, 2022, Art. no. 107100.

[35] J. Shen et al., "Depression recognition from EEG signals using an adaptive channel fusion method via improved focal loss," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 7, pp. 3234–3245, Jul. 2023.

[36] G. Zheng et al., "An attention-based multi-modal MRI fusion model for major depressive disorder diagnosis," *J. Neural Eng.*, vol. 20, no. 6, 2023, Art. no. 066005.

[37] 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.*, pp. 2352–8648, Mar. 2023.

[38] S. Ghosh et al., "COMMA-DEER: Common-sense aware multimodal multitask approach for detection of emotion and emotional reasoning in conversations," in N. Calzolari E. Santus, F. Bond, and S.-H. Na, Eds., Gyeongju, Republic of Korea, Int. Committee Comput. Linguistics, 2022, pp. 6978–6990.

[39] G. Singh, S. Ghosh, A. Verma, C. Painkra, and A. Ekbal, "Standardizing distress analysis: Emotion-driven distress identification and cause extraction (dice) in multimodal online posts," in *Proc. Conf. Empirical Methods Natural Lang. Process.,* H. Bouamor, J. Pino, and K. Bali, Eds., Singapore, Assoc. Comput. *Linguistics*, 2023, pp. 4517–4532.

[40] T. Siddharth and T. Sejnowski, "Utilizing deep learning towards multimodal bio-sensing and vision-based affective computing," *IEEE Trans. Affect. Comput.*, vol. 13, no. 1, pp. 96–107, Jan. 2022.

[41] J. Zhu, C. Yang, X. Xie, S. Wei, Y. Li, X. Li, and B. Hu, "Mutual information based fusion model (MIBFM): Mild depression recognition using EEG and pupil area signals," *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 2102–2115, Mar. 2023.

[42] X. Zhang et al., "Fusing of electroencephalogram and eye movement with group sparse canonical correlation analysis for anxiety detection," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 958–971, Feb. 2022.

[43] Y. Tao, M. Yang, H. Li, Y. Wu, and B. Hu, "DepMSTAT: Multimodal spatio-temporal attentional transformer for depression detection," *IEEE Trans. Knowl. Data Eng.*, early access, Jan. 5, 2024.

[44] S. Ghosh, A. Ekbal, and P. Bhattacharyya, "What does your bio say? Inferring Twitter users' depression status from multimodal profile information using deep learning," *IEEE Trans. Comput. Social Syst.*, vol. 9, no. 5, pp. 1484–1494, Oct. 2022.

[45] Y. Zhang, et al., "Improving brain age prediction with anatomical feature attention-enhanced 3D-CNN," *Comput. Biol. Med.*, vol. 169, 2024, Art. no. 107873.

[46] B. Zhang, D. Wei, G. Yan, T. Lei, H. Cai, and Z. Yang, "Feature-level fusion based on spatial-temporal of pervasive EEG for depression recognition," *Comput. Methods Programs Biomed.*, vol. 226, 2022, Art. no. 107113.

[47] D. Reddy, R. Ramamoorthi, and B. Curless, "Frequency-space decomposition and acquisition of light transport under spatially varying

illumination," in *Proc. Eur. Conf. Comput. Vis.*, New York, NY, USA: Springer-Verlag, 2012, pp. 596–610.

[48] R. Szeliski, *Computer Vision: Algorithms and Applications*. Cham, Switzerland: Springer Nature, 2011.

[49] M. Kwon, Y. Jeong, and H. Choi, "Feature embedding and conditional neural processes for data imputation," *Electron. Lett.*, vol. 56, no. 11, pp. 546–548, 2020.

[50] T. Chen, X. Xu, and S. Wang, "Signal processing by energy normalization method based on wavelet packet," in *Key Engineering Materials*, vol. 413, Trans. Tech. Publ., Switzerland, 2009, pp. 613–619.

[51] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1492–1500.

[52] I. Gotlib, E. Krasnoperova, D. Yue, and J. Joormann, "Attentional biases for negative interpersonal stimuli in clinical depression," *J. Abnormal Psychol.*, vol. 113, no. 1, pp. 121–135, 2004.

[53] B. Cuthbert, "International affective picture system (IAPS): Instruction manual and affective ratings," Center Res. Psychophysiol., Univ. Florida, Tech. Rep. no. a-4., 1999.

[54] M. Bradley, L. Miccoli, M. Escrig, and P. Lang, "The pupil as a measure of emotional arousal and autonomic activation," *Psychophysiology*, vol. 45, no. 4, pp. 602–607, 2010.

[55] D. Sabatinelli, M. Bradley, P. Lang, V. Costa, and F. Versace, "Pleasure rather than salience activates human nucleus accumbens and medial prefrontal cortex," *J. Neurophysiol.*, vol. 98, no. 3, pp. 1374–1379, 2007.

[56] D. Zald, "The human amygdala and the emotional evaluation of sensory stimuli," *Brain Res. Rev.*, vol. 41, no. 1, pp. 88–123, 2003.

[57] T. Libkuman, H. Otani, R. Kern, S. Viger, and N. Novak, "Multidimensional normative ratings for the international affective picture system," *Behav. Res. Methods*, vol. 39, no. 2, pp. 326–334, 2007.

[58] 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 Inform.*, early access, Oct. 5, 2023.

[59] M. Yang, Y. Gao, L. Tang, J. Hou, and B. Hu, "Wearable eye-tracking system for synchronized multimodal data acquisition," *IEEE Trans. Circuits Systems Video Technol.*, early access, Nov. 14, 2023.

[60] M. Kassner, W. Patera, and A. Bulling, "Pupil: An open-source platform for pervasive eye tracking and mobile gaze-based interaction," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2014, pp. 1151–1160.

[61] 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. Comput. Social Syst.*, vol. 10, no. 3, pp. 900–909, Mar. 2023.

[62] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.

[63] W. Zheng, X. Zhou, C. Zou, and L. Zhao, "Facial expression recognition using kernel canonical correlation analysis (KCCA)," *IEEE Trans. Neural Netw.*, vol. 17, no. 1, pp. 233–238, Jan. 2006.

[64] X. Gao, Q. Sun, and H. Xu, "Multiple-rank supervised canonical correlation analysis for feature extraction, fusion and recognition," *Expert Syst. Appl.*, vol. 84, pp. 171–185, Oct. 2017.

Xiang Liu received the Ph.D. degree in electronics science and technology from Beijing Institute of Technology, Beijing, China, in 2019.

He is currently a Teacher with Dongguan University of Technology, Dongguan, China. He had completed a general project of NFSC in 2018 and received a new general project of NFSC in 2020. His research interests include artificial intelligence, machine learning, video coding, communication, multimedia information retrieval, visual information processing, and pattern recognition.

Hao Shen received the B.E. degree in computer science from Lanzhou University, Lanzhou, China, in 2022, where he is currently working toward the M.E. degree in computer science with Gansu Provincial Key Laboratory of Wearable Computing, School of Information Science and Engineering, since 2022.

His research interests include affective computing and natural language processing.

HuiRu Li received the B.Eng. degree in computer science from Taiyuan University, Taiyuan, China, in 2020. She is currently working toward the M.E. degree with Gansu Provincial Key Laboratory of Wearable Computing, School of Information Science and Engineering, Lanzhou University, Lanzhou, China.

Her research interests include affective computing and image processing.

Yongfeng Tao (Graduate Student Member, IEEE) received the B.Ehg.Mgt. degree in information management and system from Tianjin University of Finance and Economics, Tianjin, China, in 2018. He is currently working toward the Ph.D. degree with Gansu Provincial Key Laboratory of Wearable Computing, School of information Science and Engineering, Lanzhou University, Lanzhou, China.

His research interests include affective computing, image processing, and machine learning.

Minqiang Yang (Member, IEEE) received the Ph.D. degree in computer science from Lanzhou University, Lanzhou, China, in 2022.

He is currently an Associate Professor with Gansu Provincial Key Laboratory of Wearable Computing, School of Information Science and Engineering. His research interests include affective computing, image processing, machine learning, and automatic depression detection. He has published more than 20 papers on IEEE Magazines, IEEE Journals, and leading conferences.